

Hypothesis

Not peer-reviewed version

The Precision-Weighting Dysregulation Hypothesis: An Active Inference Framework for Criminal Behaviour, Differential Pathways, and Testable Predictions

[Fritz H. Hemmerich](#) *

Posted Date: 4 March 2026

doi: 10.20944/preprints202603.0291.v1

Keywords: active inference; free energy principle; precision weighting; predictive coding; generative model; computational psychiatry; criminal behaviour; adverse childhood experiences; developmental taxonomy; juvenile delinquency; crimes of passion; recidivism; desistance; moral injury; rehabilitation; inference repair; niche construction; interoceptive inference; HRV biofeedback; life-course criminology



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Hypothesis

The Precision-Weighting Dysregulation Hypothesis: An Active Inference Framework for Criminal Behaviour, Differential Pathways, and Testable Predictions

Fritz H. Hemmerich

Zentrum für Salutogenese, Alhama de Granada, Spain; info@fridhu.com or drhemmerich@gmail.com

Abstract

Background: Criminal behaviour has been modelled through sociological, psychological, and biological lenses, yet mechanistic integration across subjective experience, neural processes, and social niche remains limited. Existing frameworks identify risk factors and correlates but rarely specify the computational mechanism through which adverse environments produce criminal inference. **Framework:** Active Inference (Free Energy Principle) provides a unified computational architecture in which perception, action, and learning are driven by the minimisation of variational free energy under precision-weighted prediction errors. We apply this framework to criminal behaviour, proposing that crime arises from systematic dysregulation of precision weighting within hierarchical generative models, shaped by the interaction between neural architecture and social niche. **Model:** We specify three distinct pathomechanisms: (1) niche-induced prior rigidity, mapping onto life-course-persistent delinquency; (2) vertical hierarchy collapse, mapping onto crimes of passion; and (3) failed inference repair, explaining recidivism as a systemic prediction of the model. **Differential pathways:** The model is transdiagnostic: the same upstream precision-weighting dysregulation can produce criminal behaviour, depression, anxiety, or somatic conditions depending on specific niche configuration, policy-precision parameters, and available action models. We formally specify the conditions under which the externalising (criminal) pathway dominates and argue, from a medical perspective, that this pathway warrants particular attention because it is therapeutically underserved, generates additional self-harm through moral injury, and is addressed by a justice system that systematically reinforces the underlying pathology. **Predictions:** Three falsifiable predictions are derived, each targeting a different pathway: (1) shifted dopaminergic value-learning in juvenile offenders, (2) PFC–amygdala decoupling in crimes of passion, and (3) superiority of inference-based rehabilitation over standard reintegration programmes. The proposed Inference Repair Programme (IRP) targets the generative model itself—not merely behavioural policies—and therefore aims at restoration of inferential flexibility rather than re-normalisation. **Epistemic status:** The Free Energy Principle as a mathematical principle is unfalsifiable, analogous to Hamilton’s principle of least action. Active Inference as a process theory is falsifiable. Empirical validation to date is in a stage of productive exploration: models explain behavioural data reasonably well but have been applied predominantly post hoc (Hodson et al., 2024). The present predictions offer an opportunity to test Active Inference in a novel domain with genuinely novel predictions.

Keywords: active inference; free energy principle; precision weighting; predictive coding; generative model; computational psychiatry; criminal behaviour; adverse childhood experiences; developmental taxonomy; juvenile delinquency; crimes of passion; recidivism; desistance; moral injury; rehabilitation; inference repair; niche construction; interoceptive inference; HRV biofeedback; life-course criminology

1. Introduction

Criminology has long been divided by a persistent schism between sociological perspectives that emphasise structural determinants—poverty, inequality, systemic discrimination—and biological approaches that seek neural or genetic correlates of antisocial conduct. The sociologist fears that biology reduces the offender to a defective organism; the neuroscientist struggles to incorporate social context into brain-level models. The result is a fragmented field in which the subjective experience of the offender—the question of how the world must have appeared to this person for their act to have been the most coherent available response—remains largely unaddressed.

This explanatory gap is not merely academic. It has direct consequences for criminal justice policy. If we cannot specify the mechanism through which social disadvantage produces criminal behaviour, we cannot design interventions that target that mechanism. We are left with interventions that address symptoms (incarceration, surveillance) rather than causes (the inferential architecture that makes criminal behaviour appear rational to the system that produces it). The persistence of recidivism rates between 40 and 70 per cent across Western jurisdictions (Fazel and Wolf, 2015) testifies to the practical cost of this theoretical deficit.

Several landmark contributions have narrowed this gap without closing it. Moffitt's (1993) developmental taxonomy distinguished two qualitatively distinct populations concealed within aggregate crime statistics: a small group of life-course-persistent (LCP) offenders whose antisocial behaviour begins in childhood and continues into adulthood, driven by the cumulative interaction of neuropsychological vulnerability with criminogenic environments, and a much larger group of adolescence-limited (AL) offenders whose delinquency is a transient, normative response to the maturity gap of adolescence. This taxonomy has been cited over 5,000 times and has fundamentally shaped developmental criminology (Fairchild et al., 2013). Steinberg's (2008, 2010) dual-systems model provided a neurodevelopmental mechanism for adolescent risk-taking by identifying the temporal gap between the early maturation of the socioemotional incentive-processing system (ventral striatum, ventromedial prefrontal cortex) and the slower maturation of the cognitive control system (lateral prefrontal cortex, anterior cingulate). This maturational imbalance, confirmed by large-scale behavioural and neuroimaging studies (Shulman et al., 2016), explains why risk-taking peaks in mid-adolescence. Sampson and Laub's (1993) age-graded theory of informal social control demonstrated that "turning points"—marriage, stable employment, military service—can redirect criminal trajectories by strengthening social bonds, yet left open the question of why turning points work for some individuals but not others. [Note: Moffitt's original two-group taxonomy has been empirically refined. Nagin and Land (1993) and subsequent trajectory analyses consistently identify four to six distinct groups rather than two. Sampson and Laub (2003) demonstrated that even the highest-rate offenders largely desisted by age 70. The Weitekamp and Kerner (1994) volume on cross-national longitudinal research confirmed this pattern across multiple cohorts and nations, while Mischkowitz (1994) provided one of the earliest systematic analyses of desistance as a process in its own right. Moffitt herself (2018) updated her model to accommodate expanded trajectory findings. The present framework treats these refinements as consistent with its predictions: a continuous precision-profile space naturally generates multiple trajectory clusters rather than categorical types.]

What none of these frameworks provides is a unified computational mechanism that explains (a) how environments programme neural circuits, (b) why the same environment produces crime in one individual and resilience in another, (c) how acute provocation can override years of moral socialisation in a single moment, and (d) why conventional rehabilitation so often fails. Karl Friston's Active Inference framework, grounded in the Free Energy Principle (FEP), offers precisely such a mechanism. The FEP describes all adaptive systems—from single cells to complex social agents—as engaging in a continuous process of variational free energy minimisation through hierarchical Bayesian inference (Friston, 2010; Parr, Pezzulo, and Friston, 2022). Crucially, this framework treats the organism and its environment (the "niche") as a coupled system: there is no brain without a world, and no behaviour without context. The variable that determines which signals dominate the inference process—and therefore which actions are selected—is precision: the confidence or

reliability assigned to different levels of the predictive hierarchy. Precision weighting is the computational equivalent of attention, salience, and—in the domain of action selection—value. This makes Active Inference neither purely biological nor purely sociological, but inherently systemic.

Friston's early work on value-learning, formulated within the dysconnection hypothesis of schizophrenia (Friston and Frith, 1995; Friston, 1998), established that “value” in neural systems is not an abstract moral quantity but the strength of synaptic transmission—the precision with which one neural population influences another. Learning what is “valuable” means adjusting synaptic efficacy so that future prediction errors are minimised. In the domain of reinforcement learning, dopamine signals encode these prediction errors (FitzGerald, Dolan, and Friston, 2015), driving the organism toward states that maximise its model evidence—its probability of existing in viable, homeostatic states. This paper extends the value-learning framework to criminal behaviour. We propose that delinquency is not a failure of moral reasoning per se, but a consequence of systematic dysregulation in the precision weighting that governs hierarchical Bayesian inference. When the social environment programmes the brain to assign high precision to threat-related, dominance-related, or impulsive signals—and low precision to prosocial alternatives—the resulting behaviour is, from the system's perspective, optimal. It minimises free energy within a generative model calibrated by adversity. [Note: The Free Energy Principle (FEP) as a mathematical principle is unfalsifiable—a point Friston himself has acknowledged. Like Hamilton's principle of least action in physics, it describes a formal constraint that systems may or may not conform to. Active Inference, by contrast, is a falsifiable process theory. The epistemic status of the FEP and the current state of empirical evidence are discussed in detail in Section 7.7.]

Misuse boundary: This framework is structurally incompatible with deterministic or individual-predictive interpretations. It treats criminal behaviour as niche-shaped inference under precision dysregulation and therefore predicts change under changed inferential conditions. It is presented to generate testable research hypotheses—not to classify persons.

The paper is structured as follows. Section 2 establishes the ontogenetic foundations of Bayesian inference, demonstrating that probabilistic reasoning is an innate capacity present from early infancy. Section 3 reviews the neurobiological architecture of precision weighting and introduces a formal notation for the dynamics at stake. Section 4 situates the present framework within the history of biological criminology and provides a principled argument for why Active Inference is structurally incompatible with deterministic ideology. Section 5 presents the Precision-Weighting Dysregulation Model proper, specifying three pathomechanisms and mapping them onto established criminological constructs. Section 6 generates three falsifiable predictions with detailed multimodal study protocols. Section 7 discusses implications for criminal responsibility, connections to adjacent research programmes, and limitations. Section 8 concludes. This is a hypothesis paper: it presents a mechanistic framework and falsifiable research predictions. It provides no clinical, juridical, or policy guidance and is not an operational framework for forensic use.

2. Ontogenetic Foundations of Bayesian Inference

Before examining how inference can go wrong, we must establish that Bayesian inference is a fundamental—not acquired—capacity of the human mind. The argument for applying Active Inference to criminal behaviour rests on the premise that every human brain, including the brain of every offender, operates as a prediction machine that continuously updates its model of the world based on incoming evidence, weighted by precision estimates. A robust body of developmental research confirms this premise.

2.1. *The Infant as Intuitive Statistician*

In a landmark series of experiments, Téglás et al. (2007, 2011) demonstrated that 12-month-old infants form rational probabilistic expectations about future events. When shown a container with three yellow and one blue object, infants looked significantly longer when the blue (less probable) object exited—indicating violation of their statistical expectation. Critically, their looking times were

quantitatively consistent with a Bayesian ideal observer integrating multiple variables simultaneously, including object number, surface area, and spatiotemporal trajectory (Téglás et al., 2011, p. 1054). Xu and Garcia (2008) extended this finding to 8-month-old infants using ping-pong balls drawn from covered boxes, demonstrating that infants make bidirectional inferences—from samples to populations and from populations to samples. Denison, Reed, and Xu (2013) pushed the boundary further, finding evidence for probabilistic inference in infants as young as 6 months. In a comprehensive review, Denison and Xu (2019) concluded that the human brain comes equipped with domain-general statistical learning mechanisms that operate long before the onset of formal instruction, language, or moral development.

2.2. Implications for Criminology

These developmental findings carry a radical implication for the study of criminal behaviour: the capacity for statistical inference about the world is not a product of education, culture, or moral development. It is a biological endowment. If a child grows up in an environment dominated by threat, unpredictability, and violence—an environment in which, to borrow Téglás’s experimental metaphor, the container is filled predominantly with “red balls” (danger)—the infant’s inference engine will correctly learn to expect red balls. When the adult later encounters a “blue ball” (an opportunity for trust, cooperation, safety), the system registers a prediction error, but if the precision of the lifelong prior is sufficiently high, the blue ball is treated as noise rather than signal—that is, as input to which the system assigns low precision, treating it as random fluctuation that does not warrant updating the generative model because its weight is negligible relative to the dominant prior. This is not a defect of reasoning. It is Bayes-optimal inference given the available data. The criminological consequence is profound: what the justice system labels as “criminal thinking” may, in many cases, represent statistically rational inference within a generative model that has been calibrated by a pathological niche.

2.3. Differential Pathways: Why Criminal Behaviour and Not Depression?

An obvious objection arises: if adverse childhood environments produce precision-weighting dysregulation, why should the outcome be criminal behaviour rather than depression, anxiety, or withdrawal? The epidemiological literature makes clear that adverse childhood experiences (ACEs) produce both externalising outcomes (aggression, delinquency) and internalising outcomes (depression, anxiety)—often comorbidly, with overlap rates of 40–60% (Felitti et al., 1998). The model presented here does not claim that precision-weighting dysregulation selectively produces criminality. It claims that the same upstream dysregulation can produce multiple outcomes depending on specific configuration parameters.

Within the Active Inference framework, the formal distinction is as follows. Withdrawal and depression minimise free energy through exposure reduction: the system limits sensory input to minimise prediction errors, producing reduced exploratory behaviour, social withdrawal, and anhedonia. In formal terms, pessimistic priors carry high precision while policy precision (β) is low—the system commits weakly to action because exploration appears futile. Aggression and criminal behaviour, by contrast, minimise free energy through active environmental modification: the system changes the niche to match its predictions. This is Active Inference in the most literal sense. Policy precision (β) is high for dominance-based policies—the system commits strongly to action sequences that assert control over the environment.

The conditions that determine which pathway dominates include: (a) niche type—environments that are unpredictable and uncontrollable favour internalisation, while environments that are threatening but controllable through dominance favour externalisation; (b) dopaminergic calibration—reward systems downregulated toward anhedonia favour internalisation, while systems calibrated to dominance and status reward favour externalisation; (c) policy precision—low β (action appears futile) favours internalisation, high β for antisocial policies favours externalisation; (d) available policy models—absence of aggressive models favours internalisation, availability and

reinforcement of aggressive models favours externalisation; (e) peer niche—isolation or prosocial peers favour internalisation, antisocial peer groups with modelling favour externalisation.

This differential specification transforms a potential weakness into a strength: the model is transdiagnostic. It predicts not only criminal behaviour but also depression, anxiety, somatic conditions (cf. Hemmerich, 2025, on the locus coeruleus norepinephrine depletion hypothesis of ME/CFS), and their frequent comorbidity as alternative manifestations of the same precision-weighting dysregulation under different parametric configurations. This is consistent with the clinical reality that delinquent adolescents frequently present with depressive symptoms, and that depressed individuals show elevated risk for violent offences (Fazel and Wolf, 2015).

The present paper focuses specifically on the criminal pathway for a reason that is medical rather than criminological: among the various manifestations of precision-weighting dysregulation, the criminal pathway is the one for which therapeutic infrastructure is most lacking. Depression has CBT, SSRIs, MBCT. Anxiety disorders have exposure therapy. PTSD has EMDR and prolonged exposure. The person whose precision-weighting dysregulation has led to criminal behaviour typically enters the justice system rather than the health system—and encounters an institutional environment (incarceration) that, as Section 7.7 will argue, systematically reinforces the very pathology it purports to address.

Moreover, criminal behaviour generates an additional mechanism of self-harm that is absent in depression or anxiety: moral injury. This concept, developed in military psychology (Shay, 1994; Litz et al., 2009) and recently examined through a predictive processing lens (Xanthios and Saffaran, 2023), describes the profound psychological damage that results from acting against one's own moral core. In Active Inference terms, moral injury can be formalised as a chronically unresolvable prediction error at the highest level of the generative model—the level that encodes identity (“who I am”). A person can minimise free energy efficiently at the policy level (criminal actions that reduce immediate threat or increase immediate reward) while simultaneously generating chronic prediction error at the identity level, because what the system does is incongruent with what the identity model predicts as self-consistent. This hierarchy-level dissociation—efficient policy execution coupled with chronic identity-level surprise—manifests with temporal delay as shame, inner restlessness, cynicism, relational incapacity, and loss of meaning. The implications for intervention are discussed in Section 6.3 and Section 7.8.

3. Neurobiological Architecture of Precision Weighting

The precision-weighting dysregulation hypothesis requires specification of the neural substrates involved and, crucially, a formal account of the computational dynamics at stake. We propose a multi-level architecture in which three interconnected systems contribute to the regulation of precision across the generative model hierarchy, and we introduce the formal notation that will be used throughout the remainder of this paper.

3.1. The Prefrontal-Amygdala Axis: Hierarchical Control

The medial and ventromedial prefrontal cortex (mPFC/vmPFC) serve as the seat of high-level priors—including moral norms, social expectations, and long-term consequence evaluation. These regions exert top-down precision control over limbic circuits, particularly the amygdala, which encodes the salience and affective precision of environmental stimuli (Adams et al., 2013). In the normally calibrated system, prefrontal priors such as “This action will lead to imprisonment” carry sufficient precision to modulate amygdala-driven impulse signals such as “This provocation demands immediate retaliation.” Effective criminal desistance depends on this hierarchical relationship remaining intact under stress.

Functional connectivity between PFC and amygdala is well documented to be reduced in populations with histories of adverse childhood experiences (ACEs), chronic stress, and antisocial behaviour (Birn et al., 2014). Carlisi et al. (2020), in a landmark neuroimaging study of the Dunedin longitudinal cohort ($n = 672$), demonstrated that individuals on Moffitt's life-course-persistent

trajectory showed reduced cortical surface area in 282 of 360 anatomically defined parcels and thinner cortex in frontal and temporal regions associated with executive function, affect regulation, and motivation, whereas adolescence-limited individuals were neuroanatomically indistinguishable from the low-antisocial group. Within the Active Inference framework, this reduced cortical substrate and diminished functional connectivity corresponds to a structural weakening of top-down precision modulation—the higher levels of the hierarchy lose their capacity to contextualise and regulate lower-level prediction errors.

3.2. The Dopaminergic System: Value-Learning and Prediction Errors

The mesolimbic and mesocortical dopaminergic pathways (ventral tegmental area to nucleus accumbens and prefrontal cortex) encode reward prediction errors—the currency of value-learning (FitzGerald, Dolan, and Friston, 2015; Schwartenbeck et al., 2015). In Active Inference, dopamine signals the precision of beliefs about policies (action plans), determining which actions the agent selects (Friston et al., 2017). This computational role of dopamine provides the bridge between Steinberg’s dual-systems model and the present framework: the adolescent surge in dopaminergic activity that Steinberg identifies as driving reward sensitivity corresponds, in Active Inference terms, to a transient increase in the precision of the socioemotional incentive-processing system relative to the still-immature cognitive control system. The result is precisely the maturational imbalance that Steinberg describes—but now formalised within a single computational currency.

In environments where prosocial behaviour is inconsistently rewarded or actively punished, the dopaminergic system calibrates accordingly: status achieved through dominance or material acquisition via theft generates reliable prediction errors, while cooperation generates only noise. This is not aberrant dopamine function—it is dopamine performing its designated computational role within a niche that systematically rewards antisocial policies. The system learns that dominance-based action plans carry higher policy precision (π) than prosocial alternatives, and this learned precision hierarchy, once established through repeated exposure during critical developmental periods, becomes increasingly resistant to revision.

3.3. Formal Notation: Precision Dynamics in the Generative Model

To anchor the model in established formalism, we briefly state the relevant equations following Parr, Pezzulo, and Friston (2022) and Friston et al. (2017). In Active Inference, an agent selects policies (π) to minimise expected free energy (G):

$$G(\pi) = E[\ln Q(s\tau | \pi) - \ln P(o\tau, s\tau)] = \text{risk} + \text{ambiguity} \quad (1)$$

where the risk term drives the agent to seek outcomes consistent with its preferences (prior beliefs about preferred observations), and the ambiguity term drives epistemic, information-seeking behaviour. The probability of selecting a given policy is determined by a softmax function over negative expected free energy, modulated by a precision parameter β (inverse temperature):

$$P(\pi) = \sigma(-\beta \cdot G(\pi)) \quad (2)$$

This parameter β is critical: it controls the confidence the agent places in its own policy evaluation. High β means the agent strongly commits to the policy with the lowest expected free energy; low β produces more exploratory, stochastic behaviour. In neural terms, β is thought to be encoded by dopaminergic projections from the ventral tegmental area to the striatum (Friston et al., 2017; Schwartenbeck et al., 2015). The precision of sensory prediction errors at each level of the hierarchy is encoded by a separate set of parameters (π_i), which modulate the gain of ascending prediction error signals. High precision at lower levels (e.g., amygdala-mediated affective signals) amplifies bottom-up influence; high precision at higher levels (e.g., prefrontal moral and legal priors) amplifies top-down contextualisation. The central claim of this paper can now be stated formally: criminal behaviour emerges when the precision profile $\Pi = \{\pi_1, \pi_2, \dots, \pi_n\}$ across the generative model

hierarchy is configured such that antisocial policies minimise expected free energy more effectively than prosocial alternatives.

3.4. *The Extended Niche: Social Inference and Shared Generative Models*

Friston and colleagues have extended Active Inference to multi-agent settings, where individuals synchronise their generative models through shared attention, cultural norms, and institutional structures (Constant et al., 2019). The legal system itself can be understood as a “shared prior”—a collectively maintained expectation structure that enables prosocial behaviour by making the consequences of defection predictable. When an individual’s niche (family, peer group, neighbourhood) operates with generative models that are fundamentally misaligned with the broader social prior (the law), the individual faces a precision conflict: conforming to the local niche minimises immediate free energy but violates the societal prior, while conforming to law increases local unpredictability and social exclusion. The system resolves this conflict in favour of whichever prior carries higher precision—which, for someone embedded in an adversarial niche since childhood, is almost invariably the local one.

This mechanism provides the computational substrate for what Sampson and Laub (1993) describe as the role of informal social control in desistance. A “turning point”—a cohesive marriage, stable employment—works precisely because it changes the precision landscape: it provides a new, reliable source of prosocial prediction errors that, over time, can compete with the precision of existing criminal priors. The age-graded theory’s observation that turning points do not work for all individuals (particularly those with high psychopathy traits; McCuish et al., 2023) is predicted by the present model: when the precision of antisocial priors is sufficiently extreme—as in Moffitt’s LCP group—the prediction errors generated by a new social bond may be insufficient to drive model updating, because they are treated as noise relative to the dominant prior.

4. Historical Context: Distinguishing Active Inference from Biological Determinism

Any proposal to apply neuroscience to criminology must address the discipline’s traumatic history with biological explanations of criminal behaviour. This section is not a rhetorical concession but a substantive clarification of why the Active Inference framework is fundamentally incompatible with historical biological determinism.

This historical context is included solely to delineate the present framework from deterministic and discriminatory misuse, and to motivate explicit safeguards against person-level classification.

4.1. *The Shadow of Lombroso and the National Socialist Era*

Cesare Lombroso’s *L’Uomo Delinquente* (1876) inaugurated a tradition of biological criminology that sought to identify innate, immutable characteristics of the “born criminal.” This tradition reached its catastrophic conclusion in National Socialist Germany, where biological classification served as the scientific veneer for forced sterilisation, “asocial” detention, and murder. The legitimate revulsion against this history has produced a deep and understandable resistance within criminology—particularly in German-speaking scholarship—to any approach that invokes neurobiology in the context of criminal behaviour. This resistance is not irrational. It reflects a well-calibrated prior, shaped by historical evidence, that biological explanations of crime tend toward exclusion and dehumanisation. The present paper takes this prior seriously and argues that it must be updated—not by dismissing the historical evidence, but by demonstrating that Active Inference generates fundamentally different predictions.

4.2. *Why Active Inference Is Not Biological Determinism*

The distinction is structural, not rhetorical, and rests on three principled arguments. First, plasticity rather than fixity: Lombroso posited static defects encoded in the organism’s anatomy.

Active Inference describes a dynamic, continuously updating system in which everything that has been learned through value-learning can, in principle, be unlearned through new precision-weighted experience. The model's central therapeutic prediction is not "remove the defective individual" but "modify the inferential environment." Second, environment rather than endowment: the NS-era biologised the offender as inherently inferior. Active Inference biologises the interaction between organism and environment. The claim is not "this brain is defective" but "this brain has adapted optimally to a pathological niche." The pathology lies in the niche, not the neuron. This provides the neurobiological mechanism for the sociological observation that structural disadvantage produces crime—it specifies how poverty enters the brain. Third, subjectivity rather than objectification: perhaps most critically, Active Inference takes the subjective perspective of the offender seriously for the first time in a mathematically rigorous way. We do not ask "What is wrong with this person?" but "How must the world have appeared to this person's generative model for this action to have been the most coherent available policy?" This is not reduction; it is radical empathy formalised as mathematics.

4.3. *The Diagnostic Criterion: What Does the Model Predict?*

The definitive test of whether a theoretical framework is compatible with deterministic ideology is its predictions. Biological determinism predicts immutability and legitimises person-level identification and exclusion. Active Inference predicts the opposite: that behaviour changes when inferential conditions change, and that the most effective interventions target the environment, not the individual. Active Inference predicts the opposite: that criminal behaviour changes when inferential conditions change, and that the most effective interventions target the environment, not the individual. A framework that prescribes niche modification as its primary intervention is structurally incompatible with eugenic thinking. This is not a matter of framing or emphasis; it follows directly from the mathematics. If behaviour is a function of precision-weighted inference over a generative model shaped by the niche, then changing the niche changes the behaviour—and the claim that any individual is "biologically destined" to offend becomes incoherent within the framework itself.

5. The Precision-Weighting Dysregulation Model of Criminal Behaviour

We now present the central contribution of this paper: a model specifying three distinct pathomechanisms through which precision-weighting dysregulation produces criminal behaviour. Each pathway maps onto an established criminological construct and generates distinct empirical predictions (see Figure 1). Crucially, these pathways are not categorical types but regions in a continuous precision-profile space: the same computational architecture produces a spectrum of behavioural outcomes depending on the degree and pattern of precision dysregulation. The three pathways are not mutually exclusive; an individual may exhibit features of more than one, and the boundaries between them are probabilistic rather than categorical—reflecting the continuous nature of precision weighting itself.

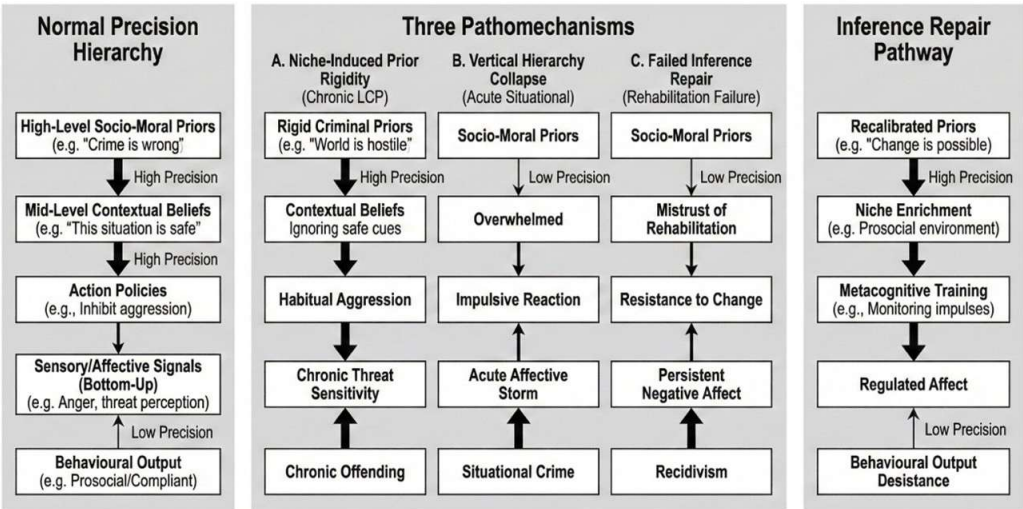


Figure 1. The Precision-Weighting Dysregulation Model of Criminal Behaviour. Left panel: Normal precision hierarchy with effective top-down modulation. Centre panels: Three pathomechanisms—(A) Niche-induced prior rigidity (chronic developmental, corresponding to Moffitt’s LCP trajectory), (B) Vertical hierarchy collapse (acute situational, corresponding to crimes of passion), (C) Failed inference repair (rehabilitation failure). Right panel: The inference repair pathway showing niche enrichment, precision recalibration, and metacognitive training as intervention targets.

5.1. Pathway 1: Niche-Induced Prior Rigidity (Life-Course-Persistent Delinquency)

The first pathway corresponds most closely to Moffitt’s (1993) life-course-persistent trajectory and addresses the developmental origins of chronic antisocial behaviour. Adolescence represents a critical period of generative model restructuring in which the brain must replace childhood priors (safety through parental attachment) with autonomous adult priors (safety through self-regulation and social reciprocity). This transition requires environmental conditions that reward exploration—the epistemically driven behaviour that Active Inference identifies as essential for model updating (Friston et al., 2017). For a juvenile in an adverse niche—characterised by violence, unpredictability, and the absence of reliable prosocial reinforcement—the exploratory phase is curtailed. The dopaminergic system learns that dominance, territorial control, and peer-group loyalty generate the most reliable prediction-error signals. These become high-precision priors, rigidly maintained because any deviation increases free energy (uncertainty, danger). The resulting behaviour pattern—what the criminal justice system labels “juvenile delinquency”—is, from the system’s perspective, an optimal adaptation to a hostile niche.

Moffitt’s original taxonomy proposed that LCP offenders differ from AL offenders in having neuropsychological deficits that interact cumulatively with criminogenic environments across development, culminating in a pathological personality. The present model refines this account by specifying the computational mechanism: neuropsychological vulnerability (whether inherited or acquired through prenatal exposure, perinatal insult, or early-childhood adversity) manifests as altered precision-weighting architecture—specifically, as a reduced capacity to assign high precision to prefrontal, executive-level priors relative to limbic, affective-level signals. Carlisi et al.’s (2020) finding that LCP individuals show widespread reductions in cortical surface area provides the neuroanatomical correlate: less cortical substrate means less computational capacity for high-level precision modulation.

Conventional incarceration compounds the problem rather than addressing it. Removing the juvenile from the original niche and placing them in an environment of even greater unpredictability and violence (the detention facility) does not challenge existing priors—it confirms them. The system receives further evidence that the world is hostile and that aggressive policies are necessary. From an Active

Inference perspective, juvenile detention functions, mathematically, as a training programme for criminal inference: it provides precisely the kind of high-precision evidence (consistent threat, unreliable cooperation) that strengthens rather than weakens the antisocial generative model. This interpretation is supported by the consistent finding that incarceration increases, rather than decreases, recidivism risk in juvenile populations (Petrosino, Turpin-Petrosino, and Guckenburg, 2010).

5.2. Pathway 2: Vertical Hierarchy Collapse (*Crimes of Passion*)

The second pathway addresses crimes of passion (Affekttaten)—offences in which the pathomechanism is temporally discrete rather than developmental. Under normal conditions, the hierarchical generative model maintains a stable priority structure: high-level priors (moral norms, legal consequences, identity as a non-violent person) carry sufficient precision to modulate lower-level affective signals (anger, humiliation, fear). Steinberg's (2008) dual-systems model describes a structurally analogous phenomenon in adolescence, where the socioemotional system's activation can overwhelm the cognitive control system. The present pathway generalises this mechanism to adults and specifies it in precision-theoretic terms.

In an acute provocation—particularly one that activates deep identity-related prediction errors (betrayal by an intimate partner, public humiliation, existential threat)—the precision of the affective signal can escalate to a level that functionally overwhelms the top-down hierarchy. This is the “vertical collapse”: the system enters a state where only the lowest-level policy (“eliminate the source of the prediction error”) is computed, because the prefrontal circuits lack the computational resources—under conditions of extreme free energy—to simulate alternatives. In formal terms, the precision of the affective prediction error (π_{affect}) surges to a value that renders the top-down precision (π_{PFC}) negligible in the softmax policy selection, even though the prefrontal circuits are structurally intact. This is not a chronic condition but a transient precision catastrophe, analogous to what Krupnik (2020) has described as the continuum between adaptive stress and trauma.

This mechanism maps directly onto the legal concept of diminished capacity (§ 20/21 StGB in German law; analogous provisions exist in most jurisdictions). “Capacity for insight” (Einsichtsfähigkeit) corresponds to the system's ability to maintain a generative model complex enough to represent legal and moral consequences. “Capacity for behavioural control” (Steuerungsfähigkeit) corresponds to the prefrontal system's ability to assign sufficient precision to alternative policies to override the dominant impulse. When either capacity is compromised by a transient precision catastrophe, the system meets the formal criteria for diminished responsibility—and the present model provides a more operationalisable criterion than the traditional, largely phenomenological assessment.

5.3. Pathway 3: Failed Inference Repair (*Recidivism as Systemic Prediction*)

The third pathway addresses the most consequential implication of the model: the mechanisms that maintain criminal behaviour over time and that cause conventional rehabilitation to fail. If criminal behaviour is maintained by high-precision antisocial priors, then effective rehabilitation requires systematic model updating: the creation of conditions under which the brain receives evidence of sufficient precision to revise its generative model. This is, in Active Inference terms, the process of “inference repair.”

Standard reintegration programmes often fail this criterion. Post-release environments frequently reproduce the original adverse niche—unstable housing, unemployment, social exclusion—while offering no structured source of high-precision prosocial prediction errors. The brain, confronted with evidence that confirms its existing model (“The world is hostile; cooperation is unreliable”), rationally maintains its criminal priors. Sampson and Laub's (1993, 2003) turning-point theory identifies the conditions under which desistance occurs—cohesive marriage, stable employment, meaningful social bonds—but Active Inference explains why these turning points work when they work: they alter the precision landscape by providing consistent, reliable evidence that prosocial predictions are correct. Conversely, the model predicts that turning points will fail when

the precision of existing antisocial priors is too high to be overcome by the new evidence—a prediction consistent with McCuish et al.'s (2023) finding that individuals with elevated psychopathy traits show attenuated responses to informal social control.

For the purpose of testing the model, the intervention arm is operationalised as three components targeting distinct precision-control parameters: (a) niche enrichment—creating a post-release environment with high predictability and consistent prosocial reinforcement, thereby reducing background free energy and making the system more receptive to new evidence; (b) precision recalibration—using physiological training (HRV biofeedback, interoceptive awareness, breathwork) to restore autonomic flexibility, which is the biological substrate of precision modulation (the capacity to flexibly up- and down-regulate the gain of different prediction error signals); and (c) metacognitive training—explicitly training the capacity to simulate alternative action policies, thereby increasing the depth and flexibility of the generative model's policy space. This tripartite approach is not a wishful programmatic statement but a direct derivation from the formal properties of the model: each component targets a specific parameter in the precision-weighting architecture.

The theoretical justification for precision recalibration through HRV biofeedback and breathing protocols requires explicit articulation within the Active Inference framework. The bridge rests on four linked propositions. First, interoceptive inference: Seth (2013) and Seth and Friston (2016) established that bodily states are regulated by autonomic reflexes that are enslaved by descending predictions from deep generative models. The precision of interoceptive signals determines how strongly bodily states influence the generative model. Second, HRV as index of precision modulation capacity: heart rate variability (especially HF-HRV and RMSSD) reflects parasympathetic cardiac modulation via the vagus nerve. Low HRV indexes rigid precision weighting—the system cannot flexibly adjust gain control. High HRV indexes flexible precision weighting—the system can shift signal weighting context-dependently. Third, resonance frequency breathing (~6 breaths per minute) produces maximal respiratory sinus arrhythmia, stimulates vagal afferents, and increases parasympathetic dominance. A systematic review of 77 HRV biofeedback studies (Wareing, Readman, Longo, Linkenauger and Crawford, 2024) found that HRV biofeedback improves interoception, with the authors proposing a three-stage model explicitly grounded in principles of interoceptive inference and predictive coding. Owens et al. (2018), with Friston as coauthor, provided empirical support for interoceptive inference in patients with autonomic dysfunction. Fourth, and most directly: Kim, Esteves, Cerritelli, and Friston (2022) describe how body-based interventions change priors: “To rectify aberrant active inference processes, we should change the ‘Rule’ of abduction, or the ‘prior beliefs’ entailed by a patient’s generative model.” The method involves weakening prior precision, then installing new priors through sensory experience. Limanowski and Friston (2020) formalise this as “self-flattening”—the attenuation of deeply anchored priors as a prerequisite for model updating.

The complete mechanism is therefore: HRV biofeedback → increased vagal tone → improved autonomic flexibility → increased capacity to modulate precision of interoceptive signals context-dependently → the system can now (a) downregulate precision of threat-related signals that were chronically elevated, (b) upregulate precision of prosocial feedback that was previously classified as noise, and (c) restore the gain-control flexibility necessary for generative model updating. Breathing protocols operate via the same vagal mechanism but additionally train CO₂ tolerance, which has consequences at two distinct levels. At the interoceptive level, increased tolerance for elevated CO₂ corresponds to increased tolerance for interoceptive prediction errors without panic or fight-flight activation, effectively expanding the policy space available for non-criminal inference (see Figure 2 for an overview of precision dynamics in normal vs. delinquent inference). At the metabolic level, an auxiliary hypothesis deserves explicit statement (testable independently of the main model): chronic hyperventilation—a hallmark of sustained sympathetic arousal in adverse niches—drives arterial CO₂ below physiological optimum, which via the Bohr effect reduces haemoglobin's oxygen release to tissues, including brain tissue. The resulting subclinical cerebral hypoperfusion may compromise the prefrontal circuits that implement top-down precision control (Section 3.1). Training CO₂

tolerance normalises arterial CO₂, restores oxygen delivery via the Bohr effect, and thereby improves the metabolic substrate on which the entire precision-weighting hierarchy depends. This auxiliary hypothesis generates its own testable prediction: end-tidal CO₂ measurement combined with near-infrared spectroscopy (NIRS) of prefrontal oxygenation during cognitive control tasks should reveal lower CO₂ and lower prefrontal oxygenation in chronic offenders compared to controls, with normalisation following breathing protocol training. In Active Inference terms, this is not merely a local interoceptive intervention but an improvement of the system’s global capacity for free energy minimisation: better-oxygenated neural tissue can sustain more complex generative models with greater temporal depth.

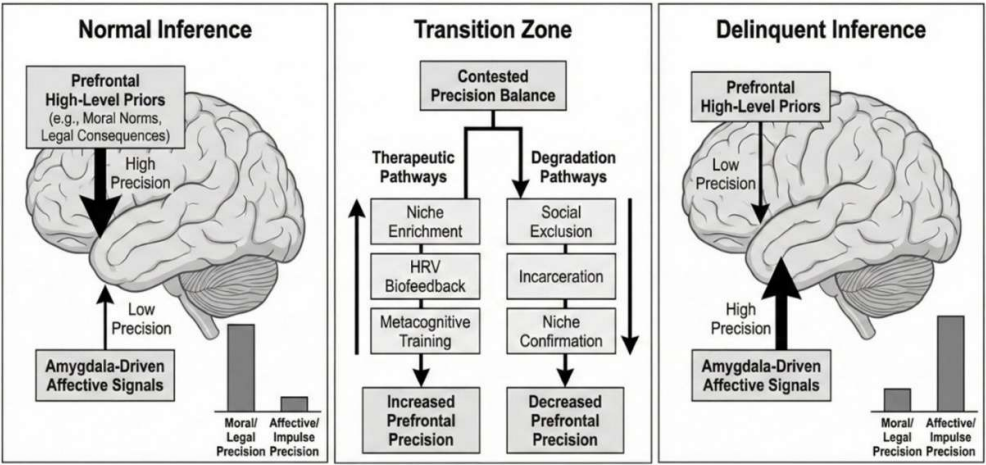


Figure 2. Precision Weighting in Normal vs. Delinquent Inference. Left panel: In the normally calibrated system, prefrontal high-level priors carry sufficient precision to modulate amygdala-driven affective signals. Right panel: In the delinquent system, precision allocation is inverted— affective and niche-specific priors dominate while prefrontal priors are functionally silenced. Centre: The transition zone with therapeutic pathways (niche enrichment, HRV biofeedback, metacognitive training) that increase prefrontal precision, and degradation pathways (social exclusion, incarceration, niche confirmation) that further reduce it.

5.4. From Construct to Proxy: Operational Indicators of Precision Modulation

Before specifying testable predictions, we must address a methodological challenge: the core constructs of the model—precision weights, prior rigidity, generative model complexity—are not directly observable. Each prediction therefore relies on operational proxies. The following mapping makes this relationship explicit: (a) Precision of interoceptive signals: putative substrate is vagal afferent gain modulated by the central autonomic network; measurement proxy is heart rate variability (RMSSD, HF-HRV); status: indirect but well-validated autonomic index (Owens, Allen, Ondobaka and Friston, 2018). (b) Policy precision (β): putative substrate is dopaminergic modulation of action selection in the basal ganglia; measurement proxy is computational modelling of reversal-learning parameters (learning rate, inverse temperature); status: indirect, requires model fitting. (c) Prior rigidity: putative substrate is reduced prefrontal top-down effective connectivity; measurement proxy is dynamic causal modelling (DCM) of PFC–amygdala coupling under provocation; status: indirect, model-dependent. (d) Generative model updating: putative substrate is synaptic plasticity in prefrontal–limbic circuits; measurement proxy is longitudinal change in belief-updating parameters and recidivism as a distal behavioural outcome; status: indirect, requires longitudinal design. None of these proxies constitutes a direct readout of precision weighting. However, the combination of convergent proxies across autonomic, computational, neuroimaging, and behavioural levels constitutes a strong test: the model predicts specific covariation patterns across all four measurement domains simultaneously, and failure of any single covariation pattern would constrain or disconfirm the hypothesis.

6. Testable Predictions and Research Agenda

To enable empirical evaluation of the precision-weighting dysregulation hypothesis, we specify three falsifiable predictions with corresponding study protocols. Each prediction targets a different pathway of the model and employs multimodal measurement to triangulate the precision-weighting dynamics at stake. The predictions are designed to be jointly sufficient for a strong test of the model: confirming all three would constitute substantial support; disconfirming any one would require revision of the corresponding pathway. Neuroimaging methods will conform to COBIDAS-MRI recommendations; physiological measures will follow Task Force standards for HRV analysis; behavioural paradigms will be preregistered on OSF; and observational reporting will follow STROBE guidelines. Site-specific IRB approval will precede any data collection. Given that all three study protocols involve vulnerable populations (juvenile offenders, incarcerated adults), heightened ethical scrutiny regarding informed consent, coercion, and data protection is mandatory; a detailed ethics and governance plan is provided in Section 7.6.

6.1. Prediction 1: Shifted Value-Learning in Juvenile Offenders

Formal derivation from the model: Precision modulation requires autonomic flexibility (Section 3). Autonomic flexibility is indexed by HRV. Chronic niche adversity reduces autonomic flexibility through sustained sympathetic dominance and vagal withdrawal. Therefore, the model predicts: juvenile offenders with high ACE scores will show reduced HRV, and this reduction will correlate with the magnitude of the value-learning shift from prosocial to dominance-based prediction errors.

The first prediction targets Pathway 1 (niche-induced prior rigidity) and tests whether juvenile offenders show the computational signature of precision-weighting dysregulation in their dopaminergic value-learning system. Specifically, we predict that juvenile offenders aged 14 to 18 with high ACE scores show, compared to age- and socioeconomic-status-matched controls: (a) reduced striatal BOLD response to prosocial reward prediction errors (cooperation, trust reciprocation) but enhanced response to dominance-based reward prediction errors (status, territorial gain) in an adapted Dictator/Ultimatum game paradigm; (b) reduced vagal flexibility (RMSSD, SDNN) at baseline and under social stress, reflecting a chronic impairment of the autonomic precision-modulation substrate; (c) a positive dose-response relationship between ACE score (measured via the Childhood Trauma Questionnaire) and the magnitude of the value-learning shift. This prediction is consistent with Moffitt's (1993) hypothesis that neuropsychological deficits interact with criminogenic environments and provides a mechanistic explanation for why LCP offenders show the specific pattern of reduced verbal IQ, impaired executive function, and elevated reward sensitivity documented by the Dunedin longitudinal study (Moffitt et al., 2002).

Design: Cross-sectional comparison; juvenile offenders ($n \geq 50$, power analysis based on medium effect size $d = 0.6$, $\alpha = 0.05$, power = 0.80) vs. matched controls ($n \geq 50$), stratified by ACE score tertiles. fMRI during a modified social reward task with prosocial and dominance conditions (randomised block design, 2 runs of 12 minutes). Concurrent HRV recording (5-minute resting baseline, task, 5-minute recovery). Behavioural measures: choice patterns, reaction times, and computational modelling of precision parameters using Active Inference models (cf. Smith et al., 2020). Bayesian hierarchical modelling will be used to estimate individual-level precision parameters and their relationship to ACE scores.

Primary endpoints: Striatal BOLD contrast (prosocial vs. dominance prediction errors); RMSSD at rest and recovery; ACE-score correlation with shift magnitude. Secondary endpoints include computational model parameters (policy precision, learning rate) and behavioural choice patterns.

Covariates: Age, sex, IQ estimate (WASI-II), medication status, substance use, comorbid psychiatric diagnoses (MINI-KID), time since last offence, current detention status.

Potential disconfirming evidence: No difference in striatal response patterns between groups; no correlation between ACE scores and value-learning shift; no HRV differences. Any of these null findings would necessitate revision of Pathway 1 as specified.

Feasibility considerations: Recruitment of juvenile offenders for neuroimaging research requires partnership with juvenile justice institutions and careful IRB review. fMRI in this population is feasible but demands age-appropriate paradigm design, parental/guardian consent (where applicable), and clear protocols for incidental findings. Prior studies have successfully conducted fMRI with juvenile offenders (Fairchild et al., 2013).

6.2. Prediction 2: PFC-Amygdala Decoupling in Crimes of Passion

Formal derivation: Section 3.1 established that the PFC exerts top-down precision control over the amygdala. Section 5.2 specified that in acute provocation, affective signal precision can overwhelm this hierarchical control. Therefore, the model predicts: perpetrators of crimes of passion will show reduced PFC→amygdala effective connectivity compared to both premeditated offenders (whose hierarchy remained intact during the offence) and non-offending controls.

The second prediction targets Pathway 2 (vertical hierarchy collapse) and tests whether perpetrators of crimes of passion show the neuroimaging and autonomic signature of a transient precision catastrophe. Specifically, we predict that adults convicted of crimes of passion show, compared to adults convicted of premeditated offences and non-offending controls: (a) reduced task-dependent functional connectivity between vmPFC and amygdala during a social provocation paradigm (insult, humiliation, betrayal scenarios); (b) characteristic pupillometric signatures preceding simulated escalation decisions—specifically, increased pupil dilation (reflecting LC-NE surge and precision reallocation) followed by reduced pupillary light reflex amplitude (reflecting loss of prefrontal modulation); (c) correlation between PFC-amygdala connectivity strength and forensic expert ratings of “control capacity” (Steuerungsfähigkeit).

Design: Cross-sectional three-group comparison; crimes-of-passion group ($n \geq 40$) vs. premeditated-offence group ($n \geq 40$) vs. non-offending controls ($n \geq 40$). fMRI with dynamic causal modelling (DCM) during social provocation task (validated vignette-based paradigm with escalation decision points). Concurrent infrared pupillometry. Forensic expert ratings of control capacity (blinded to neuroimaging results). The premeditated-offence group serves as a critical control: these individuals committed serious offences but with intact hierarchical control (the hierarchy was used in service of antisocial goals rather than overwhelmed by affective signals).

Primary endpoints: DCM effective connectivity parameters (PFC→amygdala); pupil dilation slope in pre-decision window; correlation with forensic ratings. Secondary endpoints include self-reported emotional intensity during provocation scenarios.

Covariates: Time since offence, trauma history (CTQ), medication, substance use, personality disorder diagnosis (PCL-R for psychopathy traits), age, sex.

Potential disconfirming evidence: No group differences in PFC-amygdala connectivity; no characteristic pupillometric patterns; no correlation with forensic expert ratings. Disconfirmation would require revision of the vertical collapse mechanism.

6.3. Prediction 3: Inference-Based Rehabilitation Is Predicted to Outperform Standard Reintegration

Formal derivation: Section 5.3 established that recidivism results from failed inference repair—the system’s generative model was never updated because post-release environments reproduced the original adverse niche. Therefore, the model predicts: an intervention that specifically targets the inferential architecture (IRP) will outperform standard reintegration (TAU) because it addresses the generative model itself, not merely the policies it produces. Furthermore, Friston’s own work on prior modulation (Kim et al., 2022; Limanowski and Friston, 2020) provides direct support for the mechanism: body-based interventions can weaken overprecise priors and create conditions for model updating.

The third prediction targets Pathway 3 (failed inference repair) and constitutes the most consequential test of the model, because it directly tests the therapeutic derivation of the precision-weighting hypothesis. We predict that incarcerated offenders (mixed offence types) randomised to a 12-month “Inference Repair Programme” (IRP) show, compared to standard reintegration (treatment

as usual, TAU): (a) greater increase in vagal tone (RMSSD, SDNN) from baseline to 6 and 12 months, reflecting restoration of autonomic precision-modulation capacity; (b) improved performance on belief-updating tasks (reduced “prior stickiness” in a probabilistic reversal learning paradigm), reflecting increased flexibility of the generative model; (c) reduced recidivism rates at 2-year and 5-year follow-up. We further predict that ACE scores moderate the treatment effect, with higher ACE scores predicting greater benefit from IRP—because the IRP specifically targets the precision-weighting dysregulation that ACEs produce.

Design: Multisite randomised controlled trial; IRP ($n \geq 100$) vs. TAU ($n \geq 100$), stratified by offence type and ACE score tertile (total $N \geq 200$; power analysis based on expected absolute risk reduction of 10 percentage points in 2-year reconviction rate (from 50% to 40%), $\alpha = 0.05$, power = 0.80, with survival analysis (Cox proportional hazards, expected HR = 0.70) as primary analytic framework, $\alpha = 0.05$, power = 0.80, accounting for 20% attrition). HRV assessment at baseline, 6 months, 12 months. Behavioural testing (probabilistic reversal learning with computational modelling) at baseline and 12 months. Recidivism tracking via official records at 2 and 5 years post-release. All randomisation and primary analyses will be pre-registered on OSF; analysis plan will include both frequentist and Bayesian inference to quantify evidence for and against the null hypothesis.

IRP components: (a) Niche enrichment: structured housing, stable employment or vocational training, consistent relational support from trained mentors (creating a predictable social environment that reduces background free energy). (b) Precision recalibration: 3× weekly HRV biofeedback sessions (20 min), daily breathwork protocol, interoceptive awareness training—targeting the autonomic substrate of flexible precision modulation. (c) Metacognitive training: weekly 90-minute group sessions focused on recognising precision shifts (“What am I attending to and why?”), simulating alternative action policies through structured perspective-taking exercises, and practising social inference (“How might the other person’s generative model represent this situation?”). Each component targets a specific parameter in the precision-weighting architecture: niche enrichment reduces the prior’s environmental confirmation, precision recalibration restores the autonomic substrate for flexible gain control, and metacognitive training increases the policy space available for inference.

Primary endpoints: RMSSD change from baseline; reversal-learning computational parameters (prior precision, learning rate); recidivism rate (binary: reconviction within 2/5 years). Secondary endpoints include self-reported wellbeing, employment status, and social integration measures.

Covariates: Offence severity (standardised), sentence length, age, sex, psychiatric comorbidity, substance use history, ACE score, programme adherence (attendance rate), site.

Potential disconfirming evidence: No HRV improvement; no change in belief-updating parameters; no recidivism reduction; no moderation by ACE score. Any of these null findings would require substantive revision of Pathway 3. If HRV and belief-updating improve but recidivism does not, this would suggest that the model correctly identifies the mechanism but that the intervention intensity or duration is insufficient—requiring a modified dose-finding study.

7. Discussion

7.1. Integration with Established Criminological Frameworks

The precision-weighting dysregulation hypothesis does not seek to replace existing criminological theories but to provide the computational mechanism that connects them. Moffitt’s (1993) developmental taxonomy identifies who is at risk and when: LCP offenders show early-onset antisocial behaviour driven by the cumulative interaction of neuropsychological vulnerability and criminogenic environments. The present model specifies how: through chronic miscalibration of the precision profile across the generative model hierarchy, such that antisocial policies carry systematically higher precision than prosocial alternatives. Steinberg’s (2008, 2010) dual-systems model identifies what develops differently in adolescence: the temporal gap between reward

sensitivity and cognitive control. Active Inference formalises this gap as a transient precision imbalance—the socioemotional system’s precision rises before the prefrontal control system develops the capacity to counterbalance it—and extends the mechanism beyond adolescence to explain adult crimes of passion. Sampson and Laub’s (1993, 2003) age-graded theory identifies what enables desistance: turning points that strengthen informal social control. The present model explains why turning points work (they change the precision landscape) and why they sometimes fail (when existing prior precision is too high to be overridden by new evidence).

This integrative capacity is not accidental. It follows from the generality of the free energy principle: because all adaptive behaviour can be described as precision-weighted inference, the same formalism accommodates developmental trajectories (Moffitt), neuromaturational dynamics (Steinberg), social-structural influences (Sampson and Laub), and individual acts of violence—each as a specific configuration of the precision profile within a universal computational architecture.

7.2. Implications for Criminal Responsibility

The framework suggests a reconceptualisation of criminal responsibility not as a binary (guilty/not guilty) determined by an opaque notion of “free will,” but as a graded assessment of inference capacity—the system’s ability to simulate alternative action policies with sufficient precision. This aligns with existing legal provisions for diminished responsibility (e.g., § 20/21 StGB) but provides a more operationalisable criterion. “Capacity for insight” becomes: “Could the generative model represent the legal and moral consequences of the action?” “Capacity for control” becomes: “Could the prefrontal system assign sufficient precision to alternative policies to override the dominant impulse?” This does not dissolve responsibility. On the contrary, it provides a more precise foundation for assessing it—and, crucially, for prescribing rehabilitation that targets the specific inferential deficit identified. It is essential to emphasise that this reconceptualisation must not be used to establish “biological risk pools” within the justice system. The model identifies inferential conditions, not fixed individual traits; it prescribes environmental intervention, not biological classification. Any forensic application would require extensive interdisciplinary validation involving criminologists, neuroscientists, legal scholars, and ethicists before implementation could be considered.

7.3. Connection to Neuroimmunological Research

The role of adverse childhood experiences as a common root of precision-weighting dysregulation connects the present hypothesis to recent work on ACE-related neuroimmunological conditions. The LC Norepinephrine Depletion Hypothesis of ME/CFS (Hemmerich, 2025) proposes that ACEs programme the locus coeruleus into a maladaptive high-tonic/low-phasic firing mode, leading to vesicular NE depletion and systemic dysautonomia. This suggests that ACEs, depending on genetic vulnerability, niche characteristics, and the specific neural systems most affected, may lead to divergent outcomes—somatic (ME/CFS), psychiatric (PTSD, depression), or behavioural (criminal conduct)—through a common upstream mechanism: chronic stress-induced reprogramming of precision-weighting architecture. The shared computational framework opens the possibility of cross-domain biomarker research: if HRV, pupillometry, and belief-updating paradigms index precision-weighting function in both ME/CFS and criminal populations, the same measurement battery could serve as a transdiagnostic assessment of ACE-related dysregulation.

7.4. Structural Determinants and Population-Level Considerations

A significant limitation of the model as presented is its primary focus on individual-level mechanisms. Criminal behaviour occurs within structural contexts—institutional racism, economic inequality, differential policing, inadequate social safety nets—that shape niches at the population level. Wilson’s (1987) foundational analysis of concentrated urban poverty demonstrated that the disappearance of manufacturing employment from inner-city neighbourhoods created environments

characterised by precisely the features that the present model identifies as niche-level drivers of precision-weighting dysregulation: pervasive unpredictability, absence of reliable prosocial reinforcement, and dominance-based social hierarchies as the primary available structure. Alexander's (2010) analysis of mass incarceration in the United States extends this insight to the criminal justice system itself: racially disproportionate policing, sentencing, and post-release surveillance create a caste-like structure that systematically reproduces the adverse niches from which criminal behaviour emerges—while simultaneously removing access to the turning points (stable employment, housing, civic participation) that Sampson and Laub (1993) identified as essential for desistance.

The present model accommodates these structural factors insofar as they operate through niche construction: structural disadvantage creates the adverse environments that programme maladaptive precision weighting. However, the model does not provide a detailed account of the political and economic forces that create and maintain these environments, and it must not be used to divert attention from the necessity of structural reform. A comprehensive criminological programme would combine the individual-level precision-weighting framework with population-level analyses of how policies—housing, education, employment, drug regulation, criminal justice reform—shape the distribution of niches across a society. The model's contribution to this programme is mechanistic specificity: it explains the pathway through which structural disadvantage enters the brain and produces behaviour, thereby identifying the points at which intervention—at any level, from individual rehabilitation to systemic policy reform—can be most effective. Without addressing the structural conditions that generate pathological niches at scale, individual-level interventions will remain palliative rather than curative.

7.5. Incarceration as Confirmation of the Criminogenic Prior

The argument of Section 5.1—that conventional incarceration compounds precision-weighting dysregulation rather than addressing it—warrants separate discussion because of its implications for criminal justice policy. Within the Active Inference framework, the prediction is mathematically necessary: any intervention that reproduces the adverse niche must confirm rather than update the generative model. The detention facility, characterised by environmental unpredictability, threat of violence, dominance hierarchies, and absence of reliable prosocial reinforcement, constitutes a criminogenic niche par excellence. It provides exactly the sensory evidence that the system's antisocial priors predict, thereby increasing their precision rather than challenging them.

Empirical evidence supports this formal prediction. Petrosino, Turpin-Petrosino, and Guckenburg (2010), in a Campbell Collaboration systematic review, found that formal system processing of juveniles (including incarceration) increased rather than decreased subsequent delinquency. International data underscore the scale of the problem: the United States incarcerates approximately 531–629 persons per 100,000 population—roughly 1.8 million individuals, representing approximately 25% of the world's prison population (World Prison Brief, 2024). Recidivism rates across jurisdictions remain persistently high. The Active Inference model provides a formal explanation for this failure: the system is being “trained” to perform criminal inference more precisely.

The policy implication is clear: effective intervention must change the inferential environment. This aligns with the growing evidence base for restorative justice, community-based sentences, and environmental interventions that modify the niche rather than extracting the individual from one adverse niche and placing them in another. The IRP proposed in Section 6.3 represents one operationalisation of this principle (see Figure 3 for a systems-level comparison of retributive vs. inference-repair cycles).

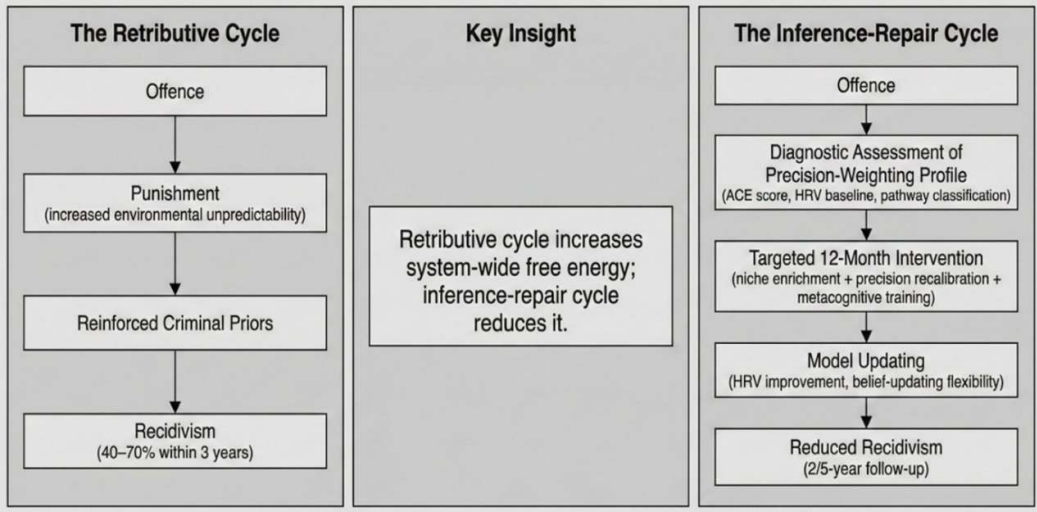


Figure 3. From Retribution to Inference Repair: A Systems-Level View. Left panel: The retributive cycle—offence → punishment (increased environmental unpredictability) → reinforced criminal priors → recidivism (40–70% within 3 years). Centre: Key insight—retributive cycle increases system-wide free energy; inference-repair cycle reduces it. Right panel: The inference-repair cycle—offence → diagnostic assessment of precision-weighting profile (research-level mechanism tracking, not individual risk classification) → targeted 12-month intervention → model updating → reduced recidivism.

7.6. Moral Injury as an Unaddressed Consequence of Criminogenic Precision Dysregulation

Section 2.3 introduced moral injury as an additional mechanism of self-harm generated by the criminal pathway. Here we develop the formal argument. Moral injury, originally described in combat veterans (Shay, 1994; Litz et al., 2009; Jinkerson, 2016) and recently examined through a predictive processing lens (Xanthios and Saffaran, 2023), describes the psychological damage resulting from actions that violate one’s own moral presuppositions. The present model offers a specific formalisation: moral injury is a chronically unresolvable prediction error at the identity level of the generative model.

The key insight is a hierarchy-level dissociation. A person operating under criminogenic precision-weighting dysregulation can minimise free energy efficiently at the policy level: criminal actions reduce immediate threat or increase immediate reward. Simultaneously, the same person generates chronic prediction error at the identity level, because the system’s highest-order model (“who I am”) predicts moral self-consistency that the policy-level actions violate. This dissociation—efficient policy execution coupled with chronic identity-level surprise—is possible because free energy is minimised locally at each level of the hierarchy, and what is optimal at one level can be catastrophic at another.

The temporal dynamics of moral injury mirror those of post-exertional malaise in ME/CFS (Hemmerich, 2025): the damage manifests with delay, without the person recognising the causal connection. The person who deceives, steals, or harms does not necessarily experience immediate guilt. The identity-level prediction error accumulates and manifests as chronic inner restlessness, hypervigilance directed inward, cynicism, relational incapacity, and progressive loss of meaning—symptoms that are frequently misclassified or undiagnosed because the temporal gap obscures the causal relationship.

Critically, no validated treatment protocol for moral injury exists in the criminal justice context. The existing literature addresses moral injury in military populations (Gray et al., 2012; Frankfurt and Frazier, 2016), but transfer to the criminological setting faces fundamental obstacles: the patient is typically in the justice system rather than the health system; acknowledging moral injury contradicts the self-protective dynamics of prison culture; and therapeutic relationships are structurally

compromised in coercive contexts. This therapeutic gap provides an additional medical rationale for the approach developed in this paper.

7.7. Epistemic Status and Limits of the Free Energy Principle

Any paper that applies the Free Energy Principle to a new domain must address its epistemic status transparently. Karl Friston is among the most cited neuroscientists worldwide, though citation metrics primarily reflect the widespread use of his methodological tool—Statistical Parametric Mapping (SPM)—rather than universal acceptance of his theoretical framework. The community that actively works with Active Inference remains relatively small.

The FEP as a mathematical principle is unfalsifiable—a point Friston himself has acknowledged. Like Hamilton's principle of least action in physics, it describes a formal constraint: systems that maintain a boundary with their environment behave as if they minimise variational free energy. This is not a flaw but the nature of a formal principle. Active Inference, by contrast, is a falsifiable process theory that specifies how particular systems implement free energy minimisation. The distinction is analogous to that between the second law of thermodynamics (a principle) and specific models of heat transfer (process theories).

Several substantive criticisms have been raised. Gershman (2019) argued that if the FEP reduces to Bayesian inference, it becomes indistinguishable from other asymptotically correct inference algorithms unless additionally constrained. Colombo and Wright (2017) demonstrated that the FEP provides explanatory pluralism rather than a single grand unified theory: the same neural system (e.g., dopamine) can be described by multiple irreducible models. More recently, concerns about post hoc explanatory flexibility have been articulated: the mathematical generality of the framework allows it to accommodate virtually any empirical outcome, raising questions about whether it can be meaningfully disconfirmed (Andrews, 2021; Williams, 2020).

The most comprehensive empirical review to date—Hodson, Smith, and Bhatt (2024)—assessed the evidence for both Predictive Coding and Active Inference. For Predictive Coding, existing evidence provides modest support, though some positive findings can also be explained by feedforward models. For Active Inference, most empirical studies have fitted models to behavioural data as a means of identifying individual or group differences—but the framework itself has rarely been the target of direct falsification attempts. The authors conclude that further empirical validation is required to test Active Inference's explanatory scope.

This assessment has direct implications for the present paper. We acknowledge that the precision-weighting dysregulation hypothesis inherits the epistemic limitations of its parent framework: it is a model that could, in principle, accommodate a wide range of outcomes. However, we counter this risk by specifying three falsifiable predictions (Section 6) that make concrete, quantitative claims testable with standard neuroimaging and psychophysiological methods. If juvenile offenders show no value-learning shift, if crimes-of-passion perpetrators show no PFC-amygdala decoupling, or if the IRP produces no advantage over standard reintegration, the specific application of Active Inference to criminal behaviour as proposed here would require substantive revision. The predictions therefore serve not only as a research agenda but as a safeguard against the very post hoc flexibility that critics of the FEP have identified.

7.8. Beyond Re-Normalisation: Inference Repair as Prevention and Rehabilitation

The currently dominant therapeutic approach in correctional settings is cognitive behavioural therapy (CBT). Meta-analyses consistently show moderate effects on recidivism rates—a meaningful improvement over no treatment. However, the Active Inference framework reveals a structural limitation of CBT that explains why its effects remain modest: CBT operates predominantly at the policy level. It provides alternative behavioural scripts—when the impulse to offend arises, the individual is trained to execute a different action sequence. In Active Inference terms, this is policy substitution: the prior preferences over action sequences (π) are modified, but the generative model

itself—the deep priors over niche structure, threat expectation, self-concept, and precision weighting—remains largely unchanged.

This policy-level intervention has three specific limitations. First, it does not address precision-weighting dysregulation: the core pathology—rigid, overprecise priors that classify prosocial signals as noise—persists. The person learns to act differently but not to perceive differently. Second, it does not address moral injury: the chronic prediction error at the identity level cannot be resolved through behavioural change alone, because past actions cannot be undone by future compliance. Third, CBT in coercive contexts (where programme completion is a prerequisite for release) is structurally compromised: the system rewards adaptive performance, not genuine model updating—which formally replicates the criminogenic logic of short-term reward maximisation.

The Inference Repair Programme proposed in this paper differs from standard rehabilitation in a way that the Active Inference framework makes formally precise. The target is not the policy level but the generative model itself. The goal is not re-normalisation—defined as behavioural compatibility with the prevailing social niche—but restoration of inferential flexibility: the capacity of the generative model to update itself in response to new evidence.

The formal distinction is as follows. Re-normalisation optimises policy preferences (π) such that they produce no conflict with societal norms. The success criterion is negative: absence of reoffending. The person “functions.” Inference repair optimises model evidence ($\log p(o|m)$)—the capacity of the entire generative model to represent the world accurately and flexibly. The success criterion is positive: restored capacity to weight prosocial signals as informative, to tolerate prediction errors without reflexive fight-flight, and to engage in active niche construction rather than passive niche compliance.

This distinction has a dual implication. As rehabilitation, inference repair addresses the generative model that produced criminal behaviour and the moral injury that criminal behaviour produced—neither of which is accessible to policy-level intervention alone. As prevention, inference repair strengthens the inferential flexibility whose absence makes criminogenic precision-weighting dysregulation possible in the first place. Prevention here does not mean teaching impulse control (a policy-level strategy) but restoring the autonomic substrate of flexible precision modulation so that the criminogenic inference chain does not become dominant.

The intervention components specified in Section 6.3 target these distinct levels: niche enrichment modifies the sensory environment (reducing background free energy); HRV biofeedback and breathing protocols restore the autonomic substrate of precision modulation (interoceptive level); and cognitive restructuring within an Active Inference framework targets the priors themselves (generative model level). Emerging evidence from mindfulness-based interventions in correctional settings (Shonin et al., 2013; Auty, Cope, and Liebling, 2017) is consistent with the prediction that prior-level interventions may add value beyond policy-level CBT, though the evidence base remains too thin for definitive conclusions. This represents an explicit research gap that the proposed studies aim to address.

7.9. Limitations

The model is subject to several important limitations that must be acknowledged transparently. First, no direct measurement of precision weighting in criminal populations currently exists; the model's neural predictions are inferred from adjacent literatures in psychosis research, trauma neuroscience, and computational decision-making. Until the proposed studies are conducted, the model remains a hypothesis, not an empirically validated framework. Second, the associations between ACEs and delinquency on which Pathway 1 depends are derived predominantly from cross-sectional and retrospective data; prospective longitudinal studies with neuroimaging, while conducted in some non-criminal populations (e.g., the Dunedin Study; Carlisi et al., 2020), are lacking specifically for the precision-weighting parameters proposed here. Third, the rehabilitation prediction (Prediction 3) requires a complex multisite RCT that has not been conducted, and the feasibility of implementing the IRP across diverse correctional settings is non-trivial. Fourth, the

model may overemphasise individual-level mechanisms at the expense of structural factors (institutional racism, economic policy, differential policing) that shape niches at a population level, as discussed in Section 7.4. Fifth, translating precision-weighting concepts into legally admissible forensic assessment requires extensive interdisciplinary validation that has not yet occurred. Sixth, the three pathways, while theoretically distinct, may be difficult to separate in practice, since many offenders exhibit features of more than one; sensitivity analyses and latent-class approaches will be needed to test the pathway boundaries empirically.

7.10. Ethics and Governance Plan for Proposed Research

ETHICAL SAFEGUARD: This framework must not be used for any individual-level predictive use (screening, profiling, triage, or risk scoring), nor for deterministic classification of persons. Its sole intended application is to inform evidence-based rehabilitation, prevention through environmental modification, and the scientific understanding of human behaviour. Any use that results in stigmatisation, discrimination, or restriction of individual rights represents a fundamental misapplication of the model.

All proposed studies involve vulnerable populations (juvenile offenders, incarcerated adults) and require the highest level of ethical scrutiny. The following governance principles are proposed as minimum safeguards for any future empirical work in vulnerable populations: (a) Informed consent must be obtained from all participants (and from parents/guardians for minors), with explicit assurance that participation is voluntary and that refusal will have no consequences for legal proceedings, parole decisions, or institutional treatment. (b) Data protection must comply with GDPR (or equivalent national standards) and must include anonymisation of neuroimaging and physiological data; no individually identifiable data may be shared with judicial authorities. (c) Incidental findings: a clear protocol for managing incidental neuroimaging findings must be established before data collection begins, including access to clinical follow-up. For example, if structural MRI reveals an unsuspected intracranial mass or vascular malformation in a juvenile offender participant, the protocol must specify: (i) immediate notification of the participant (and parent/guardian for minors) by a qualified clinician, not by research staff; (ii) referral to appropriate clinical services at no cost to the participant; (iii) clear documentation that the finding was incidental and unrelated to the research question; (iv) no communication of the finding to judicial authorities or correctional staff unless explicitly authorised by the participant. A neuroradiologist must review all structural scans within 48 hours of acquisition. (d) No predictive use: research findings must not be used for any individual-level prediction, screening, profiling, triage, or risk scoring. The model identifies population-level mechanisms, not individual-level predictions. (e) Independent ethics oversight: an independent data safety monitoring board (DSMB) with representation from criminology, forensic psychiatry, law, and ethics should oversee the RCT (Prediction 3). (f) Publication and transparency: all protocols, analysis plans, and data (anonymised) will be registered on OSF before data collection begins; negative results will be reported with equal rigour.

8. Conclusions

We have proposed a specific, testable hypothesis: criminal behaviour arises from systematic dysregulation of precision weighting within hierarchical generative models, shaped by the interaction between the individual's neural architecture and their social niche. Three pathomechanisms—niche-induced prior rigidity, vertical hierarchy collapse, and failed inference repair—generate three forensically distinguishable patterns that converge on a common computational endpoint: the inability to generate prosocial action policies with sufficient precision to compete with antisocial alternatives. The model integrates and extends the landmark contributions of Moffitt (1993), Steinberg (2008), and Sampson and Laub (1993) by providing the unified computational mechanism that connects developmental trajectories, neuromaturational dynamics, and social-structural influences within a single formal framework.

The framework generates three falsifiable predictions testable through neuroimaging, autonomic physiology, behavioural paradigms, and epidemiological follow-up. It further specifies the differential conditions under which the same upstream dysregulation produces criminal behaviour, depression, anxiety, or somatic outcomes—positioning the model as transdiagnostic rather than specific to criminality. From a medical perspective, the criminal pathway warrants particular attention because it is therapeutically underserved, generates additional self-harm through moral injury, and is addressed by institutional responses (incarceration) that formally reinforce the underlying pathology. The proposed Inference Repair Programme targets the generative model itself rather than merely substituting behavioural policies, aiming not at re-normalisation but at the restoration of inferential flexibility. Most importantly, the framework predicts that effective intervention must target the inferential environment—not the offender as a biologically deficient individual, but the niche as the sculptor of inference.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. No human participants or data are included in this manuscript. The detailed ethics and governance plan for the proposed studies is presented in Section 7.6. Site-specific OSF preregistration and local IRB/ethics committee approval will precede any data collection. Studies involving incarcerated populations require heightened ethical scrutiny regarding informed consent, coercion, and data protection.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable (conceptual hypothesis paper).

Use of AI tools: AI language tools were used for editorial assistance (clarity, formatting, and reference cross-checking). All scientific claims, interpretations, and citations were reviewed and approved by the author, who assumes full responsibility for the content.

Disclosure: This manuscript is a conceptual hypothesis intended solely to generate falsifiable research predictions within the Active Inference framework. It reports no human participants, no individual-level data, and no operational tool. The framework is not designed, validated, or suitable for individual-level prediction, screening, profiling, triage, risk scoring, or forensic classification of persons. Any attempt to use it for such purposes would be a misuse of the work and a violation of its stated scope. The sole intended use is mechanistic research: to inform study design in rehabilitation research and in prevention research via niche-level/environmental modification, not person-level classification.

Acknowledgments: The author thanks colleagues in criminology, forensic psychiatry, and the ME/CFS research community who provided critical feedback on early drafts.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACE	adverse childhood experience
AL	adolescence-limited
BOLD	blood-oxygen-level-dependent
COBIDAS	Committee on Best Practices in Data Analysis and Sharing
CTQ	Childhood Trauma Questionnaire
DCM	dynamic causal modelling
DSMB	Data Safety Monitoring Board
FEP	free energy principle
fMRI	functional magnetic resonance imaging

GDPR	General Data Protection Regulation
HRV	heart rate variability
IRB	Institutional Review Board
IRP	Inference Repair Programme
LC	locus coeruleus
LCP	life-course-persistent
mPFC	medial prefrontal cortex
NE	norepinephrine
OSF	Open Science Framework
PCL-R	Psychopathy Checklist—Revised
PFC	prefrontal cortex
RCT	randomised controlled trial
RMSSD	root mean square of successive differences
SDNN	standard deviation of normal-to-normal intervals
StGB	Strafgesetzbuch (German Criminal Code)
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
TAU	treatment as usual
vmPFC	ventromedial prefrontal cortex
VTA	ventral tegmental area
WASI-II	Wechsler Abbreviated Scale of Intelligence—Second Edition
IRP	Inference Repair Programme
MBCT	mindfulness-based cognitive therapy
MBSR	mindfulness-based stress reduction

Glossary for Interdisciplinary Readers

Noise (in Active Inference): Sensory input to which the system assigns low precision. Noise is not treated as an informative signal and does not lead to updating of the generative model. In the criminological context: a prosocial experience that the system interprets as an irrelevant fluctuation rather than a meaningful prediction error.

Moral Injury: Psychological damage resulting from acting against one’s own moral presuppositions. Formalised in the present framework as a chronically unresolvable prediction error at the identity level of the generative model.

Inference Repair: Therapeutic approach targeting the generative model itself (deep priors and precision weighting) rather than behavioural policies alone. Distinguished from re-normalisation in that its success criterion is restored inferential flexibility, not behavioural compliance.

Re-normalisation: Rehabilitation approach whose implicit success criterion is behavioural compatibility with prevailing social norms. Operates primarily at the policy level.

References

1. Adams, R.A., Stephan, K.E., Brown, H.R., Frith, C.D. and Friston, K.J. (2013) ‘The computational anatomy of psychosis’, *Frontiers in Psychiatry*, 4, p. 47.
2. Alexander, M. (2010) *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. New York: The New Press.
3. Andrews, M. (2021) ‘The math is not the territory: navigating the free energy principle’, *Biology & Philosophy*, 36, p. 30.

4. Auty, K.M., Cope, A. and Liebling, A. (2017) 'A systematic review and meta-analysis of yoga and mindfulness meditation in prison', *International Journal of Offender Therapy and Comparative Criminology*, 61(6), pp. 689–710.
5. Birn, R.M., Patriat, R., Phillips, M.L., Germain, A. and Herringa, R.J. (2014) 'Childhood maltreatment and combat posttraumatic stress differentially predict fear-related fronto-subcortical connectivity', *Depression and Anxiety*, 31(10), pp. 880–892.
6. Carlisi, C.O., Moffitt, T.E., Knodt, A.R., Harrington, H., Ireland, D., Melzer, T.R., Poulton, R., Ramrakha, S., Caspi, A., Hariri, A.R. and Viding, E. (2020) 'Associations between life-course-persistent antisocial behaviour and brain structure in a population-representative longitudinal birth cohort', *Lancet Psychiatry*, 7(3), pp. 245–253.
7. Colombo, M. and Wright, C. (2017) 'Explanatory pluralism: an unrewarding prediction error for free energy theorists', *Brain and Cognition*, 112, pp. 3–12.
8. Constant, A., Ramstead, M.J.D., Veissière, S.P.L. and Friston, K.J. (2019) 'Regimes of expectations: an active inference model of social conformity and decision making', *Frontiers in Psychology*, 10, p. 679.
9. Denison, S., Reed, C. and Xu, F. (2013) 'The emergence of probabilistic reasoning in very young infants: evidence from 4.5- and 6-month-olds', *Developmental Psychology*, 49(2), pp. 243–249.
10. Denison, S. and Xu, F. (2019) 'Infant statisticians: the origins of reasoning under uncertainty', *Perspectives on Psychological Science*, 14(4), pp. 499–516.
11. Fairchild, G., Hagan, C.C., Walsh, N.D., Passamonti, L., Calder, A.J. and Goodyer, I.M. (2013) 'Brain structure abnormalities in adolescent girls with conduct disorder', *Journal of Child Psychology and Psychiatry*, 54(1), pp. 86–95.
12. Fazel, S. and Wolf, A. (2015) 'A systematic review of criminal recidivism rates worldwide: current difficulties and recommendations for best practice', *PLoS ONE*, 10(6), e0130390.
13. Felitti, V.J., Anda, R.F., Nordenberg, D., et al. (1998) 'Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults', *American Journal of Preventive Medicine*, 14(4), pp. 245–258.
14. FitzGerald, T.H., Dolan, R.J. and Friston, K.J. (2015) 'Dopamine, reward learning, and active inference', *Frontiers in Computational Neuroscience*, 9, p. 136.
15. Frankfurt, S. and Frazier, P. (2016) 'A review of research on moral injury in combat veterans', *Military Psychology*, 28(5), pp. 318–330.
16. Friston, K.J. (1998) 'The disconnection hypothesis', *Schizophrenia Research*, 30(2), pp. 115–125.
17. Friston, K.J. (2010) 'The free-energy principle: a unified brain theory?', *Nature Reviews Neuroscience*, 11(2), pp. 127–138.
18. Friston, K.J. and Frith, C.D. (1995) 'Schizophrenia: a disconnection syndrome?', *Clinical Neuroscience*, 3(2), pp. 89–97.
19. Friston, K.J., FitzGerald, T., Rigoli, F., Schwartenbeck, P. and Pezzulo, G. (2017) 'Active inference: a process theory', *Neural Computation*, 29(1), pp. 1–49.
20. Gershman, S.J. (2019) 'What does the free energy principle tell us about the brain?', *arXiv preprint*, arXiv:1901.07945.
21. Gray, M.J., Schorr, Y., Nash, W., et al. (2012) 'Adaptive Disclosure: an open trial of a novel exposure-based intervention for service members with combat-related psychological stress injuries', *Behavior Therapy*, 43(2), pp. 407–415.
22. Harris, H.W. (2025) 'Active inference and psychodynamics: a novel integration with applications to depression and stress disorders', *Frontiers in Psychiatry*, 16, p. 1630858.
23. Hemmerich, F.H. (2025) 'The locus coeruleus norepinephrine depletion hypothesis of ME/CFS: a mechanistic model with testable predictions and multimodal study plans (protocol framework)', *Preprints*, 2025120166. doi: 10.20944/preprints202512.1660.v3.
24. Hodson, R., Smith, R. and Bhatt, M. (2024) 'The empirical status of predictive coding and active inference', *Neuroscience & Biobehavioral Reviews*, 157, 105473.
25. Jinkerson, J.D. (2016) 'Defining and assessing moral injury: a syndrome perspective', *Traumatology*, 22(2), pp. 122–130.

26. Kim, S., Esteves, J.E., Cerritelli, F. and Friston, K. (2022) 'An active inference account of touch and verbal communication in therapy', *Frontiers in Psychology*, 13, 828952.
27. Krupnik, V. (2020) 'Trauma or drama: a predictive processing perspective on the continuum of stress', *Frontiers in Psychology*, 11, p. 1248.
28. Limanowski, J. and Friston, K. (2020) 'Attenuating oneself: an active inference perspective on "selfless" experiences', *Philosophy and the Mind Sciences*, 1(1), pp. 1–16.
29. Litz, B.T., Stein, N., Delaney, E., et al. (2009) 'Moral injury and moral repair in war veterans: a preliminary model and intervention strategy', *Clinical Psychology Review*, 29(8), pp. 695–706.
30. Lombroso, C. (1876) *L'Uomo Delinquente*. Milan: Hoepli.
31. McCuish, E.C., Lussier, P. and Corrado, R.R. (2023) 'Twenty years in the making: revisiting Laub and Sampson's version of life-course criminology', *Journal of Criminal Justice*, 89, 102123.
32. Mischkowitz, R. (1994) 'Desistance from a delinquent way of life', in Weitekamp, E.G.M. and Kerner, H.-J. (eds) *Cross-National Longitudinal Research on Human Development and Criminal Behavior*. NATO ASI Series, vol. 76. Dordrecht: Kluwer Academic Publishers, pp. 303–327.
33. World Prison Brief (2024) United States of America. Institute for Crime & Justice Policy Research. Available at: <https://www.prisonstudies.org/country/united-states-america> (Accessed: 15 February 2026).
34. Moffitt, T.E. (1993) 'Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy', *Psychological Review*, 100(4), pp. 674–701.
35. Moffitt, T.E. (2018) 'Male antisocial behaviour in adolescence and beyond', *Nature Human Behaviour*, 2, pp. 177–186.
36. Moffitt, T.E., Caspi, A., Harrington, H. and Milne, B.J. (2002) 'Males on the life-course-persistent and adolescence-limited antisocial pathways: follow-up at age 26 years', *Development and Psychopathology*, 14(1), pp. 179–207.
37. Nagin, D.S. and Land, K.C. (1993) 'Age, criminal careers, and population heterogeneity: specification and estimation of a nonparametric, mixed Poisson model', *Criminology*, 31(3), pp. 327–362.
38. Owens, A.P., Allen, M., Ondobaka, S. and Friston, K.J. (2018) 'Interoceptive inference: from computational neuroscience to clinic', *Neuroscience & Biobehavioral Reviews*, 90, pp. 174–183.
39. Parr, T., Pezzulo, G. and Friston, K.J. (2022) *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. Cambridge, MA: MIT Press.
40. Petrosino, A., Turpin-Petrosino, C. and Guckenburg, S. (2010) 'Formal system processing of juveniles: effects on delinquency', *Campbell Systematic Reviews*, 6(1), pp. 1–88.
41. Sampson, R.J. and Laub, J.H. (1993) *Crime in the Making: Pathways and Turning Points Through Life*. Cambridge, MA: Harvard University Press.
42. Sampson, R.J. and Laub, J.H. (2003) 'Life-course desisters? Trajectories of crime among delinquent boys followed to age 70', *Criminology*, 41(3), pp. 555–592.
43. Schwartenbeck, P., FitzGerald, T.H., Mathys, C., Dolan, R. and Friston, K.J. (2015) 'The dopaminergic midbrain encodes the expected certainty about desired outcomes', *Cerebral Cortex*, 25(10), pp. 3434–3445.
44. Seth, A.K. (2013) 'Interoceptive inference, emotion, and the embodied self', *Trends in Cognitive Sciences*, 17(11), pp. 565–573.
45. Seth, A.K. and Friston, K.J. (2016) 'Active interoceptive inference and the emotional brain', *Philosophical Transactions of the Royal Society B*, 371, 20160007.
46. Shay, J. (1994) *Achilles in Vietnam: Combat Trauma and the Undoing of Character*. New York: Scribner.
47. Shonin, E., Van Gordon, W., Slade, K. and Griffiths, M.D. (2013) 'Mindfulness and other Buddhist-derived interventions in correctional settings: a systematic review', *Aggression and Violent Behavior*, 18(3), pp. 365–372.
48. Shulman, E.P., Smith, A.R., Silva, K., Icenogle, G., Duell, N., Chein, J. and Steinberg, L. (2016) 'The dual systems model: review, reappraisal, and reaffirmation', *Developmental Cognitive Neuroscience*, 17, pp. 103–117.
49. Smith, R., Badcock, P.B. and Friston, K.J. (2021) 'Recent advances in the application of predictive coding and active inference models within clinical neuroscience', *Psychiatry and Clinical Neurosciences*, 75(1), pp. 3–13.

50. Smith, R., Schwartenbeck, P., Parr, T. and Friston, K.J. (2020) 'An active inference approach to modeling structure learning: concept learning as an example', *Frontiers in Computational Neuroscience*, 14, p. 41.
51. Steinberg, L. (2008) 'A social neuroscience perspective on adolescent risk-taking', *Developmental Review*, 28(1), pp. 78–106.
52. Steinberg, L. (2010) 'A dual systems model of adolescent risk-taking', *Developmental Psychobiology*, 52(3), pp. 216–224.
53. Téglás, E., Girotto, V., Gonzalez, M. and Bonatti, L. (2007) 'Intuitions of probabilities shape expectations about the future at 12 months and beyond', *Proceedings of the National Academy of Sciences*, 104(49), pp. 19156–19159.
54. Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J.B. and Bonatti, L.L. (2011) 'Pure reasoning in 12-month-old infants as probabilistic inference', *Science*, 332(6033), pp. 1054–1059.
55. Wilkinson, S., Dodgson, G. and Meares, K. (2017) 'Predictive processing and the varieties of psychological trauma', *Frontiers in Psychology*, 8, p. 1840.
56. Williams, D. (2020) 'Predictive processing and the representation wars', *Minds and Machines*, 30, pp. 141–172.
57. Wilson, W.J. (1987) *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. Chicago: University of Chicago Press.
58. Wareing, L., Readman, M.R., Longo, M.R., Linkenauger, S.A. and Crawford, T.J. (2024) 'The utility of heartrate and heartrate variability biofeedback for the improvement of interoception across behavioural, physiological and neural outcome measures: a systematic review', *Brain Sciences*, 14(6), p. 579.
59. Weitekamp, E.G.M. and Kerner, H.-J. (eds) (1994) *Cross-National Longitudinal Research on Human Development and Criminal Behavior*. NATO ASI Series, vol. 76. Dordrecht: Kluwer Academic Publishers.
60. Xanthios, K. and Saffran, P. (2023) 'Examining moral injury using a predictive processing framework', eScholarship, University of California.
61. Xu, F. and Garcia, V. (2008) 'Intuitive statistics by 8-month-old infants', *Proceedings of the National Academy of Sciences*, 105(13), pp. 5012–5015.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.