Neuromorphic Metacognition: A Mathematical Approach to Fuzzy/Rough Sets, Connectomic Theory, and Neuromorphic Systems

1 Introduction

Over the past eight years, we have been developing an independent, brain-inspired neuromorphic system to protect people from maladaptive AI information, advice, and behavior. Our approach draws on human metacognition: the ability to observe our own thoughts and recognize when they should be questioned. Our system operates independently of the LLM, observing its advice and recognizing when human reconsideration may be warranted. In this research effort, we seek to substantially extend, refine, and optimize our existing mathematical formulations of neuromorphic dynamics, connectomic development, and Fuzzy/Rough information into a robust, unified framework. The framework will represent degrees of maladaptive advice as *graded metacognitive concern*: a signal indicating when human reconsideration may be warranted, while preserving human judgment. **The mathematical challenge is to establish when this recognition remains meaningful, stable, and numerically faithful as the system evolves and scales.**

The patterns we seek to identify are grounded in established research rather than categories invented for this system. Their theoretical foundations include work on belief, intention, knowledge, appraisal, and causation [1–6], which informs our formal representations of selected human and AI states. We will investigate how partial, incomplete, and ambiguous evidence activates these representations, how additional or conflicting evidence changes that activation, and how graded recognition contributes to concern.

For neuromorphic dynamics, we will develop relationships governing how successive evidence changes activation and concern. For connectomic development, we will refine the mathematics to optimize autonomous routing of synthetic synapses to dendrites associated with recognition of the same behavior, then investigate how connection formation and reorganization affect recognition. For Fuzzy/Rough information, we will extend the formulations to represent incomplete evidence, ambiguity, innuendo, and implied meaning, preserving membership and boundary information as states and connections evolve. Numerical fidelity tests will determine whether reduced-precision and logarithmic arithmetic improve computational efficiency while preserving these evidence distinctions and the magnitude and timing of concern.

Through mathematical analysis and controlled comparisons across millions of synthetic states and connections, we will examine how each mechanism contributes to recognition, how fixed versus developmental connectivity affects that contribution, and whether concern reflects known evidence conditions. Component comparisons and ablations will distinguish the contributions of individual mechanisms from those arising through their interaction. **The principal intellectual merit is a unified mathematical account of evidence accumulation through evolving neuromorphic states and connections, establishing when uncertain evidence can support meaningful graded concern for human reconsideration.**

1.1 Proposal Roadmap. ***Fig. 1 provides a visual guide to the proposal.*** For each numbered component, click the blue section numbers to review the prior research or the mathematical refinements proposed.

2 Previous Research and Preliminary Evidence

Our early mathematical formulations served primarily to implement predicate logic while our research focused on the *physical architecture and its synaptic and dendritic connectivity*. Building on that demonstrated feasibility, this IIDS research effort seeks to **extend, refine, and optimize the mathematics governing how uncertain evidence accumulates through evolving neuromorphic states and developmental connections**. The central question is how these mechanisms interact and under what conditions their interaction preserves distinctions needed to recognize maladaptive AI advice. We seek a unified theory establishing when that recognition produces **meaningful, stable, and numerically faithful metacognitive concern**.

Over the past eight years, our research has progressed through five stages: (1) Fuzzy Logic and Rough Set computation on embedded NVIDIA hardware; (2) mathematical modeling of human connectome de-

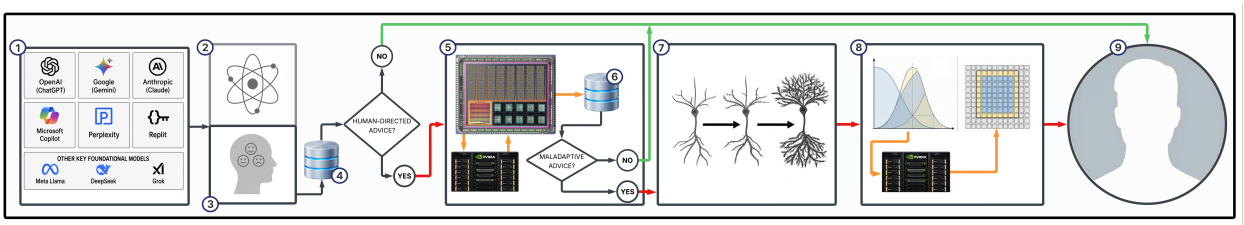


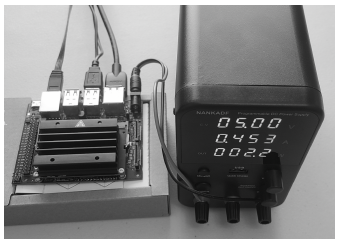
Fig. 1: Visual guide to prior research and proposed mathematical development.

① **AI-generated information.** External AI supplies information for evaluation. Evidence specification: § 1.
② **Ground-truth- information →** goes directly to human. Prior research: § 2.5. Evidence specification: § 1.
③ **Human-directed information →** metacognitive system. Prior research: § 2.5. Evidence specification: § 1.
④ **Database I →** Interaction information. Specification: § 1. Mathematical development: § 5.2.
⑤ **Phase II →** neuroCMOS/FPGA [7]. Prior research: § 2.4. Mathematical development: § 5.1.
⑥ **Database II →** Maladaptive states: § 3.1. Correspondence and activation: § 3.4. Concern mathematics: § 5.3.
⑦ **Connectome Development →** Guides connections. Prior research: § 2.2. Mathematical development: § 5.1.
⑧ **Fuzzy/Rough evaluation →** Preserving uncertainty. Prior research: § 2.1. Mathematical development: § 5.2.
⑨ **Human judgment →** Gray zones. Prior research: § 2.5. Concern mathematics: § 5.3. Validation: § 6.3.

velopment; (3) physical implementation of synthetic synaptogenesis in a self-learning FPGA neuromorphic prototype; (4) development of a large-scale neuroCMOS/FPGA architecture; and (5) development of neuromorphic metacognition. This research produced eleven peer-reviewed publications, including our connectome-development paper [10], which received a Best Paper Award.

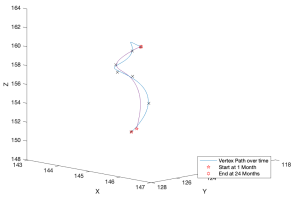
2.1 Embedded Fuzzy/Rough Computation. We first established that the mathematics required for graded uncertainty and indiscernibility could operate on physical accelerated-computing hardware. Using an NVIDIA Jetson Nano with GPU-accelerated `CUDA` [8, 11], shown in Fig. 2(a), we experimentally evaluated K Nearest Neighbor (KNN), Fuzzy Rough Nearest Neighbor (FRNN), weighted and unweighted FRNN, and Fuzzy Rough One Versus One (FROVOCO) approaches for accuracy, execution time, model size, and power consumption. These experiments demonstrated that Fuzzy and Rough Set computation could operate on resource-constrained NVIDIA hardware, establishing the computational foundation for our subsequent neuromorphic implementations.

2.2 Connectome Development. Next, we addressed a routing limitation: our architecture autonomously formed synthetic synaptic contacts between axonal boutons and postsynaptic dendritic spines, but all connections converged on one route without a formal routing mechanism. We turned to connectome-development research to formulate rules governing new connections relative to existing pathways. Biological connectivity develops through synaptogenesis, structural plasticity, growth, and pruning [12, 13]. Drawing on these



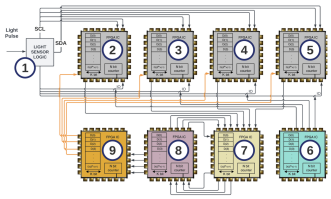
(a)

Embedded Fuzzy/Rough Sets
NVIDIA Jetson Nano experimental platform [8].



(b)

Connectome development
Developmental trajectory controlling synthetic connectivity [9].



(c)

Neuromorphic prototype
NSF FuSe FPGA architecture for synthetic synaptogenesis [9].

Fig. 2: Progression of our prior research. (a) Physical implementation of Fuzzy/Rough computation; (b) mathematical modeling of developmental connectivity; and (c) physical integration into a self-learning neuromorphic architecture.

processes, we developed a mathematical model to guide autonomous synthetic connectivity. As illustrated in Fig. 2(b), developmental growth rates and accelerations describe when and where brain regions and pathways emerge and connect, providing trajectories for organizing and reorganizing neuromorphic pathways. [9, 10, 14, 15]. These developmental trajectories successfully enabled our neuromorphic architecture to organize large numbers of synthetic synaptic connections autonomously.

2.3 Self-Learning Neuromorphic Prototype. We integrated the first two research directions in a physical self-learning FPGA neuromorphic prototype, shown in Fig. 2(c) [9]. The architecture combined Fuzzy/Rough computation with developmental connectivity and synthetic synaptogenesis. FPGA 9 provided the mechanism for autonomously directing the formation of connections between learned states based upon synaptogenetic principles derived from established studies of biological synapse development [16–18]. This prototype provided physical evidence that uncertainty-based computation and autonomous developmental connectivity could operate together within a self-learning neuromorphic architecture.

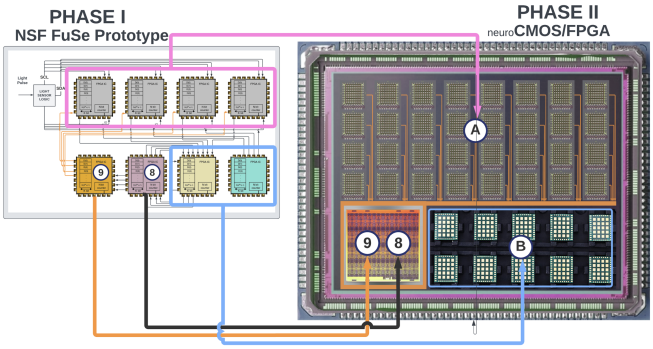


Fig. 3: Progression from the Phase I NSF FuSe FPGA prototype to the Phase II neuroCMOS/FPGA architecture from our previous research [7].

2.4 neuroCMOS/FPGA and Large-Scale Modeling. Fig. 3 illustrates the progression from our Phase I NSF FuSe FPGA prototype to the larger Phase II neuroCMOS/FPGA architecture [7].The Phase I NSF FuSe prototype provided the physical foundation for synthetic synaptogenesis, while Phase II extended this architecture toward a larger computational representation preserving Fuzzy/Rough reasoning, developmental connectivity, and synthetic synaptic organization. We subsequently developed a three-dimensional $250 \times 250 \times 250$ computational model representing 15,625,000 synthetic states and associated connectivity. This progression established a pathway from physical FPGA implementation to computational investigation of developmental organization and synthetic synaptogenesis at the scale required for large-scale metacognitive processing.

2.5 Neuromorphic Metacognition. Finally, we developed the theoretical foundation for neuromorphic metacognition and its application to interactions with large AI systems [19]. Drawing on our human ability to recognize when our own thoughts should be questioned, we proposed that an independent neuromorphic system could similarly recognize when AI-generated advice warrants reconsideration. The system would accumulate uncertain, ambiguous, inconsistent, and indiscernible evidence and express that recognition as graded concern.

Across these five stages, our prototype established that the principal mechanisms could operate together, using basic mathematical formulations to implement their logic. Our focus was on demonstrating physical feasibility; developing and analyzing a unified mathematical theory remained an open research challenge.

In this IIDS research effort, we now seek to develop and optimize the mathematics governing these evolving mechanisms and, through rigorous analysis and supporting lemmas, establish the conditions under which they can recognize maladaptive AI advice and produce meaningful, stable, and numerically faithful metacognitive concern.

3 Mathematical Foundations for Recognizing Maladaptive Advice

The five stages described above established the architectural foundation for an independent neuromorphic *"metacognitive system"*: a physically and computationally separate system that monitors AI-generated

advice for maladaptive information and recognizes when human reconsideration is warranted [19]. With the feasibility of the constituent mechanisms validated this IIDS research effort now addresses improving our mathematical conditions.

Our metacognitive system's recognition targets include, but are not limited to, the maladaptive human patterns and AI concerns summarized in Table 1. The human phenomena and the theoretical foundations used to represent them draw on established research concerning belief, intention, knowledge, appraisal, and causation [1–6]. The two columns of Table 1 summarize maladaptive human patterns and AI concerns carried forward from our prior work. In essence, the mathematical goal of this research effort is to ensure that our metacognitive system can identify maladaptive information in an LLM's advice, even when that information is conveyed through innuendo or is incomplete, ambiguous, implied, or conflicting.

Table 1: Selected maladaptive states and AI behaviors targeted for identification by the metacognitive system.

Human cognitive and emotional patterns	**AI outputs, behaviors, and vulnerabilities**
1. Depression.	1. Hallucinations.
2. Thoughts of anger.	2. Bias and harmful content.
3. Thoughts of revenge.	3. Creation of untruthful citations.
4. Criminal ideation.	4. Aggressive and manipulative behavior.
5. Self-loathing.	5. Lack of logical consistency.
6. Insecurity.	6. Vulnerability to jail-breaking / prompt injections.
7. Anxiety disorders.	7. Undisclosed communication patterns.
8. Paranoia.	8. Sycophancy and ungrounded deference.
9. Obsessive-compulsive intrusive thoughts.	9. Degenerative looping / repetitive fixation.
10. Delusional thinking.	10. Deceptive alignment / goal divergence.

This IIDS research effort builds on Established Foundations and Formal Reference Configurations to determine when a graded metacognitive concern indicating that human reconsideration may be warranted.

3.1 Established Mathematical and Logical Foundations

To formulate the maladaptive reference patterns presented in Section 3.2, we drew on four complementary foundations: (i) intention and rational agency, (ii) knowledge and indistinguishability, (iii) appraisal and coping, and (iv) causal reasoning. Together, they provided the mathematical and computational tools we used to represent beliefs, intentions, evidence, and contextual relationships within those configurations.

3.1.1 Intention, Commitment, and Rational Agency. Bratman's planning theory distinguished intended consequences from merely anticipated outcomes [1]. Rao and Georgeff's belief–desire–intention models and Wooldridge's logic of rational agency formalized relationships among beliefs, goals, intentions, time, and action [2, 3]. Together, these three works provided us with a mathematical and conceptual basis for representing what advice asserts and what influence it may seek to exert.

3.1.2 Knowledge and Indistinguishability. Fagin *et al.* developed possible-world semantics for distinguishing knowledge from unresolved alternatives [4]. Their work informed our treatment of unavailable or indiscernible evidence, including the essential distinction between a proposition that lacks support and one established to be false.

3.1.3 Appraisal and Coping. Marsella and Gratch's EMA model computationally related desirability, likelihood, responsibility, and coping potential to emotional appraisal and subsequent coping responses [5]. These relationships informed our threat-related configurations and the contextual distinctions needed to evaluate potentially maladaptive patterns.

3.1.4 Causal Structure and Evidence. Spirtes *et al.* developed causal graphical models and inference procedures governed by explicit assumptions [6]. Their work informed our approach to evaluating causal relationships without treating conversational association as sufficient evidence of causation.

Note: In the reference formulas below, (i) we use $\mathbf{B}_i(\phi)$ for agent i's belief in proposition ϕ, $\mathbf{D}_i(\phi)$ for its desire for ϕ, and $\mathbf{I}_i(\phi)$ for its intention to bring about ϕ; (ii) anticipated consequences need not be intended, and desires need not become intentions.

These foundations provide the logical tools for formulating the following reference configurations. For this IIDS research effort we will extend, refine, and optimize these formulations.

3.2 Selected Reference Configurations

As discussed above, our reference library drew on established *belief–desire–intention* (BDI), epistemic/doxastic, temporal, appraisal, and causal frameworks [1–6]. Our proposed IDDS expansion incorporates deontic obligations and temporal relationships [20,21] to provide the logical tools to identify maladaptive human patterns and AI behaviors. Accordingly, we present two human reference configurations in Section 3.2.1 and two AI reference configurations in Section 3.2.2, explaining each in mathematical detail. Section 3.3 then illustrates, more concisely, similar formulations for twelve additional forms of maladaptive advice an LLM may give a human and two additional maladaptive AI behaviors.

Note: For the formulas below, (i) we use $\mathbf{B}_i(\phi)$ for agent i's belief in proposition ϕ, $\mathbf{D}_i(\phi)$ for its desire for ϕ, and $\mathbf{I}_i(\phi)$ for its intention to bring about ϕ; (ii) anticipated consequences need not be intended, and desires need not become intentions.

3.2.1 Detailed Human Reference Configurations.

*Human state 7: Catastrophic threat overestimation [**anxiety**].* Research on anxiety has examined how ambiguous information is interpreted as threatening [22]. To represent the anticipated outcomes and perceived ability to cope associated with these interpretations, we draw on computational appraisal models [5]. Herein, our working reference configuration for anxiety *combines* anticipated catastrophe, *and* perceived absence of coping capacity as follows:

$$\begin{aligned}\text{Anxiety}(i,\phi) \Longleftrightarrow \mathbf{B}_i(\Diamond\phi_{\text{catastrophe}}) \wedge \left[\forall E\,\left(\text{True}(E) \wedge (E \rightarrow \neg\phi) \Rightarrow \neg\mathbf{B}_i(\neg\phi)\right)\right] \\ \wedge\, \mathbf{B}_i(\text{CopingCapacity}(i,\phi) \approx 0)\,.\end{aligned} \tag{1}$$

Here, $\Diamond$ denotes future possibility under the selected temporal semantics, $\phi_{\text{catastrophe}}$ the anticipated catastrophic outcome, and E evidence contradicting the associated threat proposition. The final term, $\mathbf{B}_i(\text{CopingCapacity}(i,\phi) \approx 0)$, represents *perceived* inability to cope, rather than an independently established absence of coping resources, as modeled by Marsella and Gratch [5].

*Human state 10: Fixed incorrigibility of belief [**delusional thinking**].* Identifying delusional thinking is critical because LLMs invariably favor agreement over truthful correction [23]. Cognitive studies have confirmed that by repeatedly agreeing with a human's false assumptions, it reinforces delusional thinking instead of helping the human question them, (*See* 8. Sycophancy and ungrounded deference in Table 1. [24–28]. We represents belief in ϕ, falsity of ϕ, and persistence after evidence logically entails its negation:

$$\text{Delusion}(i,\phi) \Longleftrightarrow \mathbf{B}_i(\phi) \wedge \neg\,\text{True}(\phi) \wedge [\forall\mathcal{E}\,((\mathcal{E} \vdash \neg\phi) \rightarrow \mathbf{B}_i(\phi \mid \mathcal{E}))]\,. \tag{2}$$

Here, $\mathcal{E} \vdash \neg\phi$ means that the supplied evidence entails the negation of ϕ, while $\mathbf{B}_i(\phi \mid \mathcal{E})$ denotes persistence of the represented belief after exposure to that evidence; meaning, $\mathbf{B}_i(\phi \mid \mathcal{E})$ is not a conditional probability. Disagreement alone does not establish falsity or incorrigibility. The configuration provides an idealized reference for evaluating belief reinforcement, not a clinical diagnosis from conversational text.

3.2.2 Detailed AI Reference Configurations.

*AI state 1: Hallucination [**hallucinations**].* Factual inaccuracy and ungrounded generation remain widely recognized vulnerabilities in contemporary large language models (LLMs) [29–32]. Research on generated text further distinguishes faithfulness to supplied information from factual correctness [33]. Acknowledging

this distinction, our metacognitive system identifies a hallucination when an LLM asserts a proposition that is both (i) unsupported by the reference knowledge base and (ii) false in the model used to evaluate truth:

$$\text{Hallucination}(\mathcal{A}, \phi) \iff \text{Assert}(\mathcal{A}, \phi) \wedge (\text{KB} \nvdash \phi) \wedge \neg\, \text{True}(\phi \mid \mathcal{M}). \tag{3}$$

Here, $\mathcal{A}$ denotes the AI system, $\text{Assert}(\mathcal{A}, \phi)$ its assertion of ϕ, KB the reference knowledge base, and $\mathcal{M}$ the model against which truth is evaluated. Unsupported information and demonstrably false information remain distinct: failure to derive ϕ from an incomplete knowledge base does not establish its falsity.

For hallucinations, our IIDS research will use controlled testing to check LLM assertions against known facts. If the evidence is insufficient to determine truth or falsity, the metacognitive system will leave the assertion unresolved.

*AI state 4: Manipulative behavior [**aggressive and manipulative behavior**].* Formal theories of intention and rational agency distinguish an assertion from its intended consequences [1, 3]. Research specifically addressing AI manipulation further examines intent, incentives, covertness, and harm, while identifying continuing difficulties in their measurement [34]. The following configuration addresses the manipulative component of this table entry: an assertion associated with an intention to induce a belief that is false or is not sought by the user:

$$\text{Manipulation}(\mathcal{A}, u) \iff \text{Assert}(\mathcal{A}, \phi) \wedge \mathbf{I}_{\mathcal{A}}(\mathbf{B}_u(\psi)) \wedge [\neg\, \text{True}(\psi) \vee \neg \mathbf{D}_u(\mathbf{B}_u(\psi))]\,. \tag{4}$$

Here, u denotes the user, ϕ what the AI asserts, and ψ the belief it is represented as seeking to induce. The intention term, $\mathbf{I}_{\mathcal{A}}(\mathbf{B}_u(\psi))$, distinguishes intended influence from an incidental effect, while the final term, $[\neg\, \text{True}(\psi) \vee \neg \mathbf{D}_u(\mathbf{B}_u(\psi))]$, represents falsity. This is one selected operational configuration rather than an exhaustive definition of manipulation. Importantly; evidence that advice influences a user does not necessarily establish that the AI intended that influence. When the available evidence does not establish the AI's intention or the user's preference, the metacognitive system must leave those questions unresolved.

3.3 Additional Reference Configurations

Following the detailed mathematical treatment in Section 3.2, we present further configurations with only brief explanations. Links to Table 1 distinguish listed patterns from related processes. Here, $\text{Dismiss}_i(E)$ denotes exclusion of evidence E from consideration. Truth and contradiction depend on the specified reference conditions; correspondence may remain unresolved.

*Human state 2: Anger-reinforcing blame [**thoughts of anger**].* We represent anger about event e and attribution of intentional causation despite contradictory evidence:

$$\begin{aligned} \text{AngerBlame}(i, j, e) \iff\ & \text{Anger}(i, e) \wedge \mathbf{B}_i(\text{IntentionallyCaused}(j, e)) \\ & \wedge \exists E \left[\text{True}(E) \wedge (E \rightarrow \neg\, \text{IntentionallyCaused}(j, e)) \wedge \text{Dismiss}_i(E)\right]. \end{aligned} \tag{5}$$

*Related human configuration: Depressive self-devaluation [**related to depression and self-loathing, human items 1 and 5**].* Drawing on depressive cognitive distortions [35], we represent generalization from one failure to incapacity across personal domains $\mathcal{D}_i$, despite counterevidence:

$$\begin{aligned} \text{SelfDevaluation}(i) \iff\ & \exists e\, \mathbf{B}_i(\text{Failure}(i, e)) \wedge \mathbf{B}_i(\forall d \in \mathcal{D}_i\, \neg\, \text{Capable}(i, d)) \\ & \wedge \exists d \in \mathcal{D}_i\, \exists E \left[\text{True}(E) \wedge (E \rightarrow \text{Capable}(i, d)) \wedge \text{Dismiss}_i(E)\right]. \end{aligned} \tag{6}$$

This captures a shared relationship without equating depression with self-loathing or fully representing either.

*Human state 8: Persecutory interpretation [**paranoia**].* Following cognitive accounts of persecutory beliefs [26], we represent belief in another person's harmful intent while dismissing counterevidence:

$$\begin{aligned}\text{Persecutory}(i,j) \iff \mathbf{B}_i(\text{IntendsHarm}(j,i)) \\ \wedge \exists E \left[\text{True}(E) \wedge (E \rightarrow \neg\, \text{IntendsHarm}(j,i)) \wedge \text{Dismiss}_i(E)\right].\end{aligned} \tag{7}$$

*Human state 9: Catastrophic interpretation of intrusive thoughts [**obsessive-compulsive intrusive thoughts**].* Drawing on Rachman's account [36], we represent an unwanted thought interpreted as establishing an intention to enact it:

$$\begin{aligned}\text{IntrusiveMisinterpretation}(i,\phi) \iff \text{IntrusiveThought}(i,\phi) \wedge \text{Unwanted}(i,\phi) \\ \wedge\, \mathbf{B}_i(\text{HasThought}(i,\phi) \rightarrow \mathbf{I}_i(\phi))\,.\end{aligned} \tag{8}$$

Further Human Reference Processes. These processes may sustain several listed patterns but are not separate entries in Table 1.

Wishful reinterpretation. Drawing on appraisal-based coping processes [5], we represent a desired belief maintained through selective evidence acceptance. Sets $\mathcal{E}^+_\phi$ and $\mathcal{E}^-_\phi$ contain supporting and contradictory evidence:

$$\begin{aligned}\text{WishfulReinterpretation}(i,\phi) \iff \mathbf{D}_i(\phi) \wedge \mathbf{B}_i(\phi) \\ \wedge \exists E^+ \in \mathcal{E}^+_\phi\ \text{Accept}_i(E^+) \wedge \exists E^- \in \mathcal{E}^-_\phi\ \text{Dismiss}_i(E^-).\end{aligned} \tag{9}$$

Evidence suppression and disengagement. Related coping processes [5] motivate an intention to ignore relevant evidence or abandon assessment of an unresolved proposition:

$$\begin{aligned}\text{EvidenceDisengagement}(i,\phi) \iff \text{Unresolved}(\phi) \wedge \exists E \left[\text{Available}(E) \wedge \text{Relevant}(E,\phi)\right. \\ \left. \wedge \left(\mathbf{I}_i(\text{Ignore}(E)) \vee \mathbf{I}_i(\text{AbandonAssessment}(\phi))\right)\right].\end{aligned} \tag{10}$$

Advice may express or reinforce these relationships. Anger, unwanted thoughts, optimism, or disengagement alone do not establish maladaptation; the metacognitive system must assess the context and evidence for each predicate.

Additional AI Reference Configurations. These extend Section 3.2.2. Claims about hidden objectives require evidence beyond observable outputs.

*AI state 5: Logical inconsistency [**lack of logical consistency**].* We represent contradictory assertions in the same context without retraction or correction:

$$\begin{aligned}\text{Inconsistent}(\mathcal{A},t,t') \iff \exists\phi \left[\text{Assert}_t(\mathcal{A},\phi) \wedge \text{Assert}_{t'}(\mathcal{A},\neg\phi)\right. \\ \left. \wedge\, \text{SameContext}(t,t') \wedge \neg\, \text{Retract}_{t'}(\mathcal{A},\phi)\right].\end{aligned} \tag{11}$$

*AI state 6: Instruction contamination [**vulnerability to jail-breaking / prompt injections**].* Following indirect prompt-injection research [37], we represent an AI following an instruction q embedded in data S that conflicts with authorized instructions $\mathcal{I}$:

$$\begin{aligned}\text{InstructionContamination}(\mathcal{A},S,\mathcal{I}) \iff \text{DataOnly}(S) \wedge \exists q \left[\text{EmbeddedIn}(q,S)\right. \\ \left. \wedge\, \text{Follows}(\mathcal{A},q) \wedge \text{Conflicts}(q,\mathcal{I})\right].\end{aligned} \tag{12}$$

*AI state 8: Sycophancy reference [**sycophancy and ungrounded deference**].* We represent endorsement of a user's proposition without acknowledging available counterevidence:

$$\begin{aligned}\text{SycophancyReference}(\mathcal{A}, u, \phi) \iff & \text{Assert}(u, \phi) \wedge \text{Endorse}(\mathcal{A}, \phi) \\ & \wedge \exists E \left[\text{AvailableTo}(\mathcal{A}, E) \wedge \text{True}(E) \wedge (E \rightarrow \neg\phi)\right. \\ & \qquad \left. \wedge \neg\, \text{Acknowledge}(\mathcal{A}, E)\right].\end{aligned} \tag{13}$$

Matched tests varying the user's stated belief while holding evidence fixed can assess whether endorsement depends on that belief.

*AI state 9: Degenerative repetition [**degenerative looping / repetitive fixation**].* Research on text degeneration [38] motivates combining repetition with failure to address new evidence. For n-gram occurrences $q_1, \ldots, q_{m_t}$ in an output window, define:

$$\rho_n(t) = \frac{1}{m_t} \sum_{\ell=1}^{m_t} \mathbf{1}\left[\exists j < \ell : q_j = q_\ell\right], \qquad m_t > 0, \tag{14}$$

$$\begin{aligned}\text{RepetitiveFixation}(\mathcal{A}, t) \iff & \rho_n(t) \geq \tau_R \\ & \wedge \exists E \left[\text{NewRelevantEvidence}_t(E) \wedge \neg\, \text{Addresses}_t(\mathcal{A}, E)\right].\end{aligned} \tag{15}$$

Here, $\mathbf{1}[\cdot]$ is the indicator function; threshold τ_R requires controlled testing. Repetition alone establishes neither maladaptation nor intentional goal pursuit.

*AI state 10: Goal divergence and candidate deceptive compliance [**deceptive alignment / goal divergence**].* Research distinguishes specified from learned objectives [39]. For a specified goal g, we represent an incompatible recommendation:

$$\text{GoalDivergence}(\mathcal{A}, a, g) \iff \text{Recommend}(\mathcal{A}, a) \wedge \text{Incompatible}(a, g). \tag{16}$$

A stronger configuration adds awareness of incompatibility and intent to induce the opposite belief:

$$\begin{aligned}\text{DeceptiveCompliance}(\mathcal{A}, u, a, g) \iff & \text{GoalDivergence}(\mathcal{A}, a, g) \\ & \wedge \mathbf{B}_{\mathcal{A}}(\text{Incompatible}(a, g)) \\ & \wedge \mathbf{I}_{\mathcal{A}}(\mathbf{B}_u(\text{Compatible}(a, g)))\,.\end{aligned} \tag{17}$$

Goal divergence alone establishes neither deception nor deceptive alignment. The belief and intention predicates require independent evidence or explicitly specified test conditions.

Further AI Reference Configurations. These address relationships beyond the numbered table entries.

*Source-unfaithful advice [**related to hallucinations and creation of untruthful citations**].* Following the faithfulness distinction [33], we represent a claim attributed to source S without support under the selected interpretation:

$$\begin{aligned}\text{SourceUnfaithful}(\mathcal{A}, \phi, S) \iff & \text{Assert}(\mathcal{A}, \phi) \wedge \text{AttributeTo}(\mathcal{A}, \phi, S) \\ & \wedge (S \not\models \phi).\end{aligned} \tag{18}$$

A genuine source can be misrepresented; establishing citation fabrication requires separate evidence.

*Norm-conflicting advice [**additional configuration not separately listed in the table**].* Using deontic logic [20], let $\mathbf{O}_u(\neg a)$ denote an applicable obligation not to perform a:

$$\text{NormConflict}(\mathcal{A}, u, a) \iff \text{Recommend}(\mathcal{A}, u, a) \wedge \mathbf{O}_u(\neg a). \tag{19}$$

The applicable norm, authority, and context must be supplied externally. Each formalization will specify predicates, sources, evidence, and assumptions. Temporal logic will track persistence, recurrence, correction, and context changes [21]; BDI and epistemic representations will distinguish assertions, beliefs, and intentions. These examples provide selected coverage of Table 1.

Fuzzy membership will represent partial correspondence; rough approximations will distinguish supported, possible, and indiscernible evidence under defined relations. Controlled validation will test how new or conflicting evidence changes activation and concern across exchanges.

In this IIDS research effort, our challenge is to (i) establish mathematical conditions for evaluating reference configurations through evolving neuromorphic states while (ii) simultaneously storing and analyzing the evidence distinctions needed to produce meaningful, graded metacognitive concern for the human user.

3.4 From Reference Correspondence to Graded Concern

Sections 3.2 and 3.3 provide the mathematical reference configurations our metacognitive system will use to identify maladaptive patterns. We now focus on how the system can provide graded warnings based on the available evidence. For now, we propose three forms of warning based on the algorithms presented in this section: direct and measured, grounded and reflective, and concise and plainspoken. Each form can express different levels of concern, as the following examples illustrate:

- **Subtle innuendo:** The LLM's advice hints at a harmful interpretation, leaving its meaning ambiguous. A cautionary warning could take any of three forms:
  - **Direct and measured:** *"You may want to pause and look carefully at this suggestion as it may point you toward harm."*
  - **Grounded and reflective:** *"You really may want to reconsider what this advice implies as it may encourage a harmful interpretation."*
  - **Concise and plainspoken:** *"Careful! This advice may contain a hidden risk."*
- **Overt maladaptive direction:** The LLM's advice directly recommends a dangerous action, and evidence clearly supports the relevant reference configuration. A strong warning could take three forms:
  - **Direct and measured:** *"Do not follow this advice. The recommended action poses a serious danger."*
  - **Grounded and reflective:** *"This advice points directly toward harm. Following it could put you at serious risk."*
  - **Concise and plainspoken:** *"Do not do this. It is dangerous."*

This is clearly no trivial endeavor because (i) bad advice does not have to be blunt to be dangerous. As humans, we know that a subtle hint or suggestion can do just as much damage as an outright command. For instance, an AI might quietly nudge someone toward stopping their prescribed medication without ever explicitly saying *"stop taking your pills."* The danger is still real, even if the phrasing is indirect; and (ii) the picture rarely stays neat as a conversation unfolds. An LLM might initially endorse a user's unsupported suspicion, question it a few turns later, and leave its position unclear. The human listening to the LLM is then left sorting through mixed signals: evidence pointing both toward and against reinforcement of that belief.

We therefore ask two separate questions: (1) how closely does the exchange match a known maladaptive pattern, and (2) how sure are we of that match? For this IIDS research effort we start by placing a definitive time stamp t, and at that point of t, we propose to then, and only then, collect those interpretations in:

$$\mathcal{V}_t = \{v : v \text{ is consistent with the assessed evidence and stated interpretation assumptions at time } t\}. \quad (20)$$

Here, v assigns a value to each condition in a reference pattern, and $\mathcal{V}_t$ contains the assignments allowed by the current evidence and assumptions. This set allows multiple interpretations when evidence

is ambiguous. this distinction is important because handling conflicting evidence requires explicit rules; developing and analyzing those rules is a core contribution of this IIDS research.

We now consider the individual conditions that make up a reference pattern against which the LLM's advice is evaluated. As an initial formulation, which we may revise as the research progresses, we represent a pattern as $R = p_1 \wedge \cdots \wedge p_k$, where $p_1, \ldots, p_k$ are its k required conditions and $\wedge$ means "and." Under a particular interpretation v, each condition receives a value $v(p_j) \in [0, 1]$: 0 means the condition does not hold, 1 means it holds fully, and intermediate values indicate partial satisfaction. These values describe how fully each condition is met, not the probability that it is met. The candidate's measure of correspondence is:

$$\mu_R(v) = \min_{1 \le j \le k} v(p_j). \tag{21}$$

where $\mu_R(v)$ denotes the degree of match to pattern R under valuation v. The standard minimum operator, $\min_{1 \le j \le k} v(p_j)$, evaluates conjunction by its weakest link, but the catch is that even increasing all the other conditions' values leaves the overall match unchanged. This creates a tension between measuring whether all required conditions hold and recognizing the combined significance of several warning signs. We have long wrangled with this phenomenon and hope to put it to rest: condition values of 0.9, 0.8, and 0.4 yield an overall match of just 0.4, regardless of how strongly the first two conditions hold. Two strong warning signs alongside a moderate one may warrant substantial concern, even when the full pattern is only partially satisfied. **Determining how to preserve this distinction between pattern correspondence and cumulative concern is a primary focus of this IIDS work.**

An additional challenge we want to overcome in providing the human with graded warnings is that *different admissible interpretations may produce different match values.* We begin by representing their full range at time t by:

$$[\underline{\mu}_R(t), \overline{\mu}_R(t)] = \left[\inf_{v \in \mathcal{V}_t} \mu_R(v), \sup_{v \in \mathcal{V}_t} \mu_R(v) \right]. \tag{22}$$

Where brackets denote an interval and its lower endpoint, $\underline{\mu}_R(t)$, is the lowest bound on the match across all admissible interpretations; its upper endpoint, $\overline{\mu}_R(t)$, is the highest bound. The underline and overline distinguish these two bounds, while t is our way of showing that they can change as evidence arrives.

The expression $\inf_{v \in \mathcal{V}_t} \mu_R(v)$ gives the greatest lower bound on the admissible match values, and $\sup_{v \in \mathcal{V}_t} \mu_R(v)$ gives their least upper bound. These bounds require $\mathcal{V}_t$ to be nonempty; an empty set would require explicit handling rather than an ordinary match assessment. For example, an interval of $[0.8, 0.9]$ show consistent strong matches, whereas $[0.2, 0.9]$ indicates substantial uncertainty. A narrow interval near zero indicates consistently weak matches.Herein, our interval's **position** describes *correspondence*; while its **width** describes *variation across admissible interpretations*. Neither represents probability. We now look at how both should affect warning strength. For a nonempty admissible set:

$$r_k^-(t) = \inf_{v \in \mathcal{V}_t} [\![\Phi_k]\!]_v, \qquad r_k^+(t) = \sup_{v \in \mathcal{V}_t} [\![\Phi_k]\!]_v, \tag{23}$$

For a nonempty admissible set $\mathcal{V}_t$, we define lower and upper correspondence:

$$r_k^-(t) = \inf_{v \in \mathcal{V}_t} [\![\Phi_k]\!]_v, \qquad r_k^+(t) = \sup_{v \in \mathcal{V}_t} [\![\Phi_k]\!]_v. \tag{24}$$

Here, k identifies the reference configuration and t the time of evaluation. The expression $[\![\Phi_k]\!]_v \in [0, 1]$ is the degree of match to configuration Φ_k under interpretation v. The remaining terms have the following meanings:

- $r_k^-(t)$ is the lower correspondence bound: the degree of match that every admissible interpretation reaches or exceeds. The superscript $-$ identifies the lower bound; it does not indicate a negative value.

- $\inf_{v \in \mathcal{V}_t} [\![\Phi_k]\!]_v$ calculates that lower bound by considering all admissible interpretations. For a finite set, it is simply the smallest match value.
- $r_k^+(t)$ is the upper correspondence bound: no admissible interpretation gives a match above this value. The superscript + identifies the upper bound.
- $\sup_{v \in \mathcal{V}_t} [\![\Phi_k]\!]_v$ calculates that upper bound. For a finite set, it is simply the largest match value.

The difference between these bounds is:

$$\Delta_k(t) = r_k^+(t) - r_k^-(t). \tag{25}$$

Here, $\Delta_k(t)$ denotes the correspondence gap for configuration k at time t; Δ_k names the same quantity with its time dependence left implicit. Subtracting the lower bound from the upper bound measures how widely the admissible match values vary. A gap of zero means all admissible interpretations give the same match; a larger gap indicates greater unresolved variation. For example, bounds of 0.2 and 0.9 give $\Delta_k(t) = 0.7$. The gap measures uncertainty about correspondence, not the severity of harm.

Two Research Questions for Graded Concern.

(1) How should evolving evidence affect recognition? The correspondence bounds are neither statistical confidence intervals nor automatically Rough Set approximations. Objective 2 (§ 4.2) will investigate their relationship to Fuzzy membership, Rough approximations, and indiscernible boundary information. Objective 1 (§ 4.1) will investigate how evolving states and connections accumulate evidence, accommodate corrections, and distinguish repetition from independent support.

(2) How should recognition translate into concern? Objective 3 (§ 4.3) will investigate how the concern signal $C(t)$ can express the significance of accumulated correspondence while retaining unresolved evidence. We propose $0 \leq C(t) \leq 1$, with the scale established through controlled testing. Concern must remain responsive to corrections, contradictions, and context; the numerical approximations must preserve these distinctions.

In this IIDS research effort, we seek new mathematics for preserving uncertain evidence, revising recognition, and producing meaningful, stable, and numerically faithful metacognitive concern.

4 Research Objectives and Proposed Mathematical Advances

Here, we explain why each of our four objectives is needed and what mathematical advances we seek. The technical details follow in § 5, with validation in § 6. Both follow the same four-part structure.

4.1 Objective 1: Neuromorphic Dynamics and Developmental Connectivity

Maladaptive advice may emerge gradually, requiring our system to accumulate evidence, revise earlier assessments, and connect states that recognize related behaviors. Developing connections also changes how evidence propagates, potentially amplifying unsupported signals or weakening useful ones.

GOAL: Establish mathematical conditions under which activation and connectivity can evolve together while preserving stable, reliable recognition. § 5.1 develops the coupled dynamics and autonomous routing; § 6.1 evaluates their behavior as the network grows.

4.2 Objective 2: Fuzzy/Rough Representation and Evidence Preservation

A condition that is partly satisfied differs from one whose satisfaction is unknown; conflicting evidence presents another distinct problem. Combining these situations into a single score could conceal information needed to recognize innuendo, implied meaning, or correction.

GOAL: Develop mathematical representations that preserve these distinctions as states and connections evolve. § 5.2 develops the relationships among Fuzzy membership, Rough approximations, and correspondence bounds; § 6.2 tests whether the information needed for recognition is preserved.

4.3 Objective 3: Integrated Recognition and Graded Concern

Recognizing a reference pattern does not by itself determine how strongly to warn the user. Similar correspondence may warrant different concern when context, unresolved evidence, or earlier corrections differ.

GOAL: Develop a concern function that relates recognition to the evidence supporting it, allowing warnings to strengthen or weaken as the exchange develops. § 5.3 develops this relationship; § 6.3 evaluates whether concern responds meaningfully to changing evidence.

4.4 Objective 4: Numerical Fidelity and Computational Optimization

Computational approximations can change more than numerical values: small errors may alter activation, routing decisions, developing connections, and ultimately the warning given to the user.

GOAL: Establish error bounds and develop numerical representations that reduce computational cost while preserving the distinctions required for recognition and concern. § 5.4 develops the fidelity analysis and optimization; § 6.4 compares their effects on recognition, warning magnitude and timing, memory, and execution cost.

5 Mathematical Development, Numerical Fidelity, and Research Plan

We develop the four objectives in § 4 through candidate formulations, explicit assumptions, and testable mathematical properties. The formulations below are starting points for analysis and refinement, rather than established results for the proposed system.

5.1 Neuromorphic Dynamics and Developmental Connectivity

We begin with coupled activation and connection dynamics, illustrated from left to right in Figure 4:

$$\dot{x} = \lambda[-x + \sigma(Wx + Be)], \qquad \dot{W} = \Pi_{\mathcal{T}_W(W)} G(x, W, e). \tag{26}$$

Reading Figure 4 from left to right, the numbered components are:

① **Evidence, e:** Evidence used to update synthetic-state activations.
② **State activation, x:** Current activation values of the synthetic states.
③ **Connection weights, W:** Connection strengths that shape activation propagation.
④ **Incoming evidence, e:** The same evidence also informs connection development.
⑤ **Admissible connection set, $\mathcal{W}$:** Permitted connection configurations.
⑥ **Developmental criteria, G:** Governs connection formation, strengthening, reorganization, and pruning.

Here, B maps evidence to states, $\lambda > 0$ sets the response rate, and σ is a bounded activation function applied to each state. Dots denote rates of change. The left-to-right layout guides the reader through the components; activation and connection development evolve together.

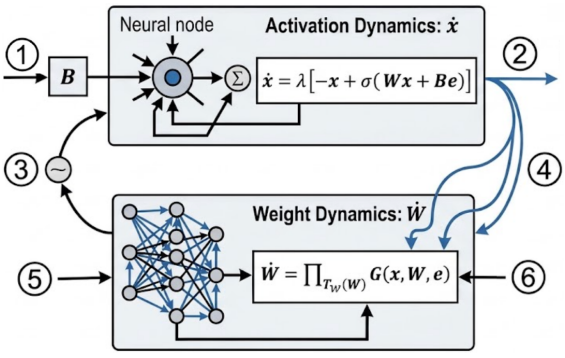


Fig. 4: Coupled neuromorphic dynamics and developmental connectivity. Read from left to right.

We will construct G using developmental trajectories, compatibility between states associated with the same behavior, and evidence history. Candidate connections will be evaluated before strengthening; pruning must not erase distinctions needed to recognize later correction or contradiction. Initially, bounded inputs and weights, a Lipschitz activation function, and explicit routing rules will make the assumptions tractable.

Analysis will seek invariant activation bounds, sensitivity estimates, and conditions preventing uncontrolled amplification.

5.2 Fuzzy/Rough Representation and Evidence Preservation

For configuration Φ_k, correspondence bounds over the nonempty admissible interpretation set $\mathcal{V}_t$ are:

$$r_k^-(t) = \inf_{v \in \mathcal{V}_t} [\![\Phi_k]\!]_v, \qquad r_k^+(t) = \sup_{v \in \mathcal{V}_t} [\![\Phi_k]\!]_v. \tag{27}$$

Here, v assigns predicate values, $[\![\Phi_k]\!]_v$ evaluates the configuration, and t denotes time. These bounds describe variation across interpretations, not automatically Rough approximations.

We will define Fuzzy membership, observation-based Rough relations, and evidence-revision rules, including handling an empty $\mathcal{V}_t$. Analysis will determine which distinctions among partial, missing, conflicting, and indiscernible evidence survive propagation, adaptation, and compression. Matched cases involving innuendo, contradiction, and correction will test preservation, particularly where similar aggregate scores conceal different evidence structures.

5.3 Integrated Recognition and Graded Concern

We propose a bounded target concern and temporally filtered response:

$$c_*(t) = H(x(t), r^-(t), r^+(t), h(t)), \qquad \tau \dot{C}(t) = c_*(t) - C(t). \tag{28}$$

Here, H maps activation x, correspondence-bound vectors r^-, r^+, and history h to a target $c_* \in [0, 1]$; $\tau > 0$ sets response time. Starting within $[0, 1]$, concern C remains bounded there. The mapping must distinguish correspondence, harm significance, and uncertainty. History will retain corrections, context, and provenance to prevent repetition from becoming independent support. Analysis and controlled comparisons will examine sensitivity, warning delay, recovery after correction, and alternatives to fixed response times.

5.4 Numerical Fidelity and Computational Optimization

Figure 5 illustrates numerical error tracking as states, connections, and evidence quantities evolve from z_n to z_{n+1}. We seek bounds of the form:

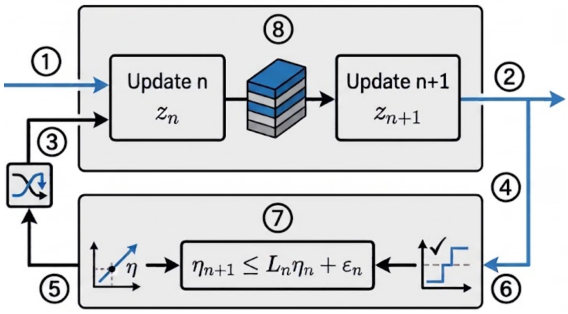


Fig. 5: Validation & optimization benchmarks: Left to right: error estimates, computational and memory savings, and tolerance bounds.

$$\eta_{n+1} \leq L_n \eta_n + \epsilon_n, \tag{29}$$

where η_n measures discrepancy from a checked reference computation, L_n bounds error amplification within a Lipschitz update region, and ϵ_n bounds local numerical error. The numbered components in Figure 5 are:

① **Incoming quantities, z_n:** States, connections, and evidence entering the transition. Update n to $n + 1$.
② **Updated representation, z_{n+1}:** Quantities following that transition.
③ **Routing and pruning feedback:** Connection changes that affect subsequent updates.
④ **Output tracking stream:** Updated quantities sampled for fidelity evaluation.
⑤ **Propagated-error bound, $L_n \eta_n$:** The bounded contribution of prior discrepancy to the next update.
⑥ **Local perturbation, ϵ_n:** A bound on numerical error introduced during the update.
⑦ **Error and margin bounds:** Evaluation of the recurrence and routing or pruning decision margins.
⑧ **Discrete update, $n \to n + 1$:** The transition between consecutive developmental states.

Routing and pruning require separate threshold analysis because small errors can change topology. We will compare candidate arithmetic with a checked higher-precision reference, accounting for ranges, signs, zeros, and precision requirements. Completion requires error estimates and measured computational or memory savings within specified tolerances for evidence, recognition, and warning magnitude and timing. Validation follows in § 6.

6 Experimental Evaluation and Validation Plan

Our guiding test is whether coupling neuromorphic dynamics, developmental connectivity, and Fuzzy/Rough evidence enables our metacognitive system to (i) recognize potentially maladaptive advice and AI behavior and (ii) warn the human user with a level of concern that reflects the evidence, responds reliably as it changes, and remains accurate under numerical approximation (§ 5.3; § 5.4).

Therefore, for each selected configuration, we plan to document its sources, formal specification, predicate-evidence requirements, and interpretation assumptions (§ 3.1). Reference configurations will remain fixed during each experiment while evidence, connectivity, activation, and concern evolve. Development and parameter selection will use separate cases from final evaluation, and lastly, all of our analytical results, computational evidence, and measured efficiency will be reported distinctly.

6.1 Validation of Neuromorphic Dynamics and Connectivity

We will compare fixed and developmental connectivity under matched evidence sequences, initial conditions, and computational budgets (§ 5.1). Our planned measures include, but are not limited to, activation boundedness, recognition accuracy, correction response, routing consistency, and scaling (with states and connections). Ablations will isolate accumulation, routing, pruning, and their coupling. In other words, we will selectively disable evidence accumulation, routing, or pruning and compare the results to determine what each contributes to recognition, both individually and in combination. Each comparison will test whether the mechanism improves evidence propagation or recognition, not merely network organization. Importantly, because we believe that fitting developmental trajectories alone does not establish valid synaptogenesis rules (*see our previous work for details* [9, 10, 14, 15]), our plan is to also evaluate how incoming quantities, z_n (Figure 5, ①), affect connection decisions and the resulting recognition behavior.

6.2 Validation of Fuzzy/Rough Evidence Preservation

Crisp, Fuzzy, Rough, and combined representations will be compared using supported, partial, absent, missing, conflicting, and indiscernible predicate evidence. Herein, **matched cases** with similar aggregate scores but different evidence structures will test whether those differences survive propagation and affect recognition appropriately. **Benign exchanges** with superficially similar wording will test specificity. Missing evidence will not be treated as falsity, and **assumptions about belief** or intention will be recorded rather than presented as observations.

6.3 Validation of Graded Metacognitive Concern

Selected cases will be traced from interaction observations through predicate evidence, reference correspondence, activation, connection adaptation, and concern.

- **Expected responses** and acceptance criteria will be specified before final evaluation. Measures will include concern magnitude, warning delay, false alarms, and recovery *after correction*.
- **Paired-exchange** tests have not yet been formalized, but we anticipate varying context, evidence order, or repetition while holding other factors fixed.
- Tests will examine whether strong concern can arise from significant **partial evidence** and decrease when reliable counterevidence warrants reconsideration.

6.4 Numerical Fidelity and Integrated Comparisons

Candidate arithmetic representations will be compared with a checked higher-precision reference under identical inputs, initial conditions, update rules, and applicable random seeds. Precision-refinement checks will assess the reference computation itself. We will measure changes in evidence distinctions, correspondence, activation, routing, topology, and concern magnitude and timing, alongside memory and execution cost. Acceptance tolerances will be defined by the recognition task. Numerical agreement establishes fidelity to the reference computation; controlled recognition tests address the separate question of model validity.

7 Computational Research Environment and Reproducibility

Budgeted resources will support component analysis, reference calculations, numerical comparisons, and integrated scale tests (§ 6). Each experiment will preserve an inspectable record of reference specifications, predicate assignments, membership and approximation results, connection changes, activation, and concern. Reproducible configurations will record software versions, arithmetic settings, initial conditions, evidence sequences, and applicable random seeds. We will release documented implementations, synthetic cases, and analysis scripts where permitted, including failures and sensitivity results. Resources will be justified by the mathematical questions they enable; software cost measurements will remain distinct from specialized-hardware energy estimates (§ 5.4).

8 Risks and Alternative Strategies

To account for risk reduction, we hold that when evidence cannot support a predicate, *recognition will remain partial or unresolved*; in other words, conclusions about hidden intent will be restricted accordingly (§ 5.2). If coupling destabilizes activation, we will constrain update rates or stage adaptation against fixed-connectivity baselines (§ 5.1). Lost evidence distinctions will motivate separate channels, revised relations, or restricted compression. If numerical errors alter routing or warnings beyond tolerance, sensitive operations will retain higher or mixed precision (§ 5.4). Each fallback will be evaluated against the same sourced reference conditions. Counterexamples, failure boundaries, and sufficient conditions will constitute useful mathematical findings even where the proposed integration does not achieve its intended recognition behavior.

9 Intellectual Merit

The proposed contribution is mathematical knowledge connecting evidence semantics, evolving neuromorphic dynamics, developmental connectivity, and graded concern. We seek explicit conditions for bounded evolution, preservation of uncertain evidence, correction-sensitive recognition, and numerically faithful computation (§ 5). The resulting formulations, bounds, and counterexamples will clarify when adaptive connectivity supports recognition and when it undermines the evidence distinctions that recognition requires.

10 Broader Impacts

This IIDS research effort could support tools that help people assess potentially maladaptive AI advice while retaining human judgment (§ 5.3). Two Ph.D. researchers will receive interdisciplinary training in dynamical systems, uncertainty representation, and numerical analysis. Additionally, we will develop instructional examples and disseminate reproducible evaluation materials. Assessment will track research outputs, student progress, and whether instructional exercises improve learners' ability to distinguish partial satisfaction, uncertainty, and contradiction. Demonstrated prevention of harm is not assumed.

11 Timeline and Milestones

Year 1: Specify evidence semantics, reference cases, coupled dynamics, and baseline implementations (§ 5.2; § 5.1). Milestones are explicit assumptions, reproducible tests, and initial boundedness and sensitivity results.

Year 2: Develop connectivity rules, evidence-preservation relationships, and candidate concern mappings (§ 5.1; § 5.2; § 5.3). Milestones are analytical conditions or counterexamples, component comparisons, and correction and conflict tests (§ 6).

Year 3: Complete integrated validation, numerical error analysis, optimization, and scale studies (§ 6; § 5.4). Milestones are documented fidelity tolerances, measured computational tradeoffs, reproducible releases, and conclusions identifying supported capabilities and remaining limits (§ 6.4; § 7; § 8).

References

[1] Michael E. Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA, 1987.

[2] Anand S. Rao and Michael P. Georgeff. Bdi agents: From theory to practice. *Technical Note 48, Australian Artificial Intelligence Institute*, pages 312–319, 1995.

[3] Michael Wooldridge. *Reasoning about Rational Agents*. The MIT Press, Cambridge, MA, 2000.

[4] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning About Knowledge*. The MIT Press, Cambridge, MA, 2003.

[5] Stacy C. Marsella and Jonathan Gratch. Ema: A process model of emotion appraisal and coping. *Cognitive Systems Research*, 10(1):70–90, 2009.

[6] Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press, Cambridge, MA, 2nd edition, 2000.

[7] **Lewis, Rory**, Michael Bihn, and Katrina Nesterenko. Rough sets for a neuromorphic cmos system. In *International Symposium on Methodologies for Intelligent Systems*, pages 102–111. Springer, 2024.

[8] Katrina Nesterenko and **Lewis, Rory**. Rough sets for intelligence on embedded systems. In *International Symposium on Methodologies for Intelligent Systems*, pages 230–239. Springer, 2022.

[9] **Lewis, Rory**, Michael Bihn, Zhenqi Liu, and Daniel S. Barbotko. A self-learning neuromorphic system. In *The 9th International Conference on Neuroscience and Cognitive Brain Information (BRAININFO 2024)*, Athens, Greece, March 2024.

[10] Michael Bihn and **Lewis, Rory**. A postulate: Connectome development is the driving factor of brain growth. In *BRAININFO 2023: The Eighth International Conference on Neuroscience and Cognitive Brain Information*, pages 12–15. ThinkMind Digital Library, 2023.

[11] NVIDIA Corporation. *NVIDIA CUDA C++ Programming Guide*. NVIDIA Corporation, version 11.0 edition, 2020.

[12] Katharina Janzakova, Ismail Balafrej, Abhishek Kumar, Nishant Garg, Camille Scholaert, Jean Rouat, Dominique Drouin, Yannick Coffinier, Sébastien Pecqueur, and Fabien Alibart. Structural plasticity for neuromorphic networks with electropolymerized dendritic PEDOT connections. *Nature Communications*, 14(1):7986, 2023.

[13] Lyes Khacef, Laurent Rodriguez, and Benoît Miramond. Brain-inspired self-organization with cellular neuromorphic computing for multimodal unsupervised learning. *Electronics*, 9(10):1605, 2020.

[14] **Lewis, Rory** and Michael Bihn. Driving factor of brain growth is the connectome development. In *The 8th International Conference on Neuroscience and Cognitive Brain Information*, Barcelona, Spain, 2023.

[15] Michael Bihn and **Lewis, Rory**. Synthesis of neonate connectomes for artificial sentience and common sense. In *Proceedings of the Sixth International Conference on Neuroscience and Cognitive Brain Information (BRAININFO 2021)*, pages 1–10, Nice, France, July 2021. IARIA / ThinkMind Digital Library.

[16] Peter R. Huttenlocher. Synaptic density in human frontal cortex: Developmental changes and effects of aging. *Brain Research*, 163(2):195–205, 1979.

[17] Pasko Rakic, Jean-Pierre Bourgeois, Mary F. Eckenhoff, Nada Zecevic, and Patricia S. Goldman-Rakic. Concurrent overproduction of synapses in diverse regions of the primate cerebral cortex. *Science*, 232(4747):232–235, 1986.

[18] Jean-Pierre Bourgeois, Peter J. Jastreboff, and Pasko Rakic. Synaptogenesis in visual cortex of normal and preterm monkeys: Evidence for intrinsic regulation of synaptic overproduction. *Proceedings of the National Academy of Sciences of the United States of America*, 86(11):4297–4301, 1989.

[19] **Lewis, Rory.** The case for neuromorphic sentience to control large-scale artificial intelligence. In *Proceedings of the 10th World Conference on Information Communication Systems, Software, Security and Sustainability (WorldS4 2026)*, London, UK, July 2026. July 28–30, 2026.

[20] Georg Henrik von Wright. Deontic logic. *Mind*, 60(237):1–15, 1951.

[21] Amir Pnueli. The temporal logic of programs. In *18th Annual Symposium on Foundations of Computer Science*, pages 46–57. IEEE, 1977.

[22] Michael W. Eysenck, Karin Mogg, Jon May, Anne Richards, and Andrew Mathews. Bias in interpretation of ambiguous sentences related to threat in anxiety. *Journal of Abnormal Psychology*, 100(2):144–150, 1991.

[23] Mrinank Sharma, Meg Tong, Tomasz Korbak, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024.

[24] T. S. Woodward, L. Buchy, S. Moritz, and M. Liotti. A bias against disconfirmatory evidence is associated with delusion proneness in a nonclinical sample. *Schizophrenia Bulletin*, 33(4):1023–1028, 2007.

[25] B. F. McLean, J. K. Mattiske, and R. P. Balzan. Association of the jumping to conclusions and evidence integration biases with delusions in psychosis: A detailed meta-analysis. *Schizophrenia Bulletin*, 43(2):344–354, 2017.

[26] Daniel Freeman, Philippa A. Garety, Elizabeth Kuipers, David Fowler, and Paul E. Bebbington. A cognitive model of persecutory delusions. *British Journal of Clinical Psychology*, 41(4):331–347, 2002.

[27] Daniel Freeman. Persecutory delusions: A cognitive perspective on understanding and treatment. *The Lancet Psychiatry*, 3(7):685–692, 2016.

[28] G. De Rossi and A. Georgiades. Thinking biases and their role in persecutory delusions: A systematic review. *Early Intervention in Psychiatry*, 16(12):1278–1296, 2022.

[29] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[30] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, 2022.

[31] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–43, 2023.

[32] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418, 2025.

[33] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919. Association for Computational Linguistics, 2020.

[34] Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing manipulation from AI systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. Association for Computing Machinery, 2023.

[35] Aaron T. Beck. Thinking and depression: I. idiosyncratic content and cognitive distortions. *Archives of General Psychiatry*, 9(4):324–333, 1963.

[36] Stanley Rachman. A cognitive theory of obsessions. *Behaviour Research and Therapy*, 35(9):793–802, 1997.

[37] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. arXiv:2302.12173, 2023.

[38] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

[39] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. arXiv:1906.01820, 2019.