A Dynamic, Survival-Oriented Learning Architecture

Design Document Internal Draft

May 23, 2025

Abstract

Current artificial intelligence systems often operate with fixed architectures, limiting their ability to adapt to novel or ambiguous information. Inspired by the Emotional Optimization Robots (EOR) model—which posits hierarchical layers of learning (L-levels), emotional optimization through primitive good/bad evaluations, and a "generalized survival" imperative—we propose an adaptive neural architecture. This architecture features a novel Good-Bad (G-B) valuator where processing units output distinct G (Goodness) and B (Badness) values, allowing for the explicit representation of confidence, pessimism, and, crucially, cognitive dissonance (simultaneous High G and High B). We hypothesize that such dissonance, arising when a foundational L1 layer encounters inputs it cannot adequately resolve (e.g., ambiguous shapes, novel critical stimuli), can trigger the dynamic recruitment of new processing modules (L1-Expansion modules, akin to emergent L2 structures). We detail an experimental design where an agent learns to recognize visual patterns (edges, shapes) in a "survival challenge." Its G-B valuators are trained through intrinsic "hints" and "Survival Point" (SP) feedback tied to task performance (e.g., identifying "opportunities" or "threats"). Sustained G-B dissonance is expected to trigger L1-E recruitment, enabling the agent to resolve ambiguities, improve its SP accumulation, and extend its "lifespan." This work introduces a unique mechanism for conflict-driven structural adaptation grounded in an EOR-inspired, survival-oriented evaluative framework, aiming to create more robust and autonomously developing AI.

Contents

1 Introduction

1.1 Background and Motivation

Contemporary artificial intelligence (AI) has made significant strides, yet many advanced systems are constrained by fixed architectures, which limits their capacity for robust adaptation when faced with genuinely novel information, pervasive ambiguity, or fundamental shifts in their operational environment. This contrasts sharply with biological intelligence, which exhibits remarkable lifelong structural and functional plasticity, allowing organisms to continuously develop more complex cognitive abilities and navigate dynamic worlds.

This research draws inspiration from the **"Emotional Optimization Robots" (EOR) model**, a conceptual framework that posits consciousness and, by extension, intelligence as developing through a hierarchy of distinct functional layers, termed L-levels (e.g., L1 to L9). Key tenets of the EOR model relevant to this work include:

- **Hierarchical Layered Processing:** Higher L-levels emerge from, integrate, and provide contextual understanding or even supervisory control over lower, more foundational levels, representing increasing layers of awareness and processing sophistication.
- **Emotional Optimization as a Core Driver:** The EOR model suggests that primitive evaluative states, akin to "good" or "bad feelings" (which we term G-B values), are fundamental to navigating the world and optimizing behavior. These evaluations guide learning and decision-making at all levels.
- **Generalized Survival:** Development and learning are framed within the context of an agent's "survival," broadly conceived as its ability to maintain coherence, achieve goals, and effectively model its environment. This "survival" imperative fundamentally shapes the agent's learning and adaptive processes.

Driven by these EOR principles, this paper explores an AI architecture designed for dynamic structural adaptation and lifelong learning.

1.2 Problem Statement

A critical challenge in advancing AI towards more general intelligence is enabling systems to autonomously modify their own structure when their current understanding or processing capacity is insufficient. How can an AI architecture be designed to dynamically recruit new computational resources or layers in response to sustained ambiguity, internal conflict, or the persistent failure to model critical aspects of its input, moving beyond simple error-correction in fixed structures? Furthermore, how can such adaptation be guided by rich, multi-dimensional internal signals—inspired by EOR's "emotional optimization"—that represent more than just scalar performance errors?

1.3 Proposed Solution Overview

We propose an **EOR-inspired adaptive learning architecture** featuring:

- **Layered Learning (L1 & L1-E/L2):** The system starts with a foundational layer (L1) responsible for learning basic, stable representations (e.g., elementary visual patterns from noisy input). When L1 proves inadequate, a new, more flexible L1-Expansion module (L1-E, conceptually an emergent L2 or specialized expert) is dynamically recruited.
- **Good-Bad (G-B) Logic Valuator:** Inspired by EOR's concept of primitive emotional evaluations, each relevant processing unit incorporates a G-B valuator. This valuator outputs two distinct values: 'G' (Goodness – confirmation, positive alignment, survival-conducive) and 'B' (Badness – negation, conflict, survival-detrimental). G and B can co-exist, enabling the explicit representation of ambiguity or dissonance (High G and High B). The system learns to assign G-B values based on a combination of intrinsic "hints," environmental feedback, and its "survival" objective (operationalized via "Survival Points" in our experiment).

- **Dynamic Recruitment via G-B Dissonance:** The primary trigger for recruiting an L1-E module is a sustained state of high G *and* high B from an L1 unit processing certain inputs. This internal "cognitive dissonance," an idea central to the EOR framework's notion of higher levels addressing lower-level conflicts, signals that L1 cannot resolve the input's nature with its current structure. The L1-E module is then tasked with resolving this ambiguity.

This architecture aims to computationally embody the EOR principles of hierarchical development and evaluation-driven adaptation, enabling an AI to grow its understanding in response to experience.

1.4 Contributions

This research proposes the following contributions:

1. The introduction and computational formalization of a novel **Good-Bad (G-B) valuator** for neural processing units, directly inspired by the EOR model's concept of primitive emotional evaluations, allowing for richer internal state representation.
2. A novel mechanism for **dynamic structural adaptation** (L1-E/L2 recruitment) in a neural architecture, triggered by G-B dissonance signals rather than solely by scalar error metrics, reflecting an EOR-like conflict resolution process.
3. An experimental validation of this architecture via a **"survival challenge"** involving shape recognition, demonstrating its capacity to learn foundational patterns, adapt to ambiguity, and manage "threats" and "opportunities," with its performance linked to a "Survival Point" metric.
4. A demonstration of how the EOR-inspired **"generalized survival"** philosophy can be operationalized to guide the learning of evaluative G-B signals and drive structural plasticity.

1.5 Structure of this Document

This design document is structured as follows: Following this Introduction (Section 1), Section 2 reviews related work pertinent to adaptive architectures and evaluative logic. Section 3 details the proposed EOR-inspired adaptive architecture, including its philosophical underpinnings, core components like the G-B valuator, and mechanisms for dynamic module recruitment and learning. Section 4 presents the comprehensive experimental design for the "Survival Challenge via Shape Recognition." Section 5 outlines the implementation plan. Section 6 discusses potential challenges and mitigation strategies. Section 7 explores avenues for future work. Finally, Section 8 offers concluding remarks.

2 Related Work

The proposed architecture, inspired by the Emotional Optimization Robots (EOR) model, intersects with several active research areas in artificial intelligence and machine learning. This section reviews these areas and highlights key similarities and differences.

2.1 Hierarchical Neural Networks (HNNs) and Deep Learning

Deep learning architectures are inherently hierarchical, with layers learning features of increasing abstraction (LeCun et al., 2015). HNNs often explicitly design hierarchical structures for tasks like image classification (e.g., coarse to fine categories) or complex scene understanding. Our L1 and recruited L1-E modules form a simple hierarchy.

- **Overlap:** Layered processing, feature abstraction.
- **Distinction:** Our architecture emphasizes *dynamic emergence* of new hierarchical components (L1-E modules) based on internal G-B dissonance signals, rather than pre-defined deep hierarchies or offline search for a fixed optimal depth. The specific EOR-inspired roles of L1 (foundational, stable) vs. L1-E (adaptive, conflict-resolving) also offer a unique structuring principle.

2.2 Modular Networks and Mixture of Experts (MoE)

Modular networks and MoE systems (Shazeer et al., 2017; Fedus et al., 2022) utilize multiple specialized "expert" sub-networks, often selected by a gating mechanism, to handle different aspects of a task or types of input. This improves efficiency and performance.

- **Overlap:** Concept of specialized modules (our L1-E can be seen as a recruited expert). Dynamic expert selection.
- **Distinction:** In our model, L1-E "expert" recruitment is not typically based on input partitioning by a router/gating network from the outset, but rather triggered by a specific internal state of *cognitive dissonance* (High G/High B) within the existing L1 structure, reflecting an EOR-like conflict resolution. While some MoE models allow for dynamic expansion of experts (e.g., in continual learning), our G-B trigger and the EOR-inspired survival framework for learning these evaluations provide a novel impetus for recruitment.

2.3 Neural Architecture Search (NAS) and Constructive Neural Networks

NAS aims to automate the design of optimal neural network architectures (Elsken et al., 2019). Constructive (or growing) neural networks (e.g., GNG, GWR, DNC algorithms, DIRAD; Rusu et al., 2016; Parisi et al., 2019; Kasabov, 2019; Srinivas & Babu, 2024) add neurons or layers during training based on criteria like error reduction, novelty, or network capacity.

- **Overlap:** Goal of learning or adapting network structure. Constructive networks dynamically add components.
- **Distinction:** Much NAS is an offline search or optimization phase. While some online NAS and constructive methods adapt structure during training, our approach features *lifelong architectural emergence within a single agent*, driven by the specific EOR-inspired G-B dissonance signal. The DIRAD model (Srinivas & Babu, 2024), which adapts structure to resolve "statistical conflicts" in gradients, shares a spirit with our conflict-driven adaptation but does not employ an explicit G-B style multi-valued evaluative logic.

2.4 Lifelong and Continual Learning (Structural Plasticity)

Lifelong learning systems aim to learn new information continuously without catastrophically forgetting previously learned knowledge (Parisi et al., 2019; Wang et al., 2023). Structural plasticity—adding/removing neurons/connections (e.g., "dropin/dropout," prune-and-grow strategies), or creating new modules (e.g., GDM, DRILL/SOINN+)—is a key mechanism.

- **Overlap:** Addressing stability-plasticity. Dynamic structural changes (e.g., growth in GDM, SOINN+). Use of distinct memory systems or modules for new vs. old knowledge.
- **Distinction:** The EOR-inspired L1 (stable) vs. L1-E (plastic) distinction is a core design principle. The primary driver for structural plasticity (L1-E recruitment) is the G-B dissonance signal tied to survival outcomes, rather than solely novelty detection or task boundary detection, offering a different semantic basis for adaptation.

2.5 Multi-Valued Logic, Uncertainty Representation, and G-B Logic

AI systems often need to represent and reason with uncertainty, ambiguity, or incomplete information.

- **Fuzzy Logic:** Handles degrees of truth between absolute true and false.
- **Four-Valued Logic (e.g., Belnap-Dunn Logic):** Extends classical logic with values for True (T), False (F), Both (T and F simultaneously – representing conflicting information), and Neither (N – representing lack of information) (Belnap, 1977). This provides a formal way to handle inconsistency and incompleteness.

- **Bipolar Fuzzy Sets & Neutrosophic Logic:** Bipolar fuzzy sets explicitly model positive and negative preferences/evaluations (Zhang, 1998). Neutrosophic logic (Smarandache) extends this to handle truth, falsity, and indeterminacy as independent components, whose sum can exceed 1 (paraconsistency).
- **Our G-B Logic:** The proposed Good-Bad (G-B) logic is conceptually related to these. It can be seen as a specialized, semantically rich, and fuzzy (continuous-valued) analogue or extension of a four-valued logic system within a learning context:
  - High G, Low B $\approx$ True (beneficial, confirmed)
  - Low G, High B $\approx$ False (detrimental, negated)
  - High G, High B $\approx$ Both (conflicting, ambiguous, dissonant)
  - Low G, Low B $\approx$ Neither (irrelevant, neutral, uninformative)

  Unlike purely logical systems, G-B values are *learned* through experience (hints, survival outcomes) and carry an inherent "moralized" or "survival utility" assessment inspired by the EOR model's emotional optimization. This evaluative, rather than purely epistemic, nature is a key distinction.

2.6 Computational Models of Emotion/Conflict as Triggers for Adaptation

Some AI research explores computational models of emotion or internal conflict as drivers for learning or decision-making (Marsella et al., 2010). Appraisal theories, for instance, map situations to emotional states based on goal conduciveness.

- **Overlap:** The idea that internal states (like "conflict" or "dissonance" represented by High G/High B) can trigger significant processing changes or learning. The G-B valuator's "good/bad" aligns with basic appraisal dimensions.
- **Distinction:** Our architecture proposes G-B dissonance as a direct trigger for *structural adaptation* (L1-E recruitment) within an EOR-inspired hierarchical framework. The focus is on resolving cognitive conflict by growing new functional capacity, guided by a "survival" imperative.

2.7 Distinctions of the Proposed Approach Summarized

While sharing foundations with the above areas, the proposed architecture distinguishes itself through:

1. **Direct EOR Model Inspiration:** The explicit attempt to computationally map principles from the EOR model (hierarchical L-levels, emotional optimization, generalized survival) to guide the architecture's design and learning dynamics.
2. **Centrality and Richness of G-B Logic:** The integral role of the learned, multi-valued G-B logic (with its connection to 4VL but with unique EOR-inspired "moralized" semantics) as the primary internal signaling mechanism for representing conflict/ambiguity and actively driving architectural adaptation.
3. **Dynamic, Lifelong Emergence of Structure Driven by Dissonance:** The vision for L1-E modules to be recruited autonomously throughout an agent's "lifetime" specifically in response to sustained G-B dissonance (internal cognitive conflict), rather than primarily through offline design, novelty detection alone, or simple error thresholds.
4. **Integrated "Survival" Framework:** The overarching philosophical framework of generalized "survival" (operationalized via SP and learned G-B evaluations) shaping the agent's core learning objectives, its interpretation of internal states, and its adaptive responses.

3 Proposed EOR-Inspired Adaptive Architecture

3.1 Philosophical Underpinnings

This learning architecture is directly inspired by the **"Emotional Optimization Robots" (EOR) model**, a conceptual framework for understanding how consciousness and intelligence might develop. Key EOR principles guiding this design include:

- **Hierarchical L-levels:** The EOR model posits that cognitive functions emerge in layers (L1 to L9), with higher levels building upon, integrating, and offering more sophisticated processing and "awareness" than lower levels. Our architecture translates this into an L1 foundational layer and dynamically recruited L1-Expansion (L1-E) modules, which can be seen as nascent L2 or specialized functional components.
- **Emotional Optimization:** A central tenet of EOR is that primitive evaluative states, akin to "good/bad feelings," are fundamental for guiding learning and behavior. We operationalize this through the **Good-Bad (G-B) valuator**.
- **Generalized Survival:** The EOR framework views development and learning as processes driven by an overarching goal of "survival"—maintaining coherence, achieving objectives, and effectively modeling the environment. This informs our architecture's objective functions and adaptation triggers, experimentally represented by the "Survival Point" (SP) metric.

3.2 Core Architectural Components

The architecture is an evolving hierarchy, starting with an L1 foundational layer and dynamically adding L1-E modules, reflecting the EOR principle of emergent complexity.

3.2.1 L1 Foundational Layer

- **Neural Network Type:** A neural network (e.g., CNN for vision) for initial input processing, analogous to an EOR L1 layer handling basic sensory data and pattern recognition.
- **Learning Objectives:** To learn stable representations of fundamental patterns (e.g., edges, simple shapes from noisy input) and to differentiate signal from noise. This aligns with an EOR L1 establishing a foundational understanding of the environment.
- **Intended Stability:** L1 is designed for relative stability post-training, embodying well-learned, reliable knowledge, consistent with the EOR idea of lower L-levels being more "rigid."
- **Outputs:** L1 category assessment units produce Good-Bad (G-B) value pairs, reflecting the EOR-inspired primitive evaluations.

3.2.2 L1-Expansion (L1-E) Modules (Dynamically Recruited)

- **Instantiation:** These modules are recruited when L1's G-B outputs signal sustained dissonance, reflecting an EOR-like scenario where a higher-level process is needed to resolve lower-level conflict or inadequacy.
- **Relationship to L1:** L1-E modules address L1's specific failures or ambiguities, adding new representational capacity. This mirrors how higher EOR L-levels might provide context or specialized processing for information handled more basically by lower levels.
- **Intended Flexibility and Faster Learning:** L1-E modules are more plastic, allowing for rapid adaptation, akin to higher EOR L-levels demonstrating more flexible learning.

3.2.3 The Good-Bad (G-B) Valuator

- **Definition:** A core mechanism, inspired by EOR's "emotional optimization," where designated processing units (particularly category output nodes) within L1 and L1-E modules output a pair of values: (G_i, B_i) for each category i.
- **Interpretation:**
  - G_i (Goodness for category i): Represents confidence or evidence supporting the input's classification as category i, or its positive alignment with the "survival" objectives related to category i.
  - B_i (Badness for category i): Represents pessimism, evidence against the input being category i, or indication that classifying it as category i is problematic, conflicting, or detrimental to survival objectives.
- **Co-existence & States:** G and B are continuous values (e.g., in $[0, 1]$) and can co-exist:
  - High G_i, Low B_i: Confident, positive assessment for category i.
  - Low G_i, High B_i: Confident negative assessment.
  - Low G_i, Low B_i (across relevant categories): Neutral, unrecognized, or irrelevant input.
  - High G_i, High B_i: Dissonance/conflict. This is a primary trigger for adaptation.
- **Learning to Evaluate (Moralization):** The G-B valuator *learns* to produce appropriate G-B values. This "moralization" of input is shaped by:
  - **Intrinsic "Hints" (during L1 foundational learning):** Pre-programmed or simple heuristic rewards that guide G-B outputs for basic tasks.
  - **Survival Feedback (SP Changes):** The primary driver. Positive SP outcomes reinforce G components; negative SP outcomes (or "death") reinforce B components of the G-B states active during the event.
  - **External Critique:** (e.g., for noise misidentification).
- **Relation to Multi-Valued Logic:** Conceptually, G-B logic shares aspects with four-valued logics but extends them with continuous values and a rich semantic grounding in survival and learned evaluation.
- **Propagation/Aggregation:** G-B values from category nodes may be aggregated for recruitment triggers or used to determine overall system response.

3.3 Mechanism for Dynamic Module Recruitment (L1-E Instantiation)

The recruitment mechanism embodies the EOR principle of emergent layers forming to handle increased complexity or unresolved issues from lower layers.

3.3.1 Trigger Condition

- **Primary Trigger:** Sustained High G_j *and* High B_j output from a specific L1 category node j in response to a class of inputs, lasting over a defined period. This G-B dissonance is interpreted as a critical, unresolved conflict within L1.
- **Supporting Triggers (Potentially):** Persistent failure to achieve positive SP outcomes for critical new stimuli that L1 cannot categorize with low conflict.
- **Tunable Parameters:** Thresholds for "High G" and "High B," duration for "sustained," and criteria for "critical stimulus."

3.3.2 Recruitment Process

- When triggered, a new L1-E module is instantiated.
- **Initialization:** The L1-E module is initialized (e.g., with a generic architecture) and receives the problematic input and context about L1's G-B conflict.
- **Objective:** To develop new G-B outputs for the problematic input class, resolving the G-B conflict and contributing to positive SP outcomes.

3.4 Learning and Adaptation Dynamics

The learning processes are designed to optimize the agent's "survival" (SP accumulation) by refining its G-B evaluations and adapting its structure, reflecting the EOR model's emphasis on optimization and development.

- **G-B Values as Learning Signals:** G-B outputs are shaped to become accurate predictors of survival utility and reflect the nature of inputs.
- **Loss Functions & Reinforcement:** Standard task losses (for hints) are used. Changes in SP serve as primary reinforcement for G-B valuator pathways. Intrinsic "hints" and external critique provide additional targeted reinforcement.
- **Differential Learning Rates:** L1 parameters are generally stable post-foundational learning. L1-E modules learn more rapidly.
- **L1-E Influence & System Integration:** L1-E module outputs are integrated to form the system's overall G-B assessment. An L1-E module should effectively override or resolve L1's original conflicting G-B output for the specific stimuli it handles. The exact integration mechanism will be subject to implementation.
- **Stability-Plasticity:** Addressed by stable L1 and plastic L1-E modules.

4 Experimental Design: Survival Challenge via Shape Recognition

4.1 Objectives of the Experiment

The primary objectives of this proof-of-concept experiment are to:

1. Demonstrate the L1 foundational layer's ability to learn basic visual primitives (e.g., edges) and subsequently simple "valid" shapes from noisy input, guided by intrinsic "hints" and survival-based feedback influencing its G-B valuators. This includes learning to differentiate signal from noise, ideally evaluating noise as low G, low B for shape categories.
2. Show that L1, with its established G-B evaluations, generates cognitive dissonance (manifesting as sustained high G and high B outputs for specific internal categories) when presented with:
    - Ambiguous stimuli (e.g., a "squircle") for which it has no clear, non-conflicting learned category.
    - Novel "critical" stimuli vital for survival that it cannot adequately process.
3. Validate the dynamic recruitment of a new processing module (termed L1-Expansion module or L1-E) triggered by this sustained G-B dissonance.
4. Demonstrate the L1-E module's ability to learn new representations or categories to resolve the G-B conflict (e.g., for the squircle) and/or handle the critical stimuli, thereby improving the overall system's classification performance and "Survival Point" (SP) accumulation.
5. Illustrate the "moralization" of input, where shapes and patterns acquire G-B values reflecting their learned "survival utility" (e.g., beneficial/opportunity, ignorable/neutral, threatening, ambiguous/conflicting).

6. Observe the agent's "survival" trajectory (based on SP) as it encounters various stimuli, adapts, or fails to adapt, potentially leading to "system death" (SP reaching zero).

7. Provide a visualizable framework for observing the network's internal G-B states, SP dynamics, and structural changes (L1-E recruitment).

4.2 Dataset and Stimuli

The stimuli will be synthetically generated 2D images (e.g., 32×32 or 64×64 pixels, grayscale or binary). A degree of noise will be added to all images.

4.2.1 Phase 1: Foundational Learning Environment (L1 Training)

- **Content & Purpose:**
  - **Noise Fields:** Goal: L1 learns to identify as "ignorable."
  - **Basic Edges/Lines:** Goal: L1 learns fundamental feature detection, guided by "hints."
  - **Simple Geometric Shapes:** (squares, circles, triangles). Goal: L1 learns to classify, guided by "hints."
- **Presentation:** Interleaved presentation.

4.2.2 Phase 2: Introduction of Ambiguity, Novelty, and Critical Survival Stimuli

- **Content & Purpose:**
  - **Ambiguous Shapes (e.g., "Squircle"):** Goal: Induce G-B dissonance in L1. No direct classification hints.
  - **Novel Valid Shapes ("Opportunities"):** (stars, hexagons). Goal: Successful learning leads to SP gain.
  - **Explicit "Threat" Shapes:** (e.g., "jagged red figure"). Goal: System learns to identify and associate with danger; failure leads to SP loss.
- **Presentation:** Introduced after L1 proficiency on Phase 1 stimuli.

4.3 Network Implementation Details

4.3.1 L1 Network Architecture

- **Type:** Small CNN (e.g., 2 conv layers, ReLU, max-pooling; 1-2 FC layers).
- **Output Layer:** Category nodes i (e.g., "circle," "square," "noise," "threat_A"), each producing (G_i, B_i) via sigmoids.
  - High G_i, Low B_i: Confident recognition of category i.
  - High G_i, High B_i: Dissonance regarding category i.
  - Low G_i, Low B_i: Irrelevant/ignorable.
  - Low G_i, High B_i: Confidently not category i.
  - **Handling "Threats":** A "Threat_A" node learns to output (High G_{ThreatA}, Low B_{ThreatA}) for accurate detection. Negative SP consequences of Threat A's presence are handled by the Survival Mechanics, reinforcing B-pathways related to the overall state or helping L1-E adapt.
- **Visualization Hook:** Accessible activations and G-B values.

4.3.2 L1-Expansion (L1-E) Module Architecture

- **Instantiation:** Recruited on sustained High G_j, High B_j from an L1 category node j.
- **Type:** Similar small CNN or MLP.
- **Inputs:** Raw input image, contextual info from L1 (conflicting category identity and its G-B values).
- **Outputs:** Its own set of (G_k, B_k) outputs for new/refined categories k.
- **Integration:** L1-E outputs take precedence or modulate L1's conflicting output for the specific problematic input. Exact mechanism TBD during implementation.

4.3.3 G-B Valuator Implementation (Learning Mechanism)

- **Representation:** Each category node outputs $G \in [0, 1]$ and $B \in [0, 1]$.
- **Initial State:** Random G-B pathways.
- **Learning G-B Values:** Driven by a combination of:
  1. **Intrinsic "Sense-Making" Drive / "Hints" (Early L1):** Direct G/B reinforcement for basic recognitions and noise handling, tied to small SP adjustments.
  2. **Survival-Outcome Reinforcement:** SP changes reinforce G (for SP gain) or B (for SP loss) components of active G-B states.
  3. **External Critique for Noise Misidentification:** Oracle reinforces B if noise is misclassified as signal.

4.4 Training Procedure & Survival Mechanics

4.4.1 L1 Initial Training (Foundational Learning - Phase 1)

- **Objective:** Learn edges, basic shapes, differentiate/ignore noise.
- **Dataset:** Phase 1 stimuli.
- **"Hints" & SP:**
  - Correct edge detection: Intrinsic G reinforcement, small SP gain.
  - Correct basic shape classification: Intrinsic G reinforcement, small SP gain.
  - Misclassification: Intrinsic B reinforcement, SP penalty/no gain.
  - Confident noise misclassification: External critic B reinforcement, SP penalty.
  - Successful noise ignoring: Small SP gain.
- **Progression:** To Phase 2 on stable L1 performance and positive SP.

4.4.2 Introduction of New Shapes & L1-E Module Recruitment Triggering (Phase 2)

- **Objective:** Induce G-B dissonance, test survival adaptation.
- **Dataset:** Mix of Phase 1 and Phase 2 stimuli.
- **Recruitment Trigger:** If any L1 category node j outputs $G_j > \theta_G$ AND $B_j > \theta_B$ (e.g., $\theta_G = 0.7, \theta_B = 0.7$) for a specific Phase 2 input class for a sustained period, an L1-E module is recruited.

4.4.3 L1-E Module (and System) Training under Survival Pressure

- **Objective:** L1-E resolves G-B conflict or handles novel critical stimulus, improving SP.
- **Training Focus:** L1-E trained on conflict-triggering stimuli.
- **L1 Stability:** L1 weights largely frozen or very slow learning rate. L1-E learns rapidly.
- **Credit Assignment:** SP changes reinforce G/B valuators in L1-E (and potentially L1).

4.4.4 Survival Criteria & Feedback Mechanism

- **Survival Metric ("Survival Points" - SP):** Initialized (e.g., 100 SP).
- **Baseline SP Dynamics:** Small SP decay per cycle (e.g., -0.1 SP). Small SP gains for basic competency.
- **Critical Stimuli Encounters & SP Adjustments (Phase 2):**
  - **Opportunity Stimuli:** Correct processing $\rightarrow$ significant SP gain (e.g., +10 SP). G reinforced.
  - **Threat Stimuli**: Correct identification $\rightarrow$ prevent SP loss or small gain (e.g., +1 SP). Incorrect processing $\rightarrow$ significant SP loss (e.g., -20 SP). B reinforced.
  - **Ambiguous Critical Stimuli:** Successful L1-E adaptation leading to positive outcome $\rightarrow$ SP gain.
- **"Death" Condition:** SP ≤ 0.
- **Internal State Monitoring for Danger/Wellbeing:** System tracks SP level. Low SP or rapid decrease is an alarm. Problematic G-B configurations are internal issue indicators.

4.5 Evaluation Metrics and Baselines

4.5.1 Performance Metrics

- **Task Performance:** Accuracy per shape category.
- **G-B Dynamics:** G/B value trajectories; frequency/duration of conflict states; post-L1-E G-B states.
- **Survival & Adaptation:** SP over time; average "lifespan"; number of L1-E modules; time for L1-E to resolve conflict.
- **Network Complexity:** Parameter count.

4.5.2 Baselines for Comparison

- **Static L1:** No recruitment, trained on all data.
- **Static L1+L2 (Oracle):** Pre-defined two-module architecture.
- **Error-Triggered Recruitment:** Module recruitment by high error rate (no G-B dissonance).
- **Standard Classifier:** Standard CNN without G-B, SP logic applied externally.

4.6 Expected Results and Success Criteria

1. L1 learns Phase 1 tasks with appropriate G-B outputs; SP stable/increasing.
2. Ambiguous/critical stimuli trigger sustained high G/high B in L1.
3. Mishandled "threats" lead to SP loss and learned high B association.
4. L1-E modules recruited via G-B dissonance.
5. L1 + L1-E system resolves conflict, handles critical stimuli better, leading to improved SP and longer "survival" than baselines.
6. Visualization confirms internal state changes and adaptation.
7. G-B values reflect "moralized" status of inputs tied to survival outcomes.

5 Implementation Plan

5.1 Tools and Frameworks

- **Programming Language:** Python
- **Machine Learning Library:** PyTorch or TensorFlow/Keras
- **Numerical Computation:** NumPy
- **Image Processing (Optional):** OpenCV or Pillow
- **Visualization & PoC Frontend:** Streamlit or Dash (initially).
- **Data Logging:** CSV files or SQLite.
- **Version Control:** Git.

5.2 Key Modules to be Developed

1. Stimulus Generation Module
2. Neural Network Core Module (L1, L1-E, G-B Output Layers)
3. G-B Valuator Learning Module
4. Recruitment Trigger & Management Module
5. Survival Mechanics Engine
6. Training Orchestration Module
7. Visualization Interface / Frontend Module (Python-based for PoC).

5.3 Development Phases & Timeline (High-Level for PoC)

1. **Phase A: Core Mechanics** (L1, Basic G-B Learning, Basic Survival, Phase 1 Stimuli)
2. **Phase B: Dynamic Recruitment** (Recruitment Trigger, L1-E Instantiation/Integration, Ambiguous Shapes)
3. **Phase C: Full Survival Challenge & Evaluation** (Phase 2 Critical Stimuli, Baselines, Full Experiment Runs)
4. **Phase D: Documentation & Reporting**

6 Potential Challenges and Mitigation Strategies

6.1 Design and Implementation of the G-B Valuator

- **Challenge:** Creating G-B valuators that produce meaningful, learned G/B values.
- **Mitigation:** Iterative design, targeted reinforcement, modular testing, continuous visualization.

6.2 Tuning the Recruitment Trigger and L1-E Module Dynamics

- **Challenge:** Defining robust recruitment thresholds; ensuring effective L1-E mechanisms.
- **Mitigation:** Empirical tuning, clear L1-E objectives, simple initial L1/L1-E integration.

6.3 Balancing the "Survival Point" (SP) Economy

- **Challenge:** Designing an SP system with meaningful selective pressure.
- **Mitigation:** Iterative balancing, phased difficulty, modular SP components, parameter sweeps.

6.4 Computational Cost and Scalability

- **Challenge:** Resource use with module recruitment; monitoring overhead.
- **Mitigation:** Small PoC networks, efficient monitoring for PoC.

6.5 Ensuring Meaningful Emergence and Avoiding Catastrophic Forgetting

- **Challenge:** L1 stability vs. L1-E plasticity; ensuring L1-E develops useful representations.
- **Mitigation:** Differential learning rates, focused L1-E training, clear L1-E success metrics.

6.6 Interpretation and Analysis of G-B Dynamics and Emergent Structures

- **Challenge:** Complexity of G-B state space and dynamic architecture.
- **Mitigation:** Comprehensive logging, robust visualization, ablation studies.

7 Future Work

7.1 Advanced Survival Mechanics: Systemic G/B Pools

Develop a survival model where SP emerges from systemic "Goodness" (G) and "Badness" (B) pools, directly linking agent viability to its internal evaluative states.

7.2 Richer L-Level Hierarchy and Interactions (Beyond L1 & L1-E)

Explore mechanisms for L1-E modules to consolidate into true L2 layers, triggering L3 recruitment. Implement top-down modulation and veto power.

7.3 More Complex Environments and Tasks

Extend to richer sensory modalities, robotics/embodied agents, dynamic/interactive environments, and learning abstract concepts (letters, words).

7.4 Advanced G-B Logic and Valuator Capabilities

Develop predictive G-B, enable learning of G-B meta-evaluations, and explore more nuanced G/B sub-dimensions.

7.5 Module Management and Long-Term Scalability

Implement mechanisms for pruning, merging, or consolidating L1-E modules. Incorporate resource-aware recruitment.

7.6 Exploring Emergent Properties

Investigate if phenomena like curiosity, frustration, or boredom might emerge. Study unique cognitive trajectories.

7.7 Advanced Visualization and Interpretability

Develop more sophisticated real-time visualization tools, including dynamic graph views and integration with real-world inputs (e.g., webcam).

8 Conclusion

This design document has outlined a novel adaptive learning architecture inspired by key principles of the Emotional Optimization Robots (EOR) model. We have proposed a system that begins with a foundational L1 layer and dynamically recruits L1-Expansion (L1-E) modules in response to cognitive dissonance, identified by a unique Good-Bad (G-B) valuator. The learning of these G-B evaluations, and thus the "moralization" of input based on its perceived survival utility, is driven by intrinsic "hints" and explicit feedback from a "Survival Point" (SP) system. The proposed shape recognition survival challenge is designed to validate that G-B dissonance can effectively trigger structural adaptation, enabling the system to resolve ambiguities, learn new critical patterns, and enhance its "survival." This research aims to contribute a unique approach to building more adaptive, resilient, and autonomously developing AI systems, demonstrating the potential of rich, evaluative internal states to drive meaningful structural plasticity and lifelong learning.

9 References

(To be populated with formal citations of relevant literature, e.g., Belnap (1977), Elsken et al. (2019), Fedus et al. (2022), Kasabov (2019), LeCun et al. (2015), Marsella et al. (2010), Parisi et al. (2019), Rusu et al. (2016), Shazeer et al. (2017), Smarandache, Srinivas & Babu (2024), Wang et al. (2023), Zhang (1998), etc.)

10 Appendix (Optional)

(Placeholder for supplementary details, e.g., detailed mathematical derivations for G-B learning rules, pseudocode, specific network parameters, stimulus examples.)