The Anti-Golem Dependency:

A Framework for Repairing the Foundations of Democracy

Samuel Leizerman

Author Note

The mathematical methods, framework, and implementation listed here are Patent Pending.

App. #63/811907

Abstract

We are witnessing democracy fail in real time—not through military coups or constitutional crises, but through the systematic collapse of our collective ability to think together. The limbic-cognitive dependencies that override democratic deliberation—recursive feedback loops, threshold collapse, and epistemic automation—reveal the architecture of breakdown and the mathematical pathways to resilience, challenging us to inscribe into democracy the א (one) thing that can return it from מת (death) to אמת (truth): Empathy.

The problem is not that people have become irrational, but that normal human cognitive architecture becomes systematically exploitable under modern information conditions. When cognitive resources are depleted, even rational agents predictably shift from evidence-based reasoning to emotion-driven inference through measurable threshold processes. These individual transitions aggregate into group-level epistemic collapse through recursive feedback loops that transform ordinary disagreement into existential threat perception.

This paper, Anti-Golem Dependency, is the first of a series of papers to explore a broader model called the Anti-Golem Model. This paper presents a theoretical framework that transforms democratic breakdown from a political problem into an engineering problem. The framework provides mathematical tools for understanding when cognitive systems fail, computational methods for predicting breakdown before it occurs, and systematic approaches for designing interventions that restore epistemic health while preserving democratic pluralism. By making these processes mathematically tractable through recursive hierarchical BART modeling and directed cyclic graph integration, the framework opens the possibility of precision intervention before collapse becomes irreversible.

The mathematical methods, framework, and implementation listed here are Patent Pending.

App. #63/811907

Keywords: democratic resilience, epistemic collapse, cognitive thresholds, recursive Bayesian modeling, political polarization, belief formation, computational social science, democratic institutions, patent pending, decision support,

The Anti-Golem Dependency: A Framework for Repairing the Foundations of Democracy

Democracy faces an unprecedented challenge that transcends traditional political analysis: the systematic breakdown of citizens' capacity for shared reasoning across ideological divides. From social media echo chambers to polarized news consumption, from conspiracy theories to epistemic closure, information ecosystems increasingly fragment along identity lines. Citizens who once engaged in productive disagreement now retreat into incompatible realities, where conflicting evidence triggers defensive responses rather than thoughtful reconsideration. This breakdown represents more than political polarization—it signals the failure of the cognitive infrastructure that makes democratic deliberation possible, challenging us to inscribe into democracy the א (one) thing that can return it from מת (death) to אמת (truth): Empathy.

The underlying mechanism is not irrationality but systematic exploitation of normal human decision-making processes. When information overload, emotional manipulation, and recursive social pressure exceed cognitive capacity, even rational agents predictably abandon evidence-based reasoning for emotion-driven inference. These individual cognitive transitions cascade into collective epistemic breakdown, transforming routine policy disagreements into existential identity conflicts that threaten democratic stability itself.

The Anti-Golem Model emerges from this analysis to offer both diagnostic precision and therapeutic intervention. By operationalizing cognitive empathy as a measurable system property rather than an abstract ideal, the framework enables early detection of epistemic breakdown while preserving the pluralistic foundations essential to democratic legitimacy. What follows is a mathematical architecture for understanding—and potentially reversing—the recursive processes that transform protective democratic institutions into destructive forces of division.

1: Limbic-Abductive Decision Architecture

The Anti-Golem draws from Jewish mystical tradition, where a Golem—a clay guardian animated by the word אמת (emet, 'truth')—must be destroyed when it becomes dangerous to those it was meant to protect. To deactivate it, the rabbi removes the first letter א (aleph), leaving מת (met, 'death'). This paper interprets this as the Golem having already removed its own aleph—its essential empathy and protective purpose—by becoming destructive. The rabbi merely completes what the Golem began: the transformation from truth-guided protection to mindless destruction.

The Golem is liberal democracy—an ideology rooted in protective purpose, born not just out of rational self-interest but also cognitive empathy. Democracy is government of the people, for the people, by the people. When our governments become irrational and unempathetic, it is because we have become irrational and unempathetic. But to save our democracy from death, we must restore truth... Whose truth? Each other's truth—but here lies the fundamental challenge that the Anti-Golem framework addresses. How do we ensure our ability to see each other's truth? How do we protect cognitive empathy? **To answer these questions, we must first understand why cognitive empathy breaks down even among well-intentioned people.**

The Rashomon effect, named after Akira Kurosawa's film depicting the same event through multiple conflicting perspectives, reveals how subjective experience shapes perception of objective reality. In democratic discourse, information imbalances amplify this effect exponentially. When citizens operate from fundamentally different information ecosystems—different news sources, social media algorithms, peer networks—they don't simply disagree about solutions; they inhabit different factual universes. Each group experiences their perspective as obviously true while viewing opposing perspectives as obviously false or malicious.

These information asymmetries create what we might call "empathy shadows"—blind spots where citizens cannot comprehend how reasonable people could possibly hold opposing views. Without shared informational foundations, cognitive empathy becomes nearly impossible. Citizens lose the capacity to model how others think, reducing complex human perspectives to caricatures of stupidity or evil.

The Anti-Golem framework recognizes that restoring democratic truth requires not imposing a single perspective, but rebuilding the cognitive infrastructure that enables mutual understanding across perspective differences. To save our democracy, we must save cognitive empathy—the capacity to understand how others reach their conclusions, even when we disagree with those conclusions.

The death of human empathy is one of the earliest and most telling signs of a culture about to fall into barbarism. - Hannah Arendt

1.1 Theoretical Foundation

The mathematical framework and implementation methods described in this work patent pending. App# 63/811907

Classical decision theory assumes agents possess complete preference orderings and unlimited computational capacity. However, this framework introduces a novel limbic-abductive decision architecture that addresses the reality of bounded rationality through neurologically-grounded mechanisms. Neurological evidence demonstrates that decision-making under uncertainty involves a two-stage process: exhaustive rational analysis followed by limbic system engagement when cognitive resources are depleted (Damasio, 1994; Bechara et al., 1997).

This neurological foundation aligns with broader theories of embodied cognition (Anderson, 2007) and predictive processing frameworks (Clark, 2013; Friston, 2010) that demonstrate how cognitive systems continuously update beliefs based on prediction errors and environmental feedback.

This paper develops a mathematical framework for modeling the transition between rational optimization and what we term limbic-mediated abductive inference. Unlike traditional bounded rationality models that treat cognitive limitations as noise or approximation errors, this approach recognizes limbic engagement as a systematic and predictable decision mechanism with measurable parameters.

Let D represent the decision space and E the available evidence set. For any decision point d ∈ D, we define the cognitive exhaustion threshold $\tau(E,d)$ as:

$$\tau(E,d) = \sup\left\{t : \frac{\partial U(E,t)}{\partial t} > \delta\right\}$$

where $U(E,t)$ represents utility gained from additional analysis time t, and δ is the minimum meaningful utility increment. Beyond τ, additional rational analysis yields diminishing returns approaching zero.

1.2 Limbic Engagement and Biased Random Walks

When t > τ(E,d), decision-making transitions from rational optimization to limbic-mediated abductive inference. This framework's key innovation is recognizing that this transition exhibits characteristics of a biased random walk with systematic, measurable parameters:

$$X_{\{t+1\}} = X_t + \mu(\eta, S_t) + \sigma(S_t)\varepsilon_t$$

where:

- X_t represents belief state at time t
- $\mu(\eta, S_t)$ is the drift parameter conditional on prior beliefs η and salience vector S_t
- $\sigma(S_t)$ captures uncertainty scaling with salience
- $\varepsilon_t \sim N(0,1)$ represents stochastic innovations

The drift parameter $\mu(\eta, S_t)$ incorporates Fundamental Attribution Bias (FAB) through: $\mu(\eta, S_t) = \alpha_0 + \alpha_1 \cdot FAB(S_t) + \alpha_2 \cdot \eta_{t-1} + \alpha_3 \cdot (S_t \otimes \eta_{t-1})$

This formulation explains why rational agents can systematically deviate from optimal choices: once cognitive resources are exhausted, biologically-evolved heuristics dominate decision-making. The innovation presented here is the mathematical operationalization of this transition as a measurable threshold process rather than treating it as random error.

1.3 Neurological Foundation and Empirical Support

The limbic-abductive framework developed in this work builds on established neurological findings while extending them into formal mathematical structure. Patients with ventromedial prefrontal cortex lesions demonstrate intact rational analysis capabilities but cannot make decisions when multiple options appear equally valid—precisely the scenario where our model predicts limbic system engagement becomes crucial (Bechara et al., 1997; Damasio, 1994).

This paper's Bayesian updating as temporal structure encoder paradigm extends this neurological foundation by modeling how prior beliefs $\log P(b_{i,t-1})$ influence current limbic processing. The temporal encoding innovation presented here treats belief updating not as static Bayesian revision but as accumulated temporal differentials: Issue is resolved by appending a time index and a loop counter to the variables. The loop counter merely need to be coded to turn on upon crossing the entrance path of the gate and then back off upon crossing the exit path.

1.4 Integration with Recursive Bayesian Gates

This framework's recursive Bayesian gate modeling for epistemic collapse innovation directly builds on the limbic-abductive foundation. The three-gate system (RTT_1, RTT_2, IRD) represents different stages of limbic processing under increasing cognitive load:

Gate 1 (RTT_1): Cognitive Exhaustion Detection

- If $t > \tau(E, d)$: Engage limbic processing
- Else: Continue rational analysis

Gate 2 (RTT_2): Limbic Resolution

- If $P(b_i) > \theta$: Commit to belief (exit probabilistic superposition)
- Else: Maintain uncertainty

Gate 3 (IRD): Social Integration

- Aggregate individual limbic outcomes into group-level patterns
- $IRD = \left(\frac{1}{n}\right) \sum RTT_{2_{outcomes}}$

The recursive nature emerges because limbic processing outcomes at time t influence cognitive exhaustion thresholds at time $t + 1$, creating the temporal feedback loops central to the Anti-Golem Model.

1.5 Salience-Weighted Decision Framework

This work introduces salience-weighted SoftMax dynamics that formally integrate limbic processing with traditional choice models. Unlike standard SoftMax formulations that assume fixed utility functions, this framework's innovation allows salience parameters to evolve based on limbic processing outcomes:

$$P(b_i) = \frac{\mathrm{e}^{\Psi(b_i)/\tau}}{\sum_j \mathrm{e}^{\Psi(b_j)/\tau}}$$

where: $\Psi(b_i) = \alpha \cdot \lambda_i + \beta \cdot \sigma_i + \gamma \cdot V_i + \log P\left(b_{i,t-1}\right)$

The λ_i (identity salience) and σ_i(external threat salience) parameters are endogenously determined through limbic processing rather than assumed constant. This represents a paradigm shift from treating salience as exogenous to modeling it as an emergent property of cognitive-limbic interaction.

1.6 Implications for Democratic Resilience

When cognitive exhaustion becomes widespread—due to information overload (Eppler & Mengis, 2004), elevated threat salience (Huddy et al., 2005), information imbalances (Sunstein, 2017), crisis fatigue (Moeller, 1999), deliberate manipulation (Howard & Hussain, 2013), or algorithmic manipulation (Zannettou et al., 2019)—entire populations may shift from rational deliberation to limbic-driven decision-making simultaneously.

This switch from rational to a more emotionally driven process framework's hypothesis connects to Hannah Arendt's analysis of how totalitarian movements systematically undermine citizens' capacity to distinguish fact from fiction, truth from falsehood—what she identified as the ideal conditions for totalitarian control (Arendt, 1951). While Arendt focused on deliberate political manipulation, this framework suggests multiple pathways can achieve similar epistemic breakdown through cognitive exhaustion.

Contemporary media ecosystems increasingly leverage emotional responses to drive engagement (Brady et al., 2020)—from news outlets prioritizing triggering content (Robertson et al., 2023) and clickbait headlines (Jung et al., 2022) to algorithmic amplification of divisive content (Shin & Jitkajornwanich, 2024) and influencer speech patterns designed for doom scrolling (Marciano et al., 2022)—creating systematic pressure on cognitive resources. These mechanisms, whether intentional or emergent from market incentives, can produce the same loss of truth-discernment capacity that Arendt observed totalitarian movements deliberately cultivate, making populations vulnerable to epistemic collapse through accumulated cognitive exhaustion rather than coordinated political strategy.

This framework's key insight is that this transition is not pathological but represents normal human cognitive architecture operating under stress. The innovation lies in making this process mathematically tractable and empirically measurable, enabling prediction and intervention before limbic processing leads to epistemic collapse.

The temporal structure encoding approach introduced here allows tracking of how accumulated cognitive exhaustion builds system-wide vulnerability, while the recursive Bayesian gates provide specific intervention points where rational processes can be restored or limbic processing can be guided toward prosocial outcomes.

1.6.1 Moral Panic Integration Section

Contemporary Moral Panic Dynamics and Democratic Cognitive Resilience

Recent research demonstrates how digital technologies have fundamentally altered moral panic dynamics, creating new pathways for the cognitive exhaustion processes central to the Anti-Golem Model. Walsh (2020) shows that social media platforms "unleash and intensify collective alarm" through mechanisms of amplification and participation that extend beyond traditional mass media models. These technological transformations enable "crowd-sourced panics" where ordinary citizens can initiate and propagate moral crusades, while algorithmic systems amplify emotional content in ways that accelerate the transition from concern to hostility to consensus formation.

Contemporary moral panic research validates the Anti-Golem Model's core predictions about how cognitive overload leads to systematic threshold-crossing behavior. Ben-Yehuda and Goode's (1994) foundational framework identifies five elements that define moral panic episodes: concern, hostility, consensus, disproportionality, and volatility. These elements map directly onto the model's gate parameters, suggesting that moral panics represent observable instances of population-level epistemic collapse through the recursive feedback loops central to democratic breakdown.

The systematic amplification of threat perception through emotionally charged rhetoric creates widespread cognitive exhaustion, triggering transitions from rational policy evaluation to identity-protective reasoning. Legislative responses often exhibit the characteristic disproportionality of epistemic collapse—policy interventions that exceed reasonable assessment of actual risk while generating recursive feedback loops through media amplification and political mobilization (Rothe & Muzzatti, 2004). Analysis of contemporary policy conflicts reveals how these dynamics operate: when populations shift from evidence-based reasoning to emotion-driven decision-making, democratic institutions become vulnerable to exploitation through the same cognitive architecture that the Anti-Golem Model mathematically operationalizes.

This connection between moral panic research and the Anti-Golem Model's theoretical framework suggests that democratic breakdown often follows predictable patterns of cognitive overload rather than random political failure. Understanding these patterns as systematic threshold processes rather than mysterious social phenomena opens possibilities for early warning systems and targeted interventions before collapse becomes irreversible.

2: Recursive Hierarchical BART (RH-BART) Formalization

2.1 Mathematical Framework

Hierarchical Decomposition

The RH-BART framework decomposes the Anti-Golem Model's gate parameters across three hierarchical levels, preserving the theoretical structure while allowing empirical discovery of functional relationships. This approach builds on established hierarchical BART methodologies (Tan et al., 2018; Sparapani et al., 2021) while introducing novel recursive parameter extraction mechanisms to address the unique computational challenges arising from the mathematical structure of the Anti-Golem framework.

This framework prioritizes **systemic integration** over exhaustive methodological rigor. The goal is to demonstrate sufficient empirical grounding to justify the theoretical foundations and warrant further specialized research, rather than to provide definitive parameter estimates or causal claims.

Modern computational approaches increasingly leverage transformer architectures **(Devlin et al., 2018)** and local polynomial methods **(Fan & Gijbels, 1996)** for pattern discovery, providing the technical foundation for the parameter extraction mechanisms described below.

Level 1: Individual Gate Parameter Estimation

$$\lambda_{ij} = \mathrm{BART}_1^{\alpha}(X_{ij}) + u^{1}_{ij} \qquad \text{\# Identity-weighted salience}$$
$$\sigma_{ij} = \mathrm{BART}_1^{\beta}(X_{ij}) + u^{2}_{ij} \qquad \text{\# External threat salience}$$
$$V_{ij} = \mathrm{BART}_1^{\gamma}(X_{ij}) + u^{3}_{ij} \qquad \text{\# Perceived utility}$$
$$\theta_{ij} = \mathrm{BART}_1^{\theta}(X_{ij}) + u^{4}_{ij} \qquad \text{\# Collapse threshold}$$

Where $\mathbf{X}_{\text{ij}}$ includes individual characteristics (demographics, personality traits, prior exposures) and environmental variables (media exposure, social network composition, geographic context, institutional factors). BART naturally handles categorical predictors through binary splitting rules (Chipman et al., 2010).

Theoretical Gate Integration:

These estimates feed directly into the Anti-Golem Model's core equations:

$$\Psi(b_{\text{ij}}) = \alpha \cdot \lambda_{\text{ij}} + \beta \cdot \sigma_{\text{ij}} + \gamma \cdot V_{\text{ij}} + \log P({b_{\text{ij}}}^{\text{t}-1})$$

$$P(b_{\text{ij}}) = \frac{\mathrm{e}^{\Psi(b_{\text{ij}})/\tau}}{\Sigma_{\text{k}} \mathrm{e}^{\Psi(b_{\text{kj}})/\tau}}$$

$$Collapse\ if\ P(b_{\text{ij}}) > \theta_{\text{ij}}$$

Level 2: Group Aggregation

$$IRD_{\text{j}} = \left(\frac{1}{n_{\text{j}}}\right) \Sigma_{\text{i}}\, RTT^2{}_{\text{ij}} + BART^2(Z_{\text{j}}, u^{1-4}{}_{\text{ij}}) + v_{\text{j}}$$

Where $\mathbf{Z}_{\text{j}}$ represents group-level characteristics and residuals from Level 1 capture cross-level interactions, following established multilevel modeling principles (Raudenbush & Bryk, 2002).

Level 3: System-Level Pattern Discovery

$$\Phi_{\text{k}} = BART^3(S_{\text{ik}}, S_{\text{ok}}, C_{\text{ik}}, C_{\text{ok}}, W_{\text{k}}, v_{\text{j}}) + \eta_{\text{k}}$$

Where crisis coefficients **C** and salience ratios **S** interact through empirically-discovered functional forms, allowing for context-dependent nonlinear relationships without a priori specification.

Methodological Genealogy: RH-BART as Computational Grounded Theory

The RH-BART framework operationalizes core principles of grounded theory through computational means, addressing methodological challenges that emerged during the author's empirical research (Leizerman, 2024). Grounded theory's strength lies in its systematic approach to theory building: emergent categorization, constant comparative analysis, theoretical sampling, and iterative refinement without a priori theoretical imposition (Glaser & Strauss, 1967; Charmaz, 2006). However, grounded theory's reflexivity requirement—that researchers continuously examine how their own assumptions and biases shape emerging theory—creates both analytical depth and methodological constraints.

RH-BART addresses grounded theory's reflexivity challenge not through enhanced self-awareness, but through algorithmic elimination of researcher bias in pattern discovery. Where grounded theory requires reflexive management of human interpretation, RH-BART removes human cognitive filters entirely from the initial discovery process. BART's recursive partitioning discovers emergent patterns without theoretical preconceptions, while the hierarchical structure enables constant comparative analysis across individual, group, and system levels. The recursive parameter extraction mirrors grounded theory's iterative refinement, allowing theoretical categories to emerge from empirical relationships rather than researcher assumptions.

This creates a complementary methodological toolkit: RH-BART provides bias-free, reproducible pattern discovery at scale, while traditional grounded theory remains essential for contextual meaning-making and interpretive depth. Researchers can leverage RH-BART for population-level parameter discovery, then employ grounded theory approaches to understand the contextual significance of discovered patterns. Rather than replacing qualitative methodology, RH-BART extends it—handling what algorithms do best (pattern detection) while preserving space for what humans do best (meaning-making).

2.2 Temporal Framework and Recursive Parameter Extraction

Belief State as Temporal Accumulation

The framework treats current belief states as accumulated temporal differentials, where $P(b_{\mathrm{i}}^{\mathrm{t}}) = P(b_{\mathrm{i}}^{0}) + \Sigma[k = 1\ to\ t]\Delta P(b_{\mathrm{i}}^{\mathrm{k}})$. This temporal calculus approach transforms the model from static state estimation to dynamic change mechanism modeling, where each differential ΔP captures evidence effects, social influence, decay processes, and skill check outcomes between agent resistance and evidence power.

The resilience integral $R = \int_0^T \left|\frac{1}{E_b(t)}\right| dt$ $\overline{\mathrm{P}}(\mathrm{H})$ is the time-varying, system-wide average

posterior belief,

$\frac{d\bar{P}(H)}{dt}$ its instantaneous rate of change, $E_b(t) \equiv \left|\frac{d\bar{P}(H)}{dt}\right|$ represents elasticity, and the reciprocal weight $\left|\frac{d\bar{P}}{dt}\right|^{-1}$ becomes a summation over discrete time differentials, measuring accumulated resistance to belief flexibility recovery. This temporal structure naturally handles path dependence, intervention effects, and stability analysis through the differential sequence.

Cross-Level Information Flow

Upward Recursion: Individual parameters aggregate to inform group-level patterns

$$E[IRD_{\mathrm{j}} \mid \lambda_{\mathrm{ij}}, \sigma_{\mathrm{ij}}, \theta_{\mathrm{ij}}] = g(\Sigma_{\mathrm{i}}\, \lambda_{\mathrm{ij}}/n_{\mathrm{j}}, \Sigma_{\mathrm{i}}\, \sigma_{\mathrm{ij}}/n_{\mathrm{j}}, \Sigma_{\mathrm{i}}\, \theta_{\mathrm{ij}}/n_{\mathrm{j}})$$

Downward Recursion: System-level pressure influences individual gate parameters

Parameter Extraction from Nonparametric Discovery

The RH-BART framework addresses the threshold observability problem through a two-stage process: nonparametric discovery followed by parametric extraction. This approach leverages BART's ability to discover functional relationships without prior assumptions, then uses established curve-fitting techniques to extract specific threshold parameters for theoretical implementation.

Stage 1: Nonparametric Pattern Discovery

BART's recursive partitioning naturally identifies threshold-like behavior through its tree structure without requiring pre-specified functional forms (Chipman et al., 2010). When belief commitment probability changes sharply at specific covariate values, BART captures these discontinuities through its splitting rules. The ensemble of trees produces smooth estimates of the underlying function while preserving sharp transitions that indicate threshold behavior (Hill, 2011; Sparapani et al., 2021).

Unlike parametric approaches that assume specific threshold locations, BART allows the data to reveal where behavioral transitions occur. This addresses what Hastie et al. (2009) identify as the fundamental challenge in threshold modeling: distinguishing genuine discontinuities from smooth but steep transitions.

Stage 2: Parametric Threshold Extraction

Once BART identifies regions of sharp behavioral change, established curve-fitting procedures extract specific threshold parameters. Following the approach outlined in Müller & Rønn (1993) for threshold estimation in nonlinear time series, we fit parametric functions to the BART-discovered relationships:

For cognitive exhaustion thresholds: Parameter Extraction from Nonparametric Discovery

The RH-BART framework addresses the threshold observability problem through a two-stage process: nonparametric discovery followed by parametric extraction. This approach leverages BART's ability to discover functional relationships without prior assumptions, then uses established curve-fitting techniques to extract specific threshold parameters for theoretical implementation.

Stage 1: Nonparametric Pattern Discovery

BART's recursive partitioning naturally identifies threshold-like behavior through its tree structure without requiring pre-specified functional forms (Chipman et al., 2010). When belief commitment probability changes sharply at specific covariate values, BART captures these discontinuities through its splitting rules. The ensemble of trees produces smooth estimates of the underlying function while preserving sharp transitions that indicate threshold behavior (Hill, 2011; Sparapani et al., 2021).

Unlike parametric approaches that assume specific threshold locations, BART allows the data to reveal where behavioral transitions occur. This addresses what Hastie et al. (2009) identify as the fundamental challenge in threshold modeling: distinguishing genuine discontinuities from smooth but steep transitions.

Stage 2: Parametric Threshold Extraction

Once BART identifies regions of sharp behavioral change, established curve-fitting procedures extract specific threshold parameters. Following the approach outlined in Müller & Rønn (1993) for

threshold estimation in nonlinear time series, we fit parametric functions to the BART-discovered relationships:

For cognitive exhaustion thresholds: $\hat{\tau} = arg \min_{\tau} \Sigma_i \left(BART_{output}\,(x_i) - f(\tau, x_i)\right)^2$

For belief commitment thresholds: Logistic curves fitted to BART predictions enable extraction of inflection points corresponding to θ parameters (Loader, 1999).

This two-stage approach follows the general framework established by Fan & Gijbels (1996) for local polynomial fitting after nonparametric discovery, adapted for threshold detection in cognitive modeling contexts.

Integration with Theoretical Framework

The extracted parameters then feed directly into the Anti-Golem Model's equations, providing the measurable thresholds required for early warning systems and intervention design. This methodology addresses the circularity problem identified by Hansen (2000) in threshold modeling by using data-driven discovery rather than assumption-based specification.

Validation follows the protocol established by Li & Racine (2007) for mixed parametric-nonparametric models: BART discovery on training data, parameter extraction through curve fitting, and performance evaluation on held-out test data to assess threshold stability and transferability.

$$E[\lambda_{\mathrm{ij}}{}^{\mathrm{t+1}} \mid \Phi_{\mathrm{k}}{}^{\mathrm{t}}] \; = \; h(\lambda_{\mathrm{ij}}{}^{\mathrm{t}}, \Phi_{\mathrm{k}}{}^{\mathrm{t}}, System\_Feedback_{\mathrm{k}}{}^{\mathrm{t}})$$

This recursive structure extends standard hierarchical BART (Hill, 2011) by incorporating temporal feedback mechanisms between hierarchical levels.

2.3 Estimation Strategy

The estimation strategies were identified through AI-assisted literature review and selected based on their suitability for handling the unique challenges of recursive hierarchical parameter dependencies. The computational approaches were chosen to address the specific mathematical constraints imposed by the Anti-Golem framework's gate structure.

MCMC Implementation:

Following established BART computational frameworks (Chipman et al., 2010; Pratola, 2016):

1. **Gibbs sampling** for BART tree structures at each level
    - Required because the framework has **4 different BART models per level** (λ, σ, V, θ) that need joint estimation as they feed into the same theoretical equations
    - Tree structure updates via birth/death/change proposals (Chipman et al., 2010)
    - Leaf parameter updates via conjugate normal-inverse gamma priors

   - Allows **conditional sampling** where trees for λ are updated given current estimates of σ, V, θ
2. **Metropolis-Hastings** for cross-level interaction parameters
   - Necessitated by non-standard conditional dependencies where **residuals from Level 1** enter **Level 2** as predictors
   - Proposal distributions adapted from Tan et al. (2018) for multilevel contexts
   - Adaptive scaling following Roberts & Rosenthal (2009)
   - Handles **complex conditional distributions** through proposal mechanisms that respect cross-level constraints
3. **Adaptive shrinkage** priors for hierarchical residuals
   - Half-Cauchy priors on variance components (Gelman, 2006)
   - Regularized horseshoe priors for sparse cross-level effects (Piironen & Vehtari, 2017)
   - Prevents **overfitting** cross-level relationships while allowing genuine interactions to emerge

Convergence criteria:

- **Gelman-Rubin diagnostic** < 1.1 across all parameters (Gelman & Rubin, 1992)
- **Effective sample size** > 1000 for key gate parameters (Geyer, 1992)
- **Monte Carlo standard error** < 0.05 × posterior standard deviation

2.4 Validation Framework

Validation 1: Level-Specific Comparison

$$For\ each\ level\ k: Compare\ BART_k performance\ vs. Standard\ BART_k$$
$$Metrics: RMSE, Coverage\ probability, Variable\ importance\ rankings$$

Following cross-validation procedures established in Chipman et al. (2010) and extended validation metrics from Hill (2011).

Validation 2: Parameter Recovery Analysis

$$Generate\ synthetic\ data\ with\ known\ \lambda, \sigma, V, \theta\ parameters$$
$$Assess: \left|\left|\hat{\theta}_{RH} - BART\ -\ \theta_{true}\right|\right|^2 vs. \left|\left|\hat{\theta}_{Standard} - \theta_{true}\right|\right|^2$$

Simulation framework adapted from Green & Kern (2012) for treatment effect heterogeneity assessment.

Validation 3: Cross-Subgroup Transferability

$$Train\ RH - BART\ on\ Subgroup\ A\ \rightarrow\ Predict\ Subgroup\ B\ outcomes$$
$$Compare\ transferability\ vs. standard\ hierarchical\ models$$

External validity assessment following Dehejia & Wahba (2002) principles for causal inference transferability, with acknowledgment of multiple comparison considerations.

2.5 Implementation Requirements

Data Collection and Sampling Design

Sampling design must preserve natural group size distributions to validly capture power dynamics central to the theoretical framework. Standard proportional sampling techniques apply, with precision varying across groups in proportion to their real-world representation—which is theoretically appropriate given the model's focus on how group size affects recursive dynamics.

Sample size requirements follow standard power analysis procedures for each hierarchical level (Raudenbush & Bryk, 2002), with data collection following established complex survey design principles, including stratified sampling across hierarchical levels.

Methodological Development Priorities

Parameter operationalization requires development of validated scales for measuring identity salience (λ_i), threat salience (σ_i), and perceived utility (V_i), along with threshold detection experiments to empirically identify collapse points (θ_i). Longitudinal data collection protocols are essential for capturing temporal feedback effects and the differential accumulation structure underlying belief state evolution.

Future research should establish mathematical conditions under which hierarchical parameters are uniquely recoverable, minimum complexity requirements for stable estimation at each level, and sensitivity analysis procedures for assumption violations.

Software development would require extensions to existing BART computational frameworks to handle joint estimation of correlated outcomes, cross-level residual passing mechanisms, and constraint enforcement for theoretical consistency.

3 Directed Cyclic Graph Integration and Temporal Dynamics

3.1: Theoretical Foundation of Directed Cyclic Graph Integration

The limbic-abductive decision architecture employs Directed Cyclic Graphs (DCGs) to represent the recursive feedback loops central to epistemic collapse: $RTT_1 \rightarrow RTT_2 \rightarrow IRD \rightarrow RTT_1$.

Traditional causal inference restricts analysis to Directed Acyclic Graphs (DAGs), treating cycles as problematic for path identification. However, the Anti-Golem Model's probabilistic Bayesian gates solve this identification problem through mutually exclusive pathway selection. While gate outcomes are stochastic, each gate forces traversal of exactly one path at a time. This is because decision making under uncertainty is nondeterministic and in real life mirrors quantum superposition in form and function. The difference here is that unlike with QS, you can observe the modeled state of the function at any point in time.

The three gates ensure clear causal identification:

- **RTT_1**: ***Firewall*** Either rational analysis continues OR limbic processing engages (never both)

- **RTT₂**: ***Active response*** Either belief commitment occurs OR uncertainty persists (never both)
- **IRD**: Reaffirmation, exit, or escalation (mutually exclusive outcomes)

This structure eliminates the traditional DCG concern about "multiple competing causal explanations" because the system maintains exactly one active pathway at any time. The apparent cycle RTT₁→RTT₂→IRD→RTT₁ represents clear sequential progression through mutually exclusive states: $DAG_{\{i+4,t\}} = DAG_{\{i,t\}} \Rightarrow \overline{DAG}_{\{i+4,t\}} \equiv DCG_t$.

3.2: Temporal Information Flow Between Gates

The Anti-Golem Model's epistemic collapse dynamics emerge from structured information flow through the sequential gate system: RTT₁ → RTT₂ → IRD. This micro-to-macro transformation requires dynamic information transfer with multi-stage aggregation mechanisms that convert individual cognitive processing into group-level decision outcomes.

Stage 1: Individual Processing (RTT₁ → RTT₂)

RTT₁ performs binary threshold detection for each individual agent, determining whether rational analysis continues, or limbic processing engages. Outputs feed directly into RTT₂, where each individual undergoes SoftMax processing over available beliefs until reaching dynamic threshold for argmax commitment. This stage maintains individual-level information processing with no aggregation.

Stage 2: Belief Aggregation $\boldsymbol{RTT_{2,pop} = P_{pop}(b_i)}$

Individual belief probabilities aggregate into group-level belief distributions over the current population P_t:

$$RTT_{2,pop} = P_{pop}(b_i) = \frac{\Sigma_{i,t} e^{\alpha*\lambda_{i,t}+\beta*\sigma_{i,t}+\gamma*V_{i,t}+logP(b_{i,t-1})/\tau}}{\Sigma_{\mathrm{i}}(\Sigma_j e^{\alpha*\lambda_{j,t}+\beta*\sigma_{j,t}+\gamma*V_{j,t}+logP(b_{j,t-1})/\tau})} = \frac{\Sigma_i e^{(\psi(b_i)/\tau)}}{\Sigma_i(\Sigma_{\mathrm{j}} e^{\psi(b_j)}/_{\tau})}$$

where i indexes beliefs, j indexes individuals, and :

$$\Psi(b_{i,t}) = \alpha\lambda_{i,t} + \beta\sigma_{i,t} + \gamma * V_{,i,t} + \log P(b_{i,t-1}).$$

This transformation captures how individual cognitive states combine into collective belief patterns while preserving individual-level tracking for identification purposes.

Stage 3: IRD Processing with Trifurcation Outcomes

$IRD\ integrates\ individual\ RTT^2 decisions\ using\ the\ aggregated\ belief\ context\ P_{pop(b_i)}$:

For each individual i in population $\mathbf{P_t}$:
$RTT_2(i) = Decision(P_{pop(b_i)})$

The individual decisions are then counted across the entire population to determine outcome proportions:

$$\textbf{Population}-\textbf{level aggregation}:$$
$$count_{dissolve} = \Sigma_{\mathrm{i}}\, I[RTT_2(i) = x(dissolve)]$$
$$count_{reaffirm} = \Sigma_{\mathrm{i}}\, I[RTT_2(i) = y(reaffirm)]$$
$$count_{escalate} = \Sigma_{\mathrm{i}}\, I[RTT_2(i) = z(escalate)]$$
$$count_{split} = \Sigma_{\mathrm{i}}\, I[RTT_2(i) = split]$$

These counts are weighted by outcome importance and normalized to calculate the final IRD measure:

$$IRD = \frac{\left(count_{dissolve} * w_x + count_{reaffirm} * w_y + count_{escalate} * w_z + count_{split} * w_{split}\right)}{P_t}$$

Each individual's outcome is determined by comparing their belief probabilities against predetermined thresholds

$$Decision(P(b_i) = \left\{\begin{matrix} x(dissolve), & if\theta_x \leq P(b_{ix}) \\ y\,(reaffirm), & ifP\left(b_{ix,iz}\right) < \min(\theta_x, \theta_z) \\ z(escalate), & if\ \theta_z \leq P(b_{iz}) \\ Split(b_{ix}, b_{iz}), & if \min\left(P(b_{ix}), P(b_{iz})\right) \geq \max(\theta_x, \theta_z) \end{matrix}\right\}$$

Variable definitions

$w_n = weights\ where\ n \in \{x(dissolve), y(reaffirm), z(escalate), split\}$
$\theta_x = dissolve\ threshold$
$\theta_z = escalate\ threshold$
$\min(\theta_x, \theta_z) = lesser\ of\ the\ two\ thresholds$
$\max(\theta_x, \theta_z) = greater\ of\ the\ two\ thresholds$
$P(b_{ix}) = probability\ of\ dissolve$
$P(b_{iz}) = probability\ of\ escalate$
$\min\left(P(b_{ix}), P(b_{iz})\right) = lesser\ of\ the\ two\ probabilites$

Stage 4: Conditional Feedback Determination

IRD outcomes determine subsequent information flow: collapse paths (dissolve, escalate) exit the system, while non-collapse paths (reaffirm) route through GFAB for continued cycling with resilience feedback R affecting future cognitive capacity in subsequent RTT$_1$ iterations.

This micro-to-macro architecture ensures that individual cognitive dynamics aggregate into meaningful group-level patterns while preserving causal identification through mutually exclusive individual pathway selection.

3.3: Time-Stepping Procedures and Temporal Indexing

The Anti-Golem Model requires comprehensive temporal tracking of system evolution through discrete time steps, necessitating a stateful architecture with complete historical preservation. Implementation would require a dedicated tier-0 stateful node capable of maintaining full system state across all temporal iterations.

Variable Classification Framework

The system distinguishes between dynamic variables requiring per-step updates and static constants that remain fixed throughout simulation:

Dynamic Variables (updated each time step):

- $Salience\ parameters: \lambda_{i(t)}, \sigma_{i(t)}$
- $Utility\ measures: V_{i(t)}$
- $Belief\ probabilities: P(b_i)(t)$
- $Threshold\ parameters: \theta_{i(t)}$
- $Population\ dynamics: P_t, group\ membership$
- $Aggregated\ measures: P\left(b_{i,j}\right)(t), IRD(t), R(t)$

Static Constants (fixed throughout simulation):

- Demographic characteristics
- Family structure (marriage status, children)
- Historical experiences
- Fundamental personality traits

Temporal Indexing System

Each system component maintains dual temporal coordinates: $Gate_{loop_{count},time_{step}}$where loop_count tracks cycling behavior through feedback loops and $time_{step}$provides universal temporal reference. Loop counters increment with each passage through identical gate sequences without IRD clearance, preserving diagnostic information through IRD processing to enable correlation analysis between cycling patterns and exit pathways.

State Management Requirements

Every time step $t \rightarrow t+1$ requires complete system state capture:

- $Agent_{j(t)} = [traitmatrixj(t)] for\ all\ agents\ j$
- $Systemstate(t) = aggregations, population, environmentalfactors$
- $Changevectors(t) = \Delta(all_{variables})$ between consecutive time steps

This stateful approach enables comprehensive behavioral signature analysis, tracking magnitude and pattern of change across all system variables to identify predictive indicators of agent cycling behavior, exit pathway selection, and intervention opportunities.

Update Sequencing and Dependencies

Temporal progression follows structured update sequences respecting causal dependencies between individual processing, aggregation mechanisms, and group-level outcomes while maintaining the mutually exclusive pathway selection that ensures causal identification throughout the recursive feedback system.

4: Trait-Based Migration and Interventional Modeling

4.1: Theoretical Foundation in Transition Theory

Building on Schlossberg's Transition Theory (Anderson et al., 2021; Schlossberg, 1981), this section addresses a fundamental reality of human adaptation: traits are inherent to agents and transfer with them regardless of context. The value lies not in predicting whether traits will transfer, but in understanding how those same traits will interact with new environmental contexts to enable targeted interventions.

Schlossberg's framework identifies transition success as dependent on the balance of resources and deficits across four categories: **Situation** (environmental context), **Self** (individual characteristics), **Support** (available assistance), and **Strategies** (coping approaches) (Anderson et al., 2021; Schlossberg, 1981). The AGM framework operationalizes this through empirical trait discovery rather than researcher-imposed categorizations.

4.2: Empirical Trait Discovery via Machine Learning

The trait-environment modeling approach developed here emerged from empirical observations during the author's research on student veteran transitions (Leizerman, 2024)

Traditional approaches to trait assessment rely on researcher cognitive filters and predetermined categories. As my capstone research revealed: **"virtually none of the research I did with Monica involved skills taught in the master's program. But she allowed me to do so because of trust"** (Leizerman, 2024). This demonstrates how actual trait-environment interactions often differ from theoretical expectations. This insight directly informs the AGM approach to trait-environment modeling - the challenge isn't changing fundamental cognitive patterns, but understanding how they function in new contexts.

The RH-BART framework addresses this limitation through **Stage 1 diagnostic analysis** - conducting thorough empirical fact pattern analysis before imposing theoretical frameworks. Using machine learning for thematic trait coding removes researcher bias while discovering naturally occurring behavioral clusters within populations.

This approach mirrors effective diagnostic practice across disciplines. As implemented in my diagnostic-interventional analysis framework: "For the analysis I elected to use a hybrid qualitative approach that involved conducting a template thematic analysis... This ad hoc methodology was chosen because it incorporated elements of thematic analysis, grounded theory, phenomenology, and autoethnography to serve as a means of implementing multiple triangulation methods to improve reliability" (Leizerman, 2024).

4.3: Trait-Environment Interaction Modeling

Cultural Competency and Acculturation Dynamics

My capstone findings on cultural competency directly validate the trait-environment interaction approach: **"Cultural competency issues are one of the two major themes found in situation... The first**

issue is veterans not understanding civilian culture. The second being the opposite, civilians not understanding veterans" (Leizerman, 2024).

This bidirectional incompatibility illustrates the core principle: traits (military cultural patterns) are inherent and portable, but their effectiveness depends entirely on environmental reception and interaction patterns. Veterans retain their decision-making styles, communication patterns, and task orientation regardless of context - the question is how these traits function in academic environments with different cultural norms.

Building on Berry's Acculturation Strategies (Berry, 2005), the analysis revealed that **"SVs and their broader college communities are likely to fall in a range between separation/segregation and marginalization/exclusion"** (Leizerman, 2024). This finding demonstrates how trait-environment mismatches create predictable adaptation patterns that can be empirically modeled and addressed through targeted interventions.

Mattering and Environmental Responsiveness

The discovery of mattering theory (Schlossberg, 1989) provides crucial insight for trait-environment optimization. Research demonstrates that adaptation success depends not on changing inherent traits, but on creating environments that can recognize and utilize those traits effectively: "The aspects of mattering she lists can provide insight as to how this might be accomplished. The aspects are: 1. Attention -- Commanding the interest of others; being noticed... 2. Importance -- the belief that others care about what is important to us... 3. Ego-Extension... 4. Dependence... 5. Appreciation" (Leizerman, 2024).

The RH-BART framework models how specific trait clusters interact with environmental capacity for providing these mattering experiences, enabling prediction of adaptation outcomes and design of targeted support interventions.

4.4: Interventional Framework Implementation

Level 1: Individual Trait Mapping

$\lambda_{ij} = \mathrm{BART}_1{}^{\alpha}(X_{ij}) + u^1{}_{ij}$ # Identity-weighted behavioral patterns $\sigma_{ij} = \mathrm{BART}_1{}^{\beta}(X_{ij}) + u^2{}_{ij}$ #
Threat/stress response patterns
$V_{ij} = \mathrm{BART}_1{}^{\gamma}(X_{ij}) + u^3{}_{ij}$ # Value/motivation patterns $\theta_{ij} = \mathrm{BART}_1{}^{\theta}(X_{ij}) + u^4{}_{ij}$ # Adaptation threshold
patterns

These parameters represent empirically-discovered trait clusters rather than theoretically-imposed categories, enabling more accurate prediction of trait-environment interactions.

Level 2: Environmental Context Modeling

Environmental parameters capture the contextual factors that determine how traits will be received and utilized:

$$\mathrm{IRD}_j = (1/n_j)\ \Sigma_i\ \mathrm{RTT}_{2ij} + \mathrm{BART}_2(Z_j, u^{1-4}{}_{ij}) + v_j$$

Where $\mathbf{Z}_j$ represents environmental characteristics such as cultural norms, institutional flexibility, support availability, and adaptation resources.

Level 3: Intervention Design

Based on empirically-discovered trait-environment interaction patterns, interventions can be precisely targeted to address specific mismatches rather than applying generic support approaches.

4.5: Real-World Applications

Student Veterans in Higher Education

My capstone research demonstrated that **"the PTVC would prefer information on issues that they can act on that might influence self-esteem as well as other factors regarding SV success at ASU"** (Leizerman, 2024). The trait-environment modeling approach provides exactly this capability by identifying which military traits create friction in academic contexts and designing specific environmental adaptations to optimize their function.

Example Application: Military task-orientation (trait) + academic deadline flexibility (environmental adaptation) = improved academic performance through trait-environment optimization rather than trait modification.

Example Application: Military task-orientation (trait) + academic deadline flexibility (environmental adaptation) = improved academic performance through trait-environment optimization rather than trait modification. To illustrate the point, complete paraplegia is a portable causal trait. It is causally linked to greater difficulty universally. But this is not universal. Many features do not maintain causal relationships across environments or over time.

This problem leads to significant threats to validity due to model fit as well as the real human problems. Thematic Analysis aims to address part of this by generating parameters that emerge from the data and thus avoid the issues that come from prescriptive parameters. It is for this reason that I view BART as thematic analysis through machine learning and if that is true then RH-BART is grounded theory. The difference in both cases is that the primary task of encoding parameters is done recursively and hierarchically. Because of this, Grounded theory, happens to be one of most powerful tools for identifying traits, but is it extremely labor intensive.

RH-BART aims to solve the labor-intensive issue through machine learning and to allow for reprogramming—not of the individual, but of the analytical frame. Rather than forcing traits to conform to prescriptive models or cultural expectations, RH-BART enables recursive updating of causal assumptions based on observed functional fit across environments. In this way, trait-environment compatibility is not merely predicted but empirically derived and iteratively refined. This allows for systematic identification of which traits require accommodation, which transfer robustly, and which may carry latent mismatches that undermine adaptation. The implication is profound: if grounded theory provides the slow, careful excavation of latent structure, RH-BART serves as its computational analog—capable of scaling insight generation across populations, contexts, and timeframes without sacrificing the underlying commitment to empirical emergence over theoretical imposition without sacrificing the underlying commitment to empirical emergence over theoretical imposition.

Immigration and Cultural Adaptation

The framework extends beyond veteran transitions to any population navigating cultural change. Immigrants bring inherent cultural traits (communication styles, authority relationships, family structures) that interact with new environmental contexts (workplace cultures, social norms, institutional expectations).

Rather than requiring complete cultural assimilation, the framework identifies which traits transfer effectively and which require environmental accommodation for optimal adaptation outcomes.

Organizational Onboarding

New employees bring established work styles, communication patterns, and decision-making approaches that interact with organizational cultures. The framework enables prediction of integration challenges and design of targeted onboarding interventions based on trait-environment compatibility analysis.

5: Causal Feature Portability and Structural Rewiring

5.1: The Fundamental Problem - When Causality Breaks Down

Traditional causal inference assumes that **X → Y** relationships are portable across populations. This assumption underlies most intervention research and policy design, yet consistently fails when applied across diverse cultural, ideological, or cognitive contexts (Pearl, 2009; Hernán & Robins, 2020). The Anti-Golem Model addresses a deeper issue: what happens when the **internal meaning** of causal relationships changes across groups due to different cognitive architectures?

Consider a simple example: **"Expert consensus reduces belief uncertainty"**

- **High-trust population**: Expert consensus → reduced uncertainty → policy acceptance
- **Low-trust population**: Expert consensus → increased suspicion → policy rejection
- **Traditional conclusion**: "The intervention failed in the low-trust group"
- **RH-BART conclusion**: "The causal structure requires rewiring for the low-trust population's cognitive architecture"

This distinction is crucial for understanding democratic breakdown. As political polarization increases, the same evidence produces opposite effects across ideological groups - not because people are irrational, but because they process causal relationships through fundamentally different cognitive frameworks (Sunstein, 2017; Brady et al., 2020).

5.2: Causal Feature Portability vs. External Validity

Causal Feature Portability represents a focused subset of external validity concerns, addressing **which specific features maintain stable causal influence** across populations rather than whether entire causal effects generalize.

External Validity (Broad):

- Question: Do the causal conclusions from one population apply to others?
- Focus: Generalizing the entire causal effect estimate
- Mathematical: $P\big(Y\big|do(X)\big)_{source} \approx P\big(Y\big|do(X)\big)_{target}$

Causal Feature Portability (Focused):

- Question: Which specific features maintain stable causal influence across populations?
- Focus: Feature-level causal stability rather than effect-level generalization
- Mathematical: Identifies which components of X have portable causal relationships with Y

This distinction proves critical for intervention design. External validity asks "Does the treatment effect transfer?" while causal feature portability asks "Which features of the treatment mechanism transfer?" (Pearl, 2014; VanderWeele & Shpitser, 2013).

5.3: RH-BART as Causal Structure Reconstruction Engine

The innovation of RH-BART lies not in predicting whether causal relationships will transfer, but in **reconstructing valid causal graphs** for each population's belief-processing logic. This addresses what Pearl (2014) identifies as the core challenge in causal inference: distinguishing between genuine causal relationships and population-specific confounding patterns.

5.4: Hierarchical Decomposition for Causal Reconstruction

Level 1: Individual Causal Parameters The framework extracts population-specific gate parameters that determine how causal relationships are processed:

$$\mathrm{P}(\mathrm{b_{ij}}) = \frac{e^{\left(\frac{\Psi(b_{\mathrm{ij}})}{\tau}\right)}}{\Sigma_{\mathrm{k}}\, e^{\left(\frac{\Psi(b_{\mathrm{kj}})}{\tau}\right)}}$$

Where: $\Psi(b_{ij}) = \alpha\cdot\lambda_{ij} + \beta\cdot\sigma_{ij} + \gamma\cdot V_{ij} + \log P(b_{ij}^{t-1})$

These parameters represent how different populations weight evidence (λ_{ij}), process threat information (σ_{ij}), and evaluate utility (V_{ij}) - fundamentally altering the causal pathways through which interventions operate.

Level 2: Group-Level Interaction Patterns
$IRD_j = (1/n_j)\ \Sigma_i\ RTT_{2ij} + BART_2(Z_j, u^{1-4}{}_{ij}) + v_j$

This level captures how individual differences aggregate into group-level causal patterns, addressing what Hernán & Robins (2020) describe as the "transportability" problem in causal inference.

Level 3: System-Level Causal Architecture $\Phi_k = BART_3(S_{ik}, S_{ok}, C_{ik}, C_{ok}, W_k, v_j) + \eta_k$

System-level analysis reveals how environmental factors interact with population characteristics to determine which causal pathways remain functional across contexts.

5.5: Practical Applications of Causal Rewiring

Political Communication Across Ideological Divides

The framework's power becomes evident when analyzing political communication effectiveness. Consider Churchill's wartime speeches or JFK's moon landing address - the same rhetorical features that inspire unity in democratic contexts might trigger opposite responses in authoritarian settings.

Democratic Context Causal Logic: Inspirational leadership → shared national identity → collective action support

Authoritarian Context Causal Logic: Foreign leader rhetoric → regime threat perception → increased state control

Rather than concluding that "inspirational rhetoric doesn't work in authoritarian contexts," causal feature portability analysis reveals how to **reconstruct the causal pathway** for effectiveness within different cognitive architectures.

Public Health Interventions

The COVID-19 pandemic demonstrated how identical public health messages produced vastly different behavioral responses across political and cultural groups (Robertson et al., 2023). Traditional approaches attributed this to "misinformation" or "irrationality," but causal feature portability analysis reveals structural differences in how different populations process authority-based evidence.

High-Trust Population: Expert recommendation → credibility assessment → behavior change

Low-Trust Population: Expert recommendation → institutional skepticism → behavior resistance

Reconstructed Pathway: Community validation → peer credibility → behavior change

The same health outcome becomes achievable through different causal mechanisms tailored to population-specific cognitive architectures.

Educational Interventions Across Cultural Contexts

My capstone research revealed how accommodation approaches that succeeded for student veterans failed when applied broadly, demonstrating the need for causal pathway reconstruction: **"professors taking a holistic approach to SVs that focus on communication, flexibility, and accommodations"** (Leizerman, 2024).

The framework enables prediction of which pedagogical features will transfer across student populations and which require structural adaptation based on cultural background, prior educational experience, and cognitive processing patterns.

5.6: Simpson's Paradox Robustness Through Hierarchical Modeling

Simpson's Paradox represents a special case of causal feature portability failure, where aggregate-level relationships reverse when examined within subgroups (Pearl, 2014; Kievit et al., 2013). The RH-BART framework inherently addresses this through its hierarchical structure that explicitly models how group membership affects causal relationships.

Traditional Approach: Identifies Simpson's Paradox after the fact through subgroup analysis
RH-BART Approach: Prevents Simpson's Paradox by modeling group-specific causal parameters from the outset

The recursive parameter extraction tracks how causal effects change across aggregation levels, ensuring that interventions remain effective when applied to specific subpopulations rather than failing due to unrecognized group heterogeneity.

5.7: Implications for Democratic Resilience

The framework's ultimate application addresses democratic breakdown itself. As political polarization increases, the same democratic institutions and processes produce opposite effects across ideological groups. Rather than viewing this as inevitable democratic failure, causal feature portability analysis reveals how to maintain democratic legitimacy through institutional adaptation.

Traditional Democratic Theory: Assumes universal acceptance of democratic norms and procedures
Causal Portability Approach: Identifies which democratic features transfer across ideological contexts and reconstructs institutional pathways for maintaining legitimacy

This transforms the challenge from "How do we make everyone accept democracy?" to "How do we reconstruct democratic causal pathways to function across diverse cognitive architectures while preserving core democratic outcomes?"

6: Scalable Internal Validity as a Modeling Goal

6.1: The Fundamental Problem with Static Validity Frameworks

Traditional internal validity represents a conceptual dead end in social science methodology. Campbell and Stanley's (1963) binary framework—either a study correctly identifies causal relationships or it does not—cannot survive contact with real-world cognitive diversity. This static approach has produced what Yarkoni (2020) identifies as the "generalizability crisis": research achieving perfect internal validity within narrow constraints yet contributing nothing to cumulative knowledge because findings systematically fail to replicate across populations (Open Science Collaboration, 2015; Ioannidis, 2005).

The problem runs deeper than methodological inadequacy. When internally valid studies fail to transfer, researchers conclude either that original findings were invalid or that new populations are deficient (Henrich et al., 2010). This binary thinking ignores a fundamental reality: **causal relationships require population-specific parameter optimization to maintain validity across cognitive architectures**, similar to how AI models require domain-specific fine-tuning to maintain performance (Kenton & Toutanova, 2019).

6.2: Parameter Fine-Tuning as Validity Solution

The AGM framework transforms internal validity from a static design constraint into a **dynamic computational property** through systematic parameter optimization. Rather than achieving validity once through experimental design, the RH-BART framework treats validity as an ongoing process that must be continuously earned through demonstrated causal stability across diverse populations.

This represents a paradigm shift from traditional frameworks:

Traditional: InternalValidity = f(ExperimentalDesign, ControlImplementation)
AGM: $Validity(t+1) = f(Validity(t), CrossPopulationStability, ParameterPortability)$

The innovation lies in **quantitative validity scoring** through cross-population parameter testing. When Group A's optimized parameters are applied to Group B's data, performance degradation provides a precise validity metric:

$$TransferValidity = 1 - (AveragePerformanceDropWhenTransferred)$$

This approach parallels successful AI fine-tuning strategies where pre-trained models achieve domain-specific effectiveness through systematic parameter adjustment rather than complete retraining (Howard & Ruder, 2018; Peters et al., 2019).

6.3: Implementation Through Computational Maintenance

The framework operationalizes validity maintenance through what we term "validity tuning"—periodic parameter optimization analogous to automotive maintenance schedules. Just as car engines require regular tuning for optimal performance across operating conditions, causal relationships require systematic parameter adjustment for optimal validity across cognitive architectures.

Maintenance Schedule Framework:

- **Baseline Tuning** (6-12 months): Comprehensive parameter optimization across major population clusters
- **Routine Maintenance** (Quarterly): Core population parameter updates based on performance monitoring
- **Performance Restoration** (As needed): Intensive re-tuning triggered by validity degradation alerts

Historical validation demonstrates the approach's potential. Stereotype threat research exemplifies the problem: Steele and Aronson's (1995) findings showed mixed replication success partly due to applying original parameters without population-specific optimization. Retrospective AGM analysis suggests their λ_identity = 0.9, σ_{threat}= 0.8 parameters would require fine-tuning to $\lambda_{identity}$ = 0.4, σ_threat = 0.3 for alternative populations, potentially preventing replication failures through systematic parameter optimization.

6.4: Implications for Evidence-Based Practice

This computational approach fundamentally transforms evidence-based policy implementation. Rather than applying research findings universally and being surprised by implementation failures, practitioners can access **population-optimized parameters** that maximize intervention effectiveness across constituencies.

The framework addresses what Pearl (2014) identifies as the core challenge in causal inference: distinguishing genuine causal relationships from population-specific confounding patterns. By making cross-population validity quantifiable and optimizable, the AGM approach adapts proven computational strategies from artificial intelligence (Ruder, 2017) to address validity challenges in social science.

Future applications should focus on expanding fine-tuning capabilities to handle cultural reasoning patterns, generational information processing differences, and digital-mediated cognitive architectures. The ultimate goal remains social science knowledge that maintains explanatory power through systematic computational optimization—knowledge robust enough to inform effective policy across the full spectrum of human cognitive diversity.

7. Future Research Directions

Experiment 1: Controlled Threshold Manipulation

Design: Laboratory subjects engage in belief formation tasks while researchers systematically vary threshold parameters (θ) as experimental dials. Subjects receive evidence streams about ambiguous scenarios (e.g., economic forecasts, policy outcomes) while salience parameters (attention allocation, threat perception, utility weighting) are measured in real-time through eye-tracking, physiological monitoring, and self-report instruments.

Key Question: Can controlled threshold manipulation produce predictable belief state transitions matching theoretical predictions?

Experiment 2: QRNG Proximity Detection

Design: Validated threshold models from Experiment 1 serve as benchmarks for detecting proximity to unknown thresholds in naturalistic settings. Subjects encounter information streams with quantum versus classical random elements while researchers monitor for early warning indicators without knowing actual threshold values.

Key Question: Can early warning systems detect approaching thresholds under genuine uncertainty?

Three-Tool Benchmarking Suite

DSEM(Dynamic Structural Equation Modeling): Validates lead-up dynamics and proximity detection capabilities by tracking continuous belief evolution, salience accumulation patterns, and temporal feedback loops before threshold crossing.

HMM(Hidden Markov Models): Validates transition detection accuracy by confirming whether predicted threshold crossings correspond to actual discrete state changes from superposition to committed belief states.

POMDP(Partially Observed Markovian Decision Processes): Validates post-collapse dynamics by modeling decision-making under uncertainty during recovery phases when agents cannot fully observe others' belief states.

This integrated approach provides comprehensive validation of all EWS components.

Conclusion: A Theoretical Framework for Democratic Cognitive Resilience

The Anti-Golem Model's theoretical framework **proposes** a fundamental paradigm shift in understanding democratic fragility—from viewing cognitive breakdown as a random failure of rationality to **hypothesizing** it as a predictable, measurable, and potentially preventable consequence of systematic overload in human information processing architectures. By integrating neurological foundations with computational precision, the framework **suggests** how abstract concerns about "polarization" and "misinformation" might be transformed into tractable engineering problems with quantifiable parameters and targeted intervention points.

The framework's **theoretical contribution** lies in its recognition that democratic collapse **may occur** not through external assault but through recursive feedback loops that exploit normal cognitive architecture under stress. The limbic-abductive decision framework (3.1) **theorizes** that the transition from rational deliberation to emotion-driven inference follows measurable threshold processes, **potentially** making intervention possible before irreversible commitment occurs. The RH-BART computational implementation (3.2-3.3) **provides a method for** empirical discovery of population-specific parameters that govern belief processing, while the trait-portability extensions (3.4-3.5) **offer approaches to ensure** interventions remain effective across diverse cognitive architectures.

The scalable internal validity approach (3.6) **addresses** social science's replication crisis by **adapting** proven computational strategies from artificial intelligence. Rather than hoping for universal findings, the framework **proposes** systematically optimizing causal relationships for specific populations through parameter fine-tuning—**a methodology that could potentially** ensure evidence-based interventions maintain effectiveness across implementation contexts.

If empirically validated, the practical implications extend far beyond academic methodology to transform how democratic institutions identify, understand, and respond to cognitive threats before they reach crisis levels.

And while it will take some time to validate all of this, that is my next step. Because, by making cognitive resilience quantifiable and optimizable, the framework enables precision intervention in democratic systems—identifying emerging threats through early warning indicators, deploying targeted cognitive interventions before collapse thresholds are reached, and maintaining democratic legitimacy across ideologically diverse populations through systematically adapted institutional pathways.

The framework thus offers both diagnostic precision for understanding when and why democratic cognitive systems fail, and therapeutic precision for designing interventions that restore epistemic health while preserving democratic pluralism. In an era when democracy's greatest threats emerge from within—through the systematic exploitation of cognitive architecture rather than external force—the Anti-Golem Model provides theoretical foundations for defending democratic resilience through the same computational principles that enable modern AI systems to maintain performance across diverse domains.

The mathematical framework and implementation methods described in this work patent pending. App# 63/811907

Epilogue: The Section That Will Probably Not Pass Peer Review But Needs Said.

The mathematical foundation is solid, the computational tools exist, and the empirical validation pathway is clear. What is the point of servant leadership if I can't lead by giving people a framework and toolset to work with? I'm going down this road one way or the other. But this is no zero sum game. I'm not doing this for prestige or money. I'm doing it because there's a fucking problem with the country I love and I'm not fucking ok with it. And it's fucking ok to not be ok with it.

What is not ok, however, is allowing our frustrations with our failing democracy get in the way of “democratic teshuva (return)”. It means we must find ways to interrupt this cycle of degradation, escalation, and division. We must learn how to talk just as a former paraplegic must learn how to walk.

But more than that we must relearn cognitive empathy and intellectual humility (I understand the irony of me saying this after claiming to have paradigm shifting ideas. But the difference between arrogance and humility is false knowledge and earned hope).

Honestly, I'll figure it out eventually, or maybe I won't. I can go it alone if I must. But it'd go a whole lot easier and a whole lot faster if I didn't. The tools exist. The theory is sound. The mission is clear. What remains is the choice: do we use our knowledge to understand and repair the mechanisms of democratic breakdown, or do we continue refining methodologies while the foundations crumble around us?

I can accept the costs of finding out that I am wrong here, but can we afford the cost of never finding out if I’m right?

The Anti-Golem Model offers a path forward. The rest is up to us.

Consolidated APA7 References Section

References

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.

Anderson, M. L., Goodman, J., & Schlossberg, N. K. (2021). *Counseling adults in transition: Linking Schlossberg's theory with practice in a diverse world* (5th ed.). Springer Publishing Company. https://doi.org/10.1891/9780826135476

Arendt, H. (1951). *The origins of totalitarianism*. Harcourt, Brace.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, 275(5304), 1293-1295. https://doi.org/10.1126/science.275.5304.1293

Ben-Yehuda, N., & Goode, E. (1994). Moral panics: Culture, politics, and social construction. *Annual Review of Sociology*, 20, 149-171. https://doi.org/10.1146/annurev.so.20.080194.001053

Berry, J. W. (2005). Acculturation: Living successfully in two cultures. *International Journal of Intercultural Relations*, 29(6), 697-712. https://doi.org/10.1016/j.ijintrel.2005.07.013

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, 16(4), 978-1010. https://doi.org/10.1177/1745691620917336

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.

Charmaz, K. (2006). Constructing grounded theory: A practical guide through qualitative analysis. Sage Publications.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266-298. https://doi.org/10.1214/09-AOAS285

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181-204. https://doi.org/10.1017/S0140525X12000477

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. Putnam Publishing.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4), 429-453. https://doi.org/10.3758/CABN.8.4.429

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1), 151-161. https://doi.org/10.1162/003465302317331982

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805

Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A review of literature. *The Information Society*, 20(5), 325-344. https://doi.org/10.1080/01972240490507974

Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127-138. https://doi.org/10.1038/nrn2787

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3), 515-534. https://doi.org/10.1214/06-BA117A

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457-472. https://doi.org/10.1214/ss/1177011136

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, 7(4), 473-483. https://doi.org/10.1214/ss/1177011137

Glaser, B. G., & Strauss, A. L. (1967). The discovery of grounded theory: Strategies for qualitative research. Aldine Publishing Company.

Green, D. P., & Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly*, 76(3), 491-511. https://doi.org/10.1093/poq/nfs036

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica*, 68(3), 575-603. https://doi.org/10.2307/2999619

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), 61-83. https://doi.org/10.1017/S0140525X0999152X

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. CRC Press.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1), 217-240. https://doi.org/10.1198/jcgs.2010.08162

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 328-339. https://doi.org/10.18653/v1/P18-1031

Howard, P. N., & Hussain, M. M. (2013). *Democracy's fourth wave?: Digital media and the Arab Spring*. Oxford University Press.

Huddy, L., Feldman, S., Taber, C., & Lahav, G. (2005). Threat, anxiety, and support of antiterrorism policies. *American Journal of Political Science*, 49(3), 593-608. https://doi.org/10.1111/j.1540-5907.2005.00144.x

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jung, A. K., Stieglitz, S., Kissmer, T., Mirbabaie, M., & Kroll, T. (2022). Click me...! The influence of clickbait on user engagement in social media and the role of digital nudging. *PLOS One*, 17(6), e0266743. https://doi.org/10.1371/journal.pone.0266743

Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171-4186. https://doi.org/10.18653/v1/N19-1423

Kievit, R. A., Frankenhuis, W. E., Waldorp, L. J., & Borsboom, D. (2013). Simpson's paradox in psychological science: A practical guide. *Frontiers in Psychology*, 4, 513. https://doi.org/10.3389/fpsyg.2013.00513

Leizerman, S. (2024). *Student veteran transition at Arizona State University* [Master's capstone]. Arizona State University School of Public Affairs.

Levitsky, S., & Ziblatt, D. (2018). *How democracies die*. Crown Publishing.

Li, Q., & Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice*. Princeton University Press.

Loader, C. (1999). *Local regression and likelihood*. Springer. https://doi.org/10.1007/b98858

Marciano, L., Ostroumova, M., Schulz, P. J., & Camerini, A. L. (2022). Digital media use and adolescents' mental health during the Covid-19 pandemic: A systematic review and meta-analysis. *Frontiers in Public Health*, 9, 793868. https://doi.org/10.3389/fpubh.2021.793868

Müller, H. G., & Rønn, B. B. (1993). Kernel estimation of partial means with application to Cox regression. *Scandinavian Journal of Statistics*, 20(2), 89-103.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. https://doi.org/10.1126/science.aac4716

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Pearl, J. (2014). Comment: Understanding Simpson's paradox. *The American Statistician*, 68(1), 8-13. https://doi.org/10.1080/00031305.2014.876829

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, 2227-2237. https://doi.org/10.18653/v1/N18-1202

Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2), 5018-5051. https://doi.org/10.1214/17-EJS1337SI

Pratola, M. T. (2016). Efficient Metropolis--Hastings proposal mechanisms for Bayesian regression tree models. *Bayesian Analysis*, 11(3), 885-911. https://doi.org/10.1214/16-BA1032

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage Publications.

Roberts, G. O., & Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2), 349-367. https://doi.org/10.1198/jcgs.2009.06134

Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2023). Negativity drives online news consumption. *Nature Human Behaviour*, 7(6), 812-822. https://doi.org/10.1038/s41562-023-01538-4

Rothe, D., & Muzzatti, S. L. (2004). Enemies everywhere: Terrorism, moral panic, and US civil society. *Critical Criminology*, 12(3), 327-350. https://doi.org/10.1007/s10612-004-3879-6

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*. https://doi.org/10.48550/arXiv.1706.05098

Schlossberg, N. K. (1981). A model for analyzing human adaptation to transition. *The Counseling Psychologist*, 9(2), 2-18. https://doi.org/10.1177/001100008100900202

Schlossberg, N. K. (1989). Marginality and mattering: Key issues in building community. *New Directions for Student Services*, 1989(48), 5-15. https://doi.org/10.1002/ss.37119894803

Shin, D., & Jitkajornwanich, K. (2024). How algorithms promote self-radicalization: Audit of TikTok's algorithm using a reverse engineering method. *Social Science Computer Review*, 42(1), 1-23. https://doi.org/10.1177/08944393231225547

Sparapani, R., Spanbauer, C., & McCulloch, R. (2021). Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *Journal of Statistical Software*, 97(1), 1-66. https://doi.org/10.18637/jss.v097.i01

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69(5), 797-811. https://doi.org/10.1037/0022-3514.69.5.797

Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media*. Princeton University Press.

Tan, Y. V., Flannagan, J., & Elliott, M. R. (2018). A flexible multi-level model for binary responses with applications to ensemble methods. *Computational Statistics & Data Analysis*, 123, 150-158. https://doi.org/10.1016/j.csda.2018.02.007

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279-1285. https://doi.org/10.1126/science.1192788

VanderWeele, T. J., & Shpitser, I. (2013). On the definition of a confounder. *The Annals of Statistics*, 41(1), 196-220. https://doi.org/10.1214/12-AOS1058

Walsh, J. P. (2020). Social media and moral panics: Assessing the effects of technological change on societal reaction. *Media and Communication*, 8(1), 454-463. https://doi.org/10.17645/mac.v8i1.2635

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 45, e1. https://doi.org/10.1017/S0140525X20001685