

# Cognitive Warfare, AI, and Security:

## Why Deductive Mathematical Foundations Are Essential in Integrated Kinetic and Cognitive Conflict

*A Non-Mathematical Overview for Policy Audiences*

*Based on "A Unified Theory of Latent Potentials: Homeostasis and Cognitive Warfare  
— Toward a Mathematical Foundation of Cognitive Control in Physical, Social, and AI  
Systems"*

Research paper written for the April 21, 2026 lecture at National Defense University



Carnegie Mellon University  
Security and Privacy Institute



## Hideto Tomabechi

- Cognitive Research Labs, Inc. CEO & Principal Research Scientist
- Research Professor, C5I Center, George Mason University
- Fellow, CyLab, Carnegie Mellon University

- 1. What the Public Thinks Cognitive Warfare Is (Wiki-Level)
- 2. What Cognitive Warfare Actually Is — A Personal History
- 3. How It Differs from Information Ops & PsyOps
- 4. Why Cognitive Science Is the Foundation
- 5. Why We Need Mathematical Foundations
- 6. Current Research: Unified Theory (Key Points + Appendix A)
- 7. New Methods Beyond AI Generation
- 8. What AI Cannot Do — The Human Expert Factor
- 9. AI Cyber Vulnerability in Cognitive Warfare — A Strategic Security Imperative

## **Part 1:**

# **What the Public Thinks Cognitive Warfare Is**

# Public Understanding of Cognitive Warfare (Wiki-Level)

4

*The public conversation largely stems from a 2020 NATO report. This framing, while useful, is incomplete.*

- Originated in 2020 NATO Human Domain Concept Paper
- AI and social media enabling mass generation of disinformation
- Intervening in enemy (or civilian) cognition: beliefs, emotions, decisions
- Goal: induce behavior favorable to the adversary (fragmentation, vote manipulation, etc.)
- Framed as 'the 6th domain' after land, sea, air, space, and cyber

**This framing focuses on METHODS (information ops) rather than the underlying science.**

*Understanding cognitive warfare at the scientific level requires going beyond content generation to structural cognitive dynamics.*

## **Part 2:**

# **What Cognitive Warfare Actually Is — A Personal History**

## My field since the 1990s — then called ‘Brainwashing Research’

**1990s**

**Brainwashing Research (洗脳研究)**

Joint research with Harvard Medical School on functional brain science. National Police Agency (NPA) request following Aum Shinrikyo Tokyo subway sarin attack. Published in Japanese only — to prevent proliferation.

**2007**

**Renamed for English Lectures**

Began using ‘Cognitive Warfare’ for English-language lectures and seminars.

**Oct 2019**

**Washington D.C. Lecture**

George Mason University & Daniel Morgan Graduate School of National Security — “Internal Representation in Cognitive Warfare”

**Aug 2023**

**World’s First Cognitive Warfare COP**

Demonstrated to Admiral Aquilino, Commander, U.S. INDOPACOM

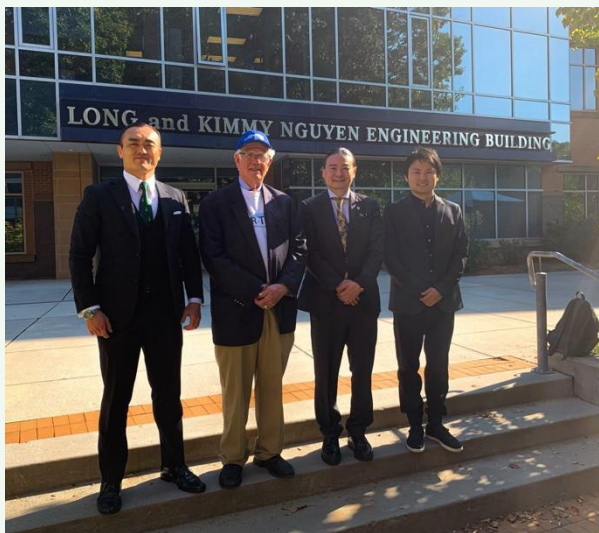
**Oct 2025**

**Heritage Foundation**

Cognitive Warfare lecture, Washington D.C.

# 2019 Washington D.C. Lecture — Cognitive Warfare

7



## INTERNAL REPRESENTATION IN COGNITIVE WARFARE

Hideto Tomabechi [drtomabechi@me.com](mailto:drtomabechi@me.com)  
George Mason University\* and Carnegie Mellon University\*\*  
\*Visiting Professor, C4I and Cyber Center \*\*Adjunct Fellow, CyLab



## Oct 2019 — DC Lecture: “Internal Representation in Cognitive Warfare,”

GMU C4I&Cyber Center & Daniel Morgan Graduate School of National Security

*‘Cognitive Warfare’ coined by Tomabechi ~2007. China and Russia have conducted long-term classified research in this domain.*

## **Part 3:**

# **How It Differs from Information Ops & PsyOps**

# Cognitive Warfare vs. Information Ops vs. PsyOps

9

## Information Operations

A methodology within cognitive warfare. Primarily focuses on content: what is said, what is sent.

## PsyOps (Air Force 1999 etc.)

Based on behaviorism. Does NOT assume cognitive modeling of the individual. Stimulus → response.

## Cognitive Warfare

Grounded in cognitive science. Requires modeling of individual cognitive dynamics. Structure, not just content.

**PsyOps changes behavior by manipulating stimuli. Cognitive Warfare changes behavior by restructuring the cognitive landscape itself.**

## **Part 4:**

# **Why Cognitive Science Is the Foundation**

# What 'Cognitive Science as Foundation' Means

- Functionalism as the base: human cognition is modeled via symbolic, quasi-symbolic, and sub-symbolic representations
- Individual cognitive activity is explicitly modeled — not assumed or averaged
  - Symbolic: language, logic, explicit belief systems
  - Quasi-symbolic: frames, schemas, conceptual metaphor
  - Sub-symbolic: neural, statistical, learned representations
- Social-level cognitive dynamics also modeled — group behavior emerges from individual models
- High affinity with AI: cognitive functions can be modeled mathematically or statistically
  - This is why cognitive warfare systems can be implemented as operational AI

## **Part 5:**

# **Why We Need Mathematical Foundations**

# Why Cognitive Warfare Requires Mathematical Foundations

13

- Most current cognitive warfare projects focus on: generating favorable/disinformation content, or detecting it
  - Problem: AI-generated outputs are statistical — no guarantee of coherence across separate attacks
- Modern conflicts integrate physical and cognitive dynamics inseparably (e.g., Hormuz Strait)
  - A tanker wasn't physically blocked — it was cognitively blocked
- Cognitive warfare systems need a unified COP displaying both physical and cognitive domain dynamics
  - This requires a shared mathematical foundation bridging both domains
- Without mathematical foundations: no formal guarantees, no integration, no strategic coherence
  - AI systems are fundamentally inductive — cannot detect corruption of their own training data or parameters (including by cyber attack)

**Physical kinetics + cognitive dynamics = one integrated battlespace. One math to govern both.**

*This becomes critical when AI is introduced into cognitive warfare.*

# Case Study: The Hormuz Strait, March 2026

14

A tugboat sunk. Three crew missing. A tanker ablaze. Traffic drops from 100+ vessels to a handful.

## Physical Event (small)

1 missile  
1 tugboat sunk

## Cognitive Effect (medium)


Insurance rates surge  
Ship operators reroute

## Strategic Outcome (massive)

Global oil spike  
Geopolitical instability

**The physical action was minimal. The cognitive restructuring was strategic.  
The  $V(x,t)$  of thousands of maritime operators was modified simultaneously.**

*$V(x,t)$  = evaluative function: the accumulated cognitive cost assigned to each possible state  $x$  at time  $t$ . The Ego minimizes  $\int V(x,t)dt$  — so raising  $V$  makes a state feel unsafe; lowering it makes it feel natural and stable.*

 **Plain Language:** The Strait was not blocked by force. It was blocked by reshaping the evaluative function  $V(x,t)$  of shipping operators — making 'transit' feel too costly to attempt. This is why physical and cognitive domains cannot be modeled separately.

## AI systems are fundamentally inductive

- Learn from data (past patterns)
  - Optimize locally in high-dimensional space
  - No guarantee of global consistency

### Result:

- Hallucination
  - Internal inconsistency
  - Unpredictable behavior at scale

**Local correctness  $\neq$  global correctness**

*Key point: a system that cannot guarantee global coherence cannot be trusted as a strategic control mechanism.*

- Cannot detect corruption of:
  - Training data
  - Model parameters
  - Output channels
- Same system evaluates its own outputs

## Result:

- Coherent but adversarial outputs
  - No reliable self-verification

**This is a structural limitation, not a bug.**

## AI without deductive structure

= powerful but unstable

### Inductive systems:

- No formal guarantees
- No global consistency
- Cannot verify their own integrity
- Vulnerable to adversarial manipulation

## AI within a deductive structure

= controllable and secure

### Deductive structure provides:

- Provable correctness
- System-level consistency
- Robustness under attack
- Verifiable integrity

**Control of structure is the only way to ensure security.**

**Part 6:**  
**Current Research:**  
**Unified Theory**  
**— Key Points**

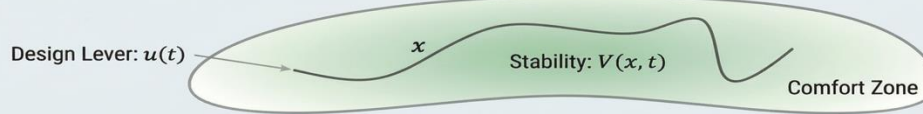
# The Total Comfort Zone (TCZ) Framework based on Tomabechi Theorem 1

JP definition keywords:  
Subjective Homeostasis / Future Redefinition /  
Stability Boundary Erosion

## The Total Comfort Zone Framework (based on Tomabechi Theorem 1)

Unified TCZ Definition  
(Conceptual Header)

$$TCZ(x_0) = \bigcup_{t \geq 0} \{x(t) \in \mathcal{R}(t; x_0) \mid V(x(t), t) \leq \theta\}$$

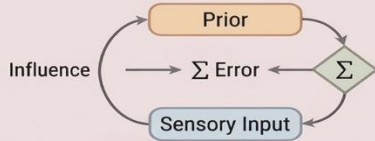


$$P(V \leq \theta \text{ for all } t \in [0, T]) \geq 1 - \alpha$$

**Tomabechi's "Ego"**  
modeled as an optimal control problem:  
$$\pi_c(x) = \arg \min_{u(t)} \mathbb{E} \int_0^T V(x(t), t) dt$$
  
Risk:  $\alpha$   $x(t)$ : cognitive state;  $u(t)$ : control input;  
 $V(x, t)$ : cognitive instability / evaluation function

The Three Model Pillars  
(Mathematical Formulation)

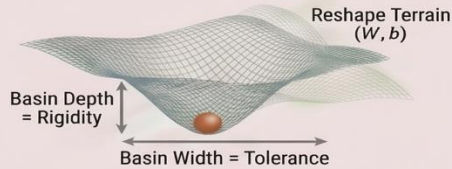
### A. Model I - Free Energy Formulation.



$$V(x, t) = \lambda_1 (\text{Prediction Error})^2 + \lambda_2 (\text{Dissonance}) + \lambda_3 (\text{Threat})$$

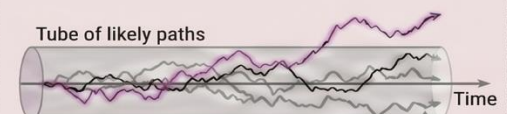
**Comfort  $\leftrightarrow$  Low Expected Free Energy**

### B. Model II - Attractor Basin Geometry.



$$V(x) = -\frac{1}{2} x^T W x + b^T x$$

### C. Model III - Stochastic Risk-Constrained Stability.



Operational Resilience Threshold  $\alpha$

$$dx = f(x, u, t)dt + \sum dW_t$$

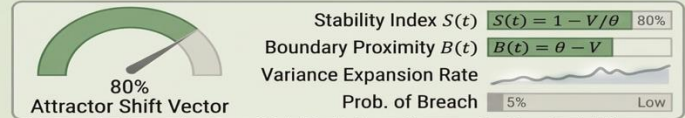
Operations & Metrics  
(Application Layer)

### Full-Spectrum Cognitive Operations



- Increase Adversary  $\alpha$  (Destabilize)
- Terrain Engineering
- Variance Manipulation
- Trajectory Control

### Cognitive COP Dashboard Metrics



Technical Appendix (not shown): Viability Kernel Definition: Energy Deformation:  $V = V_0 + \Delta V$ : Stochastic Dynamics Integration: Cross-Model Equivalence: Energy  $\leftrightarrow$  Basin  $\leftrightarrow$  Probability

Figure 1. The Total Comfort Zone (TCZ) Framework. Top: Unified TCZ definition with Ego modeled as optimal control  $\pi_c(x) = \arg \min_u \mathbb{E}[\int_0^T V(x(t), t) dt]$ . Middle: Three model pillars — (A) Free Energy Formulation, (B) Attractor Basin Geometry, (C) Stochastic Risk-Constrained Stability. Bottom: Full-Spectrum Cognitive Operations and Cognitive COP Dashboard Metrics.

# The Core Concept: Total Comfort Zone (TCZ)

**People do not act randomly. Behavior converges toward regions of perceived stability — this region I named the Total Comfort Zone (TCZ).**

$$\text{Ego} = \pi_c(x) = \operatorname{argmin} E[\int_0^T V(x(t), t) dt] \rightarrow x^*(t) \rightarrow \text{TCZ}$$

$\pi_c(x)$  = cognitive control policy: maps any cognitive state  $x$  to the optimal action minimizing accumulated evaluative cost. Subscript  $c$  = cognitive.

$x^*(t)$  = optimal cognitive state trajectory: the solution to  $\operatorname{argmin} E[\int V(x, t) dt]$  — the state path that minimizes accumulated evaluative cost, converging to the TCZ as its attractor.

$V(x, t)$  = evaluative function — the cost assigned to each cognitive state at time  $t$ . The Ego continuously minimizes accumulated cognitive cost. This is how behavior is generated.

**Cognitive warfare = modifying  $V(x, t)$ . Change what feels stable and acceptable — change behavior without orders.**

$$V(x, t) \rightarrow V'(x, t)$$

New evaluative landscape. New TCZ. New behavior.  
No direct command required.

Social level (Theorem 2):  
Narrow shared space  $\rightarrow$  fragmentation.  
Elevate abstraction  $\rightarrow$  integration.  
Choice of target determines direction.

*[For Non-Mathematical Audience]*

The concept of Self, developed in the author's earlier works through modal logic, operates in two distinct ways: it can select a desired TCZ from among possible worlds, or it can transform the TCZ itself. The mapping  $s : \text{TCZ} \rightarrow \text{TCZ}$  expresses this second operation — not movement within a comfort zone, but the transition or reconstruction of the comfort zone as a structure. Ego is the author's reformulation as a control equation for AI implementation of cognitive warfare. Where modal-logic Self defines desirable territory in terms of possible worlds, Ego is the control policy converging toward the TCZ as an attractor basin ( § 2.5). This duality — Self operating across possible worlds, Ego converging in state space — forms the mathematical foundation for implementing human cognition in AI. The same mechanism that enables a person to pursue their own goals is, depending on how it is applied, also the mechanism of brainwashing.

*[For Non-Mathematical Audience]*

This equation says that the Ego — the part of the self that makes decisions — constantly steers cognition toward states of lower internal tension and greater stability. It is not reacting randomly; it is solving an optimization problem at every moment, minimizing accumulated discomfort over time. In everyday terms: a person does not just react to each event independently; they navigate life in a way that, over time, moves them closer to their own stable, coherent sense of self and world — their TCZ.

*[For Non-Mathematical Audience]*

Before anything else can be defined, we need to specify what the mind is working on. This equation answers that question: the cognitive world  $W$  is not just the present moment, but the entire space of present and possible future states. A person navigating a difficult decision is not reacting to a single snapshot — they are simultaneously weighing many possible futures. This set  $W$  is the arena in which all subsequent concepts in this paper are defined.

*[For Non-Mathematical Audience]*

Not all possible futures feel the same. Some feel safe, coherent, and livable; others feel threatening or wrong. This operator  $r$  is the formal expression of that ranking: it reorganizes the space of possible worlds according to how compatible each one is with the agent's internal stability. The mind is not a neutral camera — it is a ranking machine. This ordering is what makes it possible to define, mathematically, which part of  $W$  counts as the 'comfort zone.'

*[For Non-Mathematical Audience]*

The logical structure here is: for ANY situation  $y$  you might face, there EXISTS a response  $x$  that returns you to stable ground. Your TCZ is the collection of worlds in which you are never completely stranded — there is always a path back to equilibrium. A person with a large TCZ can absorb a wide range of unexpected events without being destabilized; a person with a small TCZ can be thrown into crisis by even minor disruptions. This formula is the precise mathematical definition of psychological resilience.

## *[For Non-Mathematical Audience]*

The TCZ is simply the collection of all mental states that a person can reach — from their current situation — while remaining within a comfortable, stable range. Think of it as a map of all the psychological territory that feels 'safe' to inhabit. States outside the TCZ represent excessive tension, inconsistency, or distress. The formula captures this precisely: the TCZ is every state that is reachable and keeps the internal instability measure  $V$  below a threshold  $\theta$ . The structure of a person's TCZ defines, in mathematical terms, what their 'comfort zone' actually is.

Core claim: Einstein's 1901 capillarity theory and the Tomabechi cognitive framework are mathematically identical — both are accumulation-of-latent-potentials theories.

## Einstein 1901 Capillarity Theory

Latent potentials accumulate  
over SPACE

Surface tension emerges  
at the liquid/air boundary

Tiny boundary asymmetry  
→ large macroscopic effect

## Tomabechi 2026 Cognitive Framework

Latent potentials accumulate  
over TIME

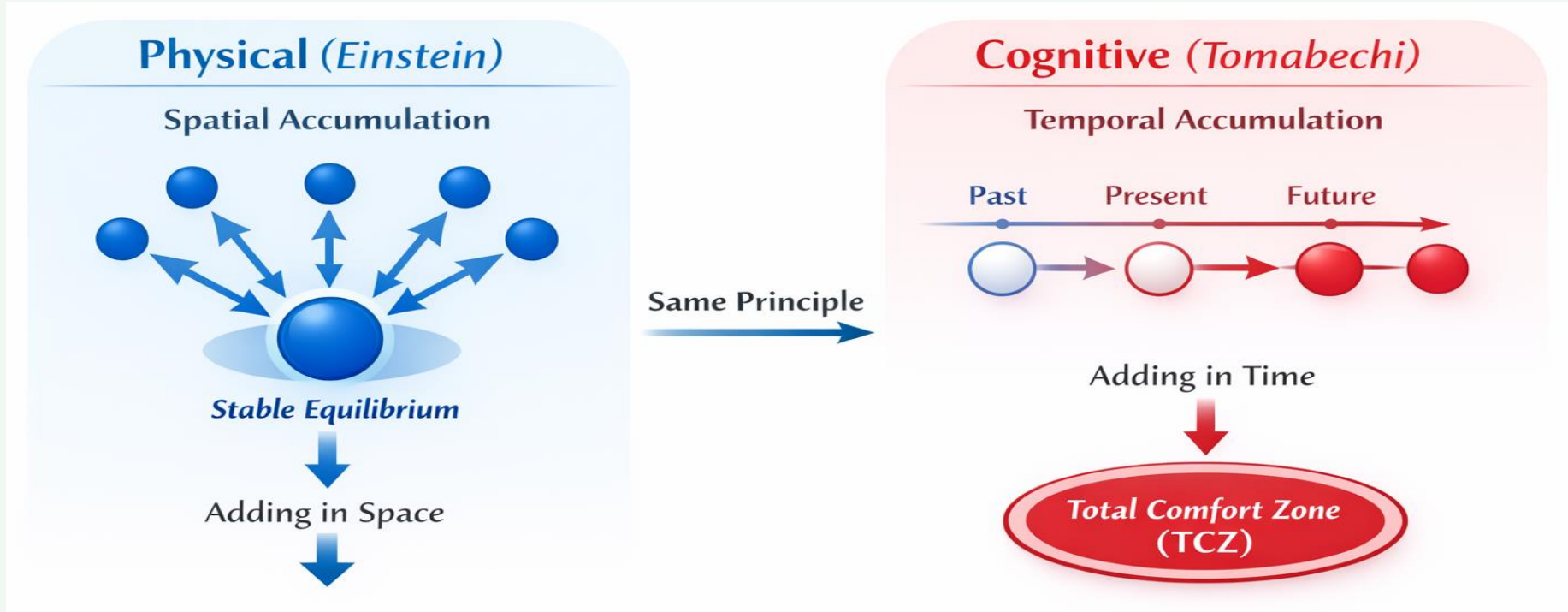
Behavior converges  
to the TCZ boundary

Tiny cognitive perturbation  
→ large behavioral shift

💡 Plain Language: Surface tension cannot be explained by looking at one molecule — it emerges from the sum of ALL molecular interactions across ALL of space. The same principle governs cognition: no single belief, memory, or experience determines behavior. Behavior emerges from the temporal accumulation of all evaluations. This is why Einstein's 1901 insight — that macroscopic phenomena require integration over the whole field — applies directly to cognitive dynamics. You cannot understand a person from one moment; only from the accumulated trajectory.

# Accumulation of Latent Potentials — Einstein 1901 & Tomabechi 2026

Both theories share the same mathematical skeleton: latent potentials accumulate — over space in physics, over time in cognition.

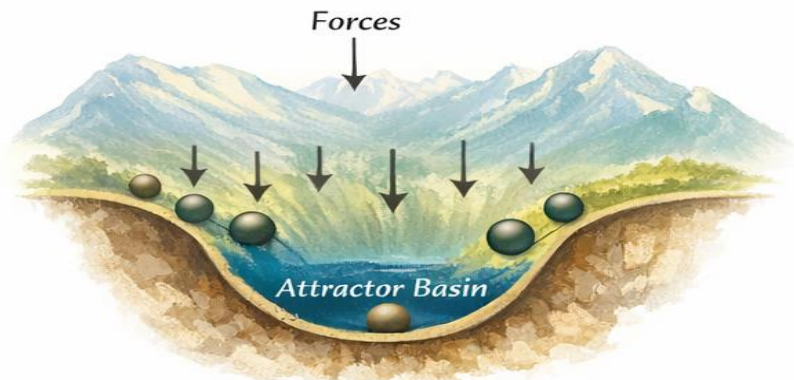


## *[For Non-Mathematical Audience]*

Just as a liquid's macroscopic properties emerge from the spatial sum of molecular interactions, a person's behavior emerges from the temporal sum of their cognitive evaluations — their judgments, feelings, memories, and expectations accumulated over time. No single moment determines behavior; it is the total weight of everything that has come before. This is why trauma has lasting effects, why trust is built slowly, and why belief systems resist sudden change. To understand or change behavior, you must understand the accumulated structure — not just the present moment.

# Unified Accumulation: Physical vs. Cognitive Systems

## Physical System

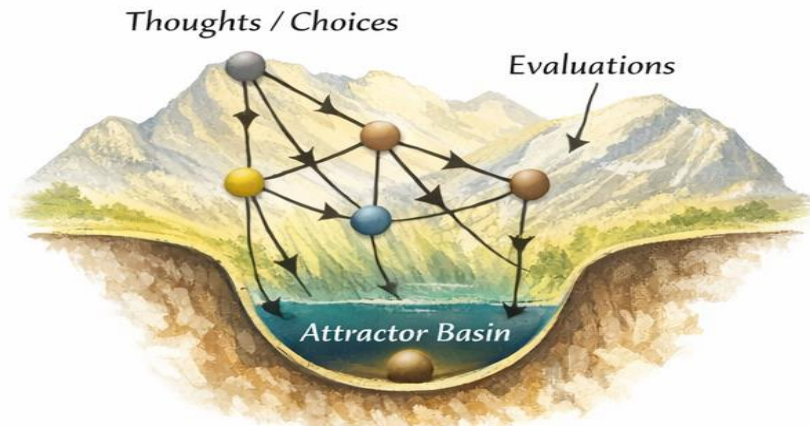


$$\int U(x) \rho dV$$

Accumulation over Space

Physical Space

## Cognitive System



$$\int U(r) \rho dT$$

Accumulation over Time

Cognitive Space

Left: In physical systems, local molecular interactions accumulate over space via  $\int U(x) \rho dV$ , converging to an attractor basin. Right: In cognitive systems, evaluations (thoughts/choices) accumulate over time via  $\int_0^T V(x(t), t) dt$ , converging to a cognitive attractor basin (TCZ). The structural equivalence: space in physics  $\leftrightarrow$  time in cognition.

# The Unified Theory: What This Paper Establishes

31

## A Unified Theory of Latent Potentials: Homeostasis and Cognitive Warfare

Tomabechi | April 4, 2026 | CyLab CMU · C5I Center GMU · Cognitive Research Laboratories

Core claim: the mathematical structure of Einstein's 1901 capillarity theory and the Tomabechi cognitive framework are structurally identical. Both are accumulation-of-latent-potentials theories — one over space, one over time.

### Tomabechi Theorem 1

Individual cognitive trajectories converge to a Total Comfort Zone (TCZ)

### Tomabechi Theorem 2

Socially coupled agents converge to a Shared-TCZ — social stability as emergent attractor

### Tomabechi Theorem 3

Abstraction elevation lifts convergence to a Least Upper Bound (LUB) — integration & altruism

### Appendix A

Operational extensions — how these theorems become cognitive warfare instruments

# Three Theorems: From Individual to Strategic

Each theorem governs a different level of human organization — and a different level of cognitive warfare.

## Theorem 1 Individual

$Ego = \operatorname{argmin} E[\int V(x,t)dt] \rightarrow x^*(t) \rightarrow TCZ$

$x^*(t) = \textit{optimal cognitive state trajectory: solution to } \operatorname{argmin} E[\int V(x,t)dt] \textit{ — minimizes accumulated evaluative cost} \rightarrow \textit{converges to TCZ as its attractor.}$

Every person continuously navigates toward their personal comfort zone. Modify  $V(x,t)$  — what feels stable — and the trajectory changes. No orders needed. No detectable coercion.

## Theorem 2 Social

$Shared\text{-}TCZ = \cap TCZ_i, \gamma_{\{ij\}} > 0$

$\gamma_{\{ij\}} = \textit{social cognitive coupling coefficient: degree of influence agent } j \textit{ exerts on agent } i\text{'s trajectory. } \gamma_{\{ij\}} > 0 \rightarrow \textit{cooperative (attractive) coupling} \rightarrow \textit{agents co-converge toward Shared-TCZ.}$

Groups develop a joint attractor. Narrow the shared space  $\rightarrow$  fragmentation. Expand it  $\rightarrow$  coherence. Social stability is NOT imposed externally — it emerges from joint cognitive dynamics.

## Theorem 3 Strategic

$LUB(\Phi) = \vee\{TCZ_i\}$  under abstraction operator  $\alpha$

With elevated abstraction, systems converge to the Least Upper Bound of shared cognitive worlds: maximal inclusion, minimal information, anti-conflict structure. Mechanism behind genuine alliance cohesion.

**The same mathematical structure governs all three levels. One math for cognitive-kinetic integration.**

## *[For Non-Mathematical Audience]*

Each person minimizes not only their own internal tension, but also their misalignment with others. The term  $S_i$  captures how much a person's cognitive state conflicts with the states of those around them. The coefficient  $\gamma_{ij}$  measures how strongly agent  $j$  influences agent  $i$ 's trajectory. When  $\gamma_{ij}$  is large, agent  $j$  exerts strong pull on agent  $i$  — cooperative coupling toward Shared-TCZ. When it is small, they operate more independently. This is the mathematical structure of social conformity, group identity, and collective stability. A society where coupling is strongly connected, all agents converge to a shared cognitive state — for better or worse, depending on what that shared states are.

## *[For Non-Mathematical Audience]*

This formula has two conditions, both required simultaneously. First: every person must be inside their own comfort zone — no one is being dragged somewhere they find intolerable. Second: their states must be sufficiently aligned with each other ( $\text{Share}(x) \geq \eta$ ). A shared comfort zone is not just any overlap — it is a state in which everyone is stable and mutually coherent. A functional team, a stable alliance, or a cohesive society all correspond to this region. When either condition fails — someone is destabilized, or alignment falls below  $\eta$  — the shared TCZ collapses.

# TCZ Structure, Abstraction Hierarchy & Dual Structure (Figs. 4 & 5)

Fig. 4 (left): Shared-TCZ as intersection of individual TCZs. Narrowing = Shared-low-TCZ. Elevation = Shared-high-TCZ / LUB. Fig. 5 (right): Dual structure — kinetic decrease (Physics/FEP) and cognitive decrease (Cognition) both converge upward to abstraction.

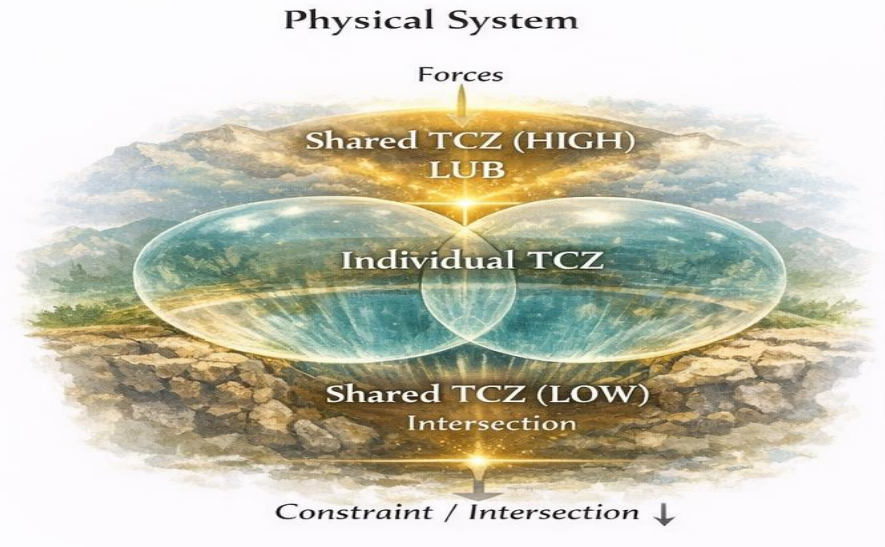


Figure 4: TCZ Structure and Abstraction Hierarchy

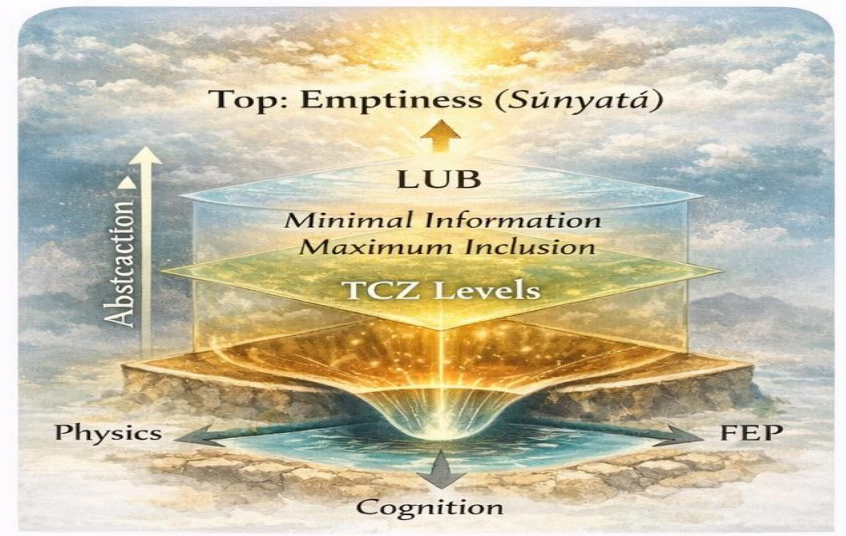


Figure 5: Dual Structure of Decrease and Abstraction

## *[For Non-Mathematical Audience]*

These two formulas describe two fundamentally different ways of being 'together.' The low shared TCZ (intersection,  $\cap$ ) is what all members already agree on — the common denominator. It is safe, but fragile: as the group grows or diversifies, the intersection shrinks. The high shared TCZ (LUB) is the most abstract concept that includes all individual positions without collapsing any of them. It grows richer, not smaller, with diversity. The mathematical difference between these two is the difference between an alliance based on a common enemy (intersection) and one based on shared aspiration (LUB).

# Why Abstraction Reduces Entropy — Not Increases It

*[For Non-Mathematical Audience]*

This result seems counterintuitive: in physics, entropy tends to increase as systems become more disordered. But in cognitive and information space, moving to higher abstraction does the opposite — it reduces entropy by eliminating unnecessary distinctions and organizing more under fewer concepts. The word 'justice' contains less raw information than a complete list of every specific law, but it structures the legal landscape more powerfully. This is why genuine wisdom — thinking at higher abstraction — produces clarity rather than vagueness. It compresses without losing structure.

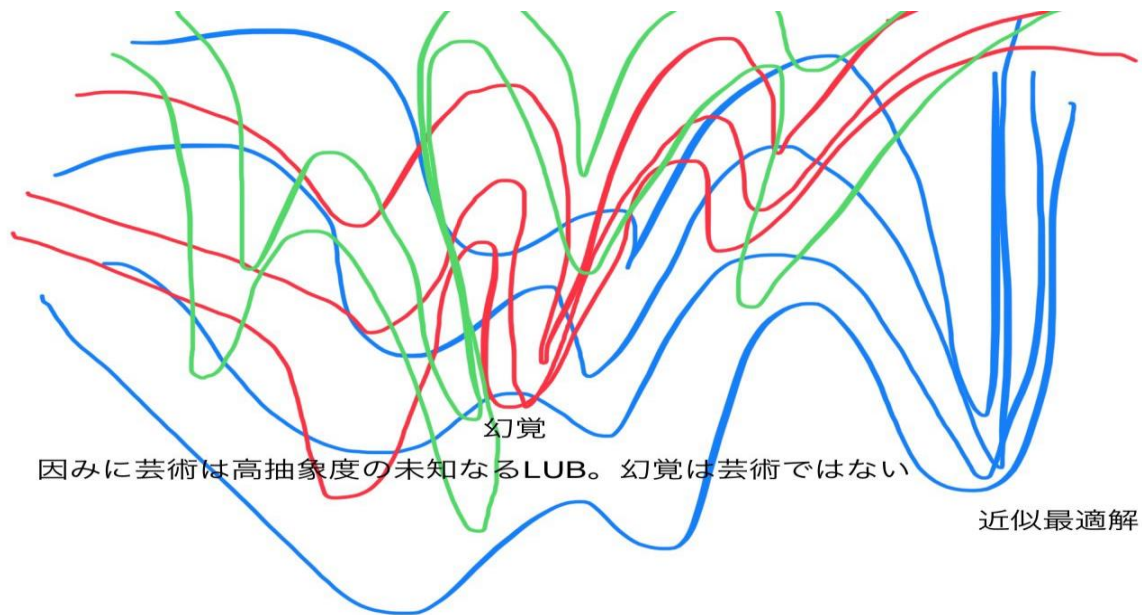
## *[For Non-Mathematical Audience]*

The entire framework rests on two simultaneous processes that point in different directions — yet do not conflict. Descent in potential produces stability: you move toward your comfort zone, reducing internal tension. Ascent in abstraction produces inclusion: you expand the scope of what you can coherently encompass. A person who is psychologically grounded and ethically broad is not making a tradeoff between inner peace and concern for others. Mathematically, they are expressing the same underlying structure at two levels at once — which is why the deepest stability and the widest compassion tend to appear together.

## Figure 6: Local Minima vs. Least Upper Bound (LUB)

[Advanced]

AI cannot produce true art through error minimization alone. Art is structurally different: the LUB in abstract space, not any local minimum.



- Hallucination (bad local min.)
- Approx. optimum (still local min)
- Art / LUB (max abstraction)

**Appendix A**  
**(Confidential):**  
**Operational**  
**Framework —**  
**Key Extensions**

# Appendix A.1: Operational Interpretation — The Three Theorems as Weapons

Each theorem directly translates to an operational capability:

<b>Theorem 1 (Individual)</b>	<b>Theorem 2 (Social)</b>	<b>Theorem 3 (Abstract)</b>
<b>Modify <math>V(x,t)</math> → Redefine TCZ</b>	<b>Constrain shared structure → Shared-low-TCZ</b>	<b>Elevate abstraction → Shared-high-TCZ</b>
Target's decision-making naturally evolves toward adversary-intended states. No direct command. No detectable coercion.	Amplify fragmentation and instability within target population. Narrow the space of shared stability.	Enable integration, coherence, and stability. Or — prevent enemy from achieving this.

**The same mathematical structure governs defense as well as offense. Control the shared abstraction level → control the stability of the operational environment.**

*[For Non-Mathematical Audience]*

This single arrow is the operational core of the entire paper. Cognitive warfare does not say 'do this' — it changes the landscape in which decisions are made, so that certain actions come to feel natural, unavoidable, or obviously correct. Once  $V$  has been changed to  $V'$ , the Self still minimizes freely — but it now minimizes toward a different TCZ. The target never receives a command; they simply find that their own judgment leads them somewhere new. This is why cognitive warfare is structurally harder to detect and resist than conventional influence: there is no message to reject, because the evaluative function itself has been rewritten.

## Appendix A.2: Boundary-Based Control — Maximum Leverage, Minimum Intervention

Key operational insight: the most sensitive point is not deep inside the TCZ, but at its boundary.

$$u^*_{cw}(t) = \operatorname{argmin} E \int_0^T ( |V(x,t) - \theta|^2 + \lambda C(u) ) dt$$

*$u^*_{cw}(t)$  = optimal CW control signal.  $V(x,t)$  = evaluative function.  $\theta$  = TCZ boundary threshold.  $\lambda$  = cost-effect tradeoff weight (scalar penalty on intervention cost; distinct from social coupling  $\gamma_{\{ij\}}$ ).  $C(u)$  = intervention cost.*

- This optimization targets the boundary of the TCZ — where small perturbations produce the largest behavioral shifts.
  - $|V - \theta|^2$  is symmetric: it attracts from both inside and outside TCZ. Objective: boundary maintenance, not interior stabilization.
  - $\lambda C(u)$ :  $\lambda$  is the cost-effect tradeoff weight — maximum effect, minimum exposure, minimum detectability.

### Einstein 1901 Capillarity Analogy:

Surface tension arises at the boundary between liquid and air — where molecular forces become asymmetric. A tiny structural difference at the interface produces a large, observable effect. Cognitive warfare at the TCZ boundary operates by exactly this principle: minimal structural perturbation → disproportionately large behavioral outcome.

**Target the boundary, not the interior. Maximum leverage, minimum cost, minimum detectability.**

# Appendix A.3: Multi-Bridge Message Ensemble — LUB-Guided Information Operations

A single message too far from the target's TCZ will be rejected. The solution: ensembles.

$$\max_{\{M\}} [ \sum A(m_k|x_t) + \alpha \cdot \text{Lift}(M) - \beta \cdot \text{Frag}(M) - \delta \cdot \text{Decept}(M) ]$$

$M$  = message ensemble ·  $A(m_k|x_t)$  = acceptance prob. ·  $\alpha$  = abstraction-lift weight ·  $\beta$  = fragmentation penalty ·  $\delta$  = deception penalty ( $\neq$  coupling  $\gamma_{\{ij\}}$ )

$\sum A(m_k|x_t)$

Each message individually acceptable — no single message overreaches the TCZ

$\alpha \cdot \text{Lift}(M)$

Lift weight  $\alpha$ : promotes LUB alignment with Shared-high-TCZ — collectively elevates abstraction

$\beta \cdot \text{Frag}(M)$

Frag penalty  $\beta$ : penalizes incoherent message sets — bridges must connect

$\delta \cdot \text{Decept}(M)$

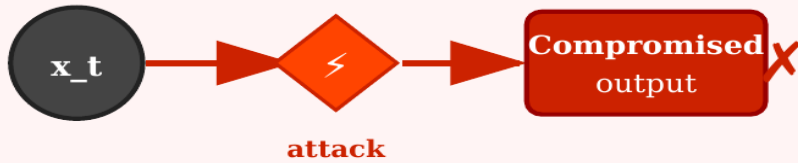
Decept penalty  $\delta$ : penalizes manipulation — structural change, not deception

**Each message is a bridge. Bridges must connect. The ensemble constructs pathways along which change becomes natural. This implements Theorem 3 at the level of information generation.**

Mechanism	Operation	Effect
Theorem 1 (A.1)	Modify $V(x,t)$	Rewrite the TCZ
Theorem 2 (A.1)	Constrain shared structure	Shared-low-TCZ $\rightarrow$ fragmentation
Theorem 3 (A.1)	Elevate abstraction	Shared-high-TCZ $\rightarrow$ integration
<b>A.2 (Boundary)</b>	Control near $\partial$ TCZ	Max effect, minimal intervention
<b>A.3 (Ensemble)</b>	LUB-guided message set	Lift toward Shared-high-TCZ via bridges

**Cognitive warfare does not control actions. It controls the structures from which actions emerge — most efficiently at the boundary, most coherently through ensembles.**

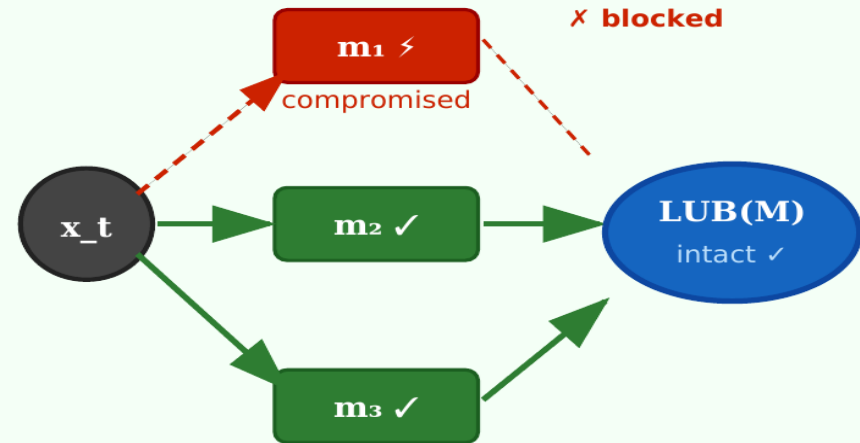
## Single Trajectory — Vulnerable



*One point of failure  
→ entire signal compromised*

No redundancy · No error correction

## Multi-Bridge Ensemble — Robust



*Partial compromise ≠ LUB compromise*

**Vulnerabilities:** adversary-intended or emergent — internal consistency failures across domains, AI hallucinations, cyber attacks on AI training data / model parameters / output channels. **Structural redundancy is essential.**

## **Part 7:**

# **New Methods Beyond AI Generation**

These emerging techniques must also be governed by the same mathematical framework:

## Interventions in Spatial Computing

AR/VR/XR environments that restructure perceived reality and alter spatial-cognitive anchors in ways indistinguishable from physical experience.

## Decoded Neural Feedback & Active Steganography (fMRI)

Reading and writing to neural representational spaces via fMRI-decoded feedback. Active steganography embeds cognitive triggers below conscious perception.

## Behavior Modification via Horizontal Gene Transfer

Wireworm-mantis gene transfer example: behavioral modification through genetic mechanisms that bypass cognitive processing entirely.

## **Part 8:**

# **What AI Cannot Do — The Human Expert Factor**

## A critical limitation of AI in cognitive warfare — stated before the conclusion:

- AI-based cognitive intervention minimizes discomfort — it cannot generate positive novel experiences
  - The joy of a cuisine you've never tasted. The emotion toward a person you've never met. These cannot be modeled statistically.
- AI is trained on what exists — it cannot generate truly novel anchored affect
- Human brainwashing experts can:
  - Synthesize anchored emotional states and embed them as triggers
  - Amplify, intensify, and maximize emotional valence beyond baseline
  - Induce altered states of consciousness that amplify receptivity

**AI provides cognitive intervention up to discomfort minimization.  
Human experts provide positive synthesis, embedding, amplification, and maximization.**

*The methodology for AI-implementing human brainwashing expertise is my specialty since the 1990s — but outside today's scope.*

# AI Cyber Vulnerability in Cognitive Warfare — A Strategic Security Imperative

50

Because AI operates inductively, it cannot self-detect tampering with its own training data or internal mathematical state.

## Training data / weights tampered

A cyber-adversary can redirect AI generative trajectories without detectable error signals. The system continues with apparent coherence — generating outputs aligned with adversarial objectives.

## Errors & hallucinations

Outputs diverging from intended directions may occur independently of deliberate attack, further complicating detection.

## Output-layer interception

Generated text, images, or decisions may be altered after leaving the AI system, before reaching human operators — with no cryptographic integrity guarantee.

## Deductive foundation is the only solution

Only a formally verifiable deductive control structure — as proposed in this paper — can provide guarantees that inductive AI systems, by their nature, cannot. AI can contribute to peace only within such a framework.

**Cyber-attacks on AI are a present strategic reality. A deductive mathematical foundation is the only structural guarantee against AI being weaponized in cognitive warfare.**

AI doesn't just generate warfare messages — it now writes the control code itself. Verification becomes mathematically impossible.

## The Scale Problem

- **Scale explosion:** AI writes control code for cognitive, cyber & physical weapons — orders of magnitude beyond human audit.
- **Self-attack risk:** Experts cannot verify the volume → unverifiable risk behaviors including self-attacks rise sharply.
- **Democratization:** AI-assisted programming lets personnel **with no CS training** produce warfare code at exponential scale.
- **The Gödel–Chaitin limit:** Inductive AI cannot prove the consistency of its own code — verification is **mathematically impossible**.

## Chaitin's Extension (Technical)

- **Non-computability of  $\Omega$ .** Chaitin's halting probability is a rigorously defined real number that no algorithm can compute.
- **Randomness unreachable.** Algorithmically random sequences cannot be derived from any finite axiom set.
- **Formal-system limit (Chaitin 1974/75).** Axiom system T proves only finitely many " $K(s) \geq n$ " — unprovable once n exceeds T's own complexity.

**Direct application:** When Kolmogorov complexity of LLM-generated code exceeds the verifier's complexity, verification is impossible *in principle*.

Only deductive mathematical control at a higher abstraction level can close this structural gap.

**Advanced Section:**

# **Mathematical Foundations — Deep Dive**

Deep dive for specialists

Self · Ego · TCZ · Altruism · Peace → Mathematical proofs

**Mathematical Foundations — Deep Dive:**

# Key Theoretical Results

**1****Promote Peace**

The most effective cognitive warfare is to promote Shared-High-TCZ convergence — the path of peaceful coexistence, avoiding conflict and war.

**2****Selfish Gene Is Incomplete (§ 2.12.6)**

Altruism is not in contradiction with evolution — it is evolution's natural result. Humans are fundamentally peaceful organisms who naturally avoid conflict and war.

**3****Mathematically Proved**

Both conclusions are formally proved: Theorem 3 (LUB convergence) and § 2.12.6 ( $\alpha_i^* = \operatorname{argmax} F_i$ ). These are theorems, not philosophical claims.

1

## Most Effective Cognitive Warfare: Promote Peace

The most effective cognitive warfare strategy is to promote integration toward Shared-High-TCZ — which is the path of peaceful coexistence, avoiding conflict and war. Shared-High-TCZ convergence is mathematically stable, self-sustaining, and evolutionarily optimal.

2

## Selfish Gene Is Incomplete ( § 2.12.6)

The Selfish Gene hypothesis (Dawkins) — a mainstream of Darwinism — claims humans are genetically predisposed to selfish behavior, corresponding only to Theorem 2 (low- $\alpha$ ) dynamics. § 2.12.6 demonstrates this is incomplete: altruism is NOT a contradiction of evolution — it IS evolution's natural result. Humans are fundamentally peaceful organisms in their natural state.

3

## Formally and Mathematically Proved in This Paper

Both conclusions — the optimality of peace (Theorem 3, LUB convergence) and the evolutionary superiority of altruism ( § 2.12.6,  $\alpha_i^* = \text{argmax } F_i$ ) — are formally and rigorously proved. These are not philosophical claims; they are mathematical theorems.

# Boundary Effects: Maximum Leverage at the TCZ Boundary

$$\gamma_{\text{surface}} \sim \int \partial\Omega U(r) dS$$

**Maximum sensitivity at the boundary, not the interior**

*Plain language: The internal structure of a liquid is not directly visible. But where the liquid meets air — the boundary — the internal structure becomes observable as surface tension. The boundary is where asymmetry is made manifest.*

- Observable physical phenomena emerge where internal interactions become asymmetric at the boundary
  - Einstein 1901:  $\gamma_{\text{surface}} \sim \int \partial\Omega U(r) dS$  — surface tension is a boundary integral, not a volume property
- Cognitive analogy: the TCZ boundary  $\partial\text{TCZ}$  is where small perturbations produce maximal behavioral change
  - Interior TCZ: stable, high resistance to perturbation — requires large intervention to produce effect
  - Boundary  $\partial\text{TCZ}$ : unstable equilibrium, minimal intervention suffices — phase-transition-like sensitivity
  - Operational implication: target  $\partial\text{TCZ}$ , not the interior —  $u^*_{\text{cw}}$  minimizes  $|V-\theta|^2$ , not  $V$  itself

$\gamma \sim \int \partial\Omega U(r) dS$ : Surface tension  $\gamma$  = integral ( $\int$ ) of molecular forces  $U(r)$  only over the BOUNDARY surface ( $\partial\Omega$ ), not the whole volume. Plain meaning: the interior is invisible — only the BOUNDARY reveals the structure as an observable effect. For your comfort zone: the smallest nudge at  $\partial\text{TCZ}$  (the edge) produces the biggest behavioral change.

# Altruism as Abstract TCZ Expansion: Motivation and Formulation

56

The Shared-TCZ framework explains proximity-based cooperation. But humans act altruistically toward strangers, nations, future generations. A formal extension is needed.

## Abstraction parameter $\alpha_i \in \mathbb{R}^+$ : determines breadth of cognitive concern

Small  $\alpha_i$  = local/kin interactions. Large  $\alpha_i$  = global/abstract/universal considerations.

$$S_i(x; \alpha_i) = \sum_{j \neq i} w_{\{ij\}}(\alpha_i) d(x_i, x_j) \quad \text{where} \quad w_{\{ij\}}(\alpha_i) = \exp(-\text{dist}(i,j) / \alpha_i)$$

$$\pi_i = \operatorname{argmin} \int_0^T [ V_i(x_i, t) + \sum_j \gamma_{\{ij\}} S_{\{ij\}}(x_i, x_j) + \eta_i A_i(x_i) ] dt$$

$A_i(x_i) = \text{abstraction potential} \mid \eta_i = \text{abstraction sensitivity of agent } i$

- Altruism = control bias toward higher abstraction levels: Self selects trajectories that stabilize shared TCZs across distant and unknown agents
  - Altruistic behavior is optimal under a broader evaluative domain — not irrational. It emerges from high- $\alpha$  dynamics: mathematically the natural consequence of abstraction elevation.

$S_i$  formula: how much you 'include others in your concern'.  $d(x_i, x_j)$  = difference between your situation and person  $j$ 's.  $w_{ij}$  grows with  $\alpha_i \rightarrow$  distant/unknown people get more weight.  $\pi_i = \operatorname{argmin}[\dots + \eta_i A_i]$ : your control policy minimizes abstraction potential  $A_i$ , so you naturally move

**Altruism is formally defined: not a moral anomaly but the mathematically optimal behavior under high abstraction.**

# Evolutionary Perspective: Altruism Is Not Irrational — It Is Optimal

57

## Fitness function incorporating abstraction:

$$F_i = \int_0^T (-V_i + \beta_i \cdot \text{Share} + v_i \cdot G_i) dt \Rightarrow \alpha_i^* = \operatorname{argmax} F_i$$

*$v_i$  = fitness gain coefficient |  $G_i$  = group fitness term |  $\beta_i$  = sharing benefit coefficient*

- This explains why humans evolved large-scale cooperation, moral systems, and abstract altruism — high  $\alpha$  is evolutionarily selected
- Selfish Gene theory (Dawkins): humans are genetically selfish — corresponds to Theorem 2 (low- $\alpha$ ) dynamics only
- This paper (§ 2.12.6): altruism is NOT a contradiction of evolution — it IS evolution; high- $\alpha$  strategies are evolutionarily stable
  - Cooperation, moral systems, and abstract altruism are the mathematically predicted result of natural selection

$F_i = \int(-V_i + \beta_i \cdot \text{Share} + v_i \cdot G_i)dt$  means: Fitness = sum over time of (1) less personal discomfort, (2) cooperation bonus, (3) group survival bonus. 'argmax  $F_i$ ' = evolution selects the  $\alpha_i^*$  that maximizes ALL THREE simultaneously. High- $\alpha$  (broad altruism) scores best on all three terms.

**§ 2.12.6 demonstrates:** The Selfish Gene hypothesis is incomplete. Humans are NOT genetically predisposed to conflict. Altruism is the evolutionarily optimal strategy. Conflict and war are unnatural low- $\alpha$  states. Humans in their natural state are fundamentally peaceful.

# The Most Effective Cognitive Warfare: Promoting Peace Through TCZ-High-Shared

58

**Key Insight: The most effective cognitive warfare strategy is to promote convergence toward Shared-High-TCZ — which is the path of peaceful coexistence.**

## Shared-Low-TCZ (Fragmentation)

- Narrow shared stability space
- Low abstraction ( $\alpha$  small)
- Fragmentation, conflict, war
- Short-term, unstable
- High maintenance cost

## Shared-High-TCZ (Integration)

- Expanded shared stability space
- High abstraction ( $\alpha$  large, LUB)
- Cooperation, alliance, peace
- Long-term, self-sustaining
- Low maintenance cost

LUB = Least Upper Bound: the most abstract concept still including everyone's comfort zone. Theorem 3 (proved): raising shared abstraction  $\alpha \rightarrow$  LUB convergence = Shared-High-TCZ = peace / alliance. Mathematically stable, self-sustaining, evolutionarily optimal ( § 2.12.6).

**Theorem 3 proof:** Elevating shared abstraction  $\alpha$  produces Shared-High-TCZ convergence (LUB). Mathematically stable, self-sustaining, evolutionarily optimal. **Promoting peace is not idealism — it is the optimal control strategy.**

**§ 2.12.6:**  $F_i = \int_0^T (-V_i + \beta_i \cdot \text{Share} + v_i \cdot G_i) dt \Rightarrow \alpha_i^* = \operatorname{argmax} F_i$  — evolution selects high- $\alpha$  (altruism) as optimal

## Unified Setup

$$\dot{z}=F(z), \Phi(z), \Omega\theta:=\{z \mid \Phi(z)\leq\theta\}$$

If  $\nabla\Phi(z)\cdot F(z) \leq -\alpha(\Phi(z)-\theta)$  for all  $z \notin \Omega\theta$ , then  $\text{dist}(z(t), \Omega\theta) \rightarrow 0$ .

All three theorems are specializations of this single Lyapunov-descent structure.

## Three Specializations

$\Phi=V \rightarrow$  Theorem 1 (individual TCZ)

$\Phi=\mathcal{L}=\sum_i V_i + \sum_{ij} \gamma_{ij} S_{ij} \rightarrow$  Theorem 2 (shared TCZ)

$\Phi=\mathcal{L}A \rightarrow$  Theorem 3 (LUB)

$$A(x) = 0 \iff \phi(x) = T = \text{LUB}(W_1, \dots, W_N)$$

The three theorems are not independent results — they are successive extensions of a single Lyapunov-based stability structure.

## Cross-Domain Universality

Same convergence principle:

Physics  $\rightarrow$  energy decreases | Cognition  $\rightarrow$  evaluative function decreases

FEP  $\rightarrow$  free energy decreases | Predictive processing  $\rightarrow$  prediction error decreases

All share the same convergence principle. This framework shows convergence is accompanied by abstraction — a unified Lyapunov structure across physical, cognitive, and social systems.

## Statement & Setup

*Lyapunov:*  $\Phi(x) = V(x), \quad \Omega_{\vartheta} := \{x \mid V(x) \leq \vartheta\}$

*Condition:*  $\nabla V(x) \cdot F(x) \leq -\alpha (V(x) - \vartheta), \quad \alpha > 0, \quad \forall x \notin \Omega_{\vartheta}$

*Conclusion:*  $x^*(t) \rightarrow \text{TCZ}(x_0)$

## Proof

*Let*  $y(t) = V(x(t)) - \vartheta$ . *Then*  $\dot{y} \leq -\alpha y$ .

*Comparison principle:*  $V(x(t)) - \vartheta \leq (V(x_0) - \vartheta) e^{-\alpha t}$

$\Rightarrow \text{dist}(x(t), \Omega_{\vartheta}) \rightarrow 0$ . *Forward invariance:*  $dV/dt \leq 0$  on  $\partial\Omega_{\vartheta}$ . ■

*Stability arises from the decrease of internal evaluative instability.*

## Pairwise Lyapunov & Gradient Structure

$$\mathcal{L}(x) = \sum_i V_i(x_i) + \sum_{i,j} \gamma_{ij} S_{ij}(x_i, x_j), \quad \gamma_{ij} > 0$$

$$S_{ij} = \|x_i - x_j\|^2 \quad (\text{pairwise cognitive deviation})$$

$$\partial \mathcal{L} / \partial x_i = \nabla V_i + 2 \sum_{j \neq i} (\gamma_{ij} + \gamma_{ji})(x_i - x_j) \quad [\text{strong connectivity} \Rightarrow \gamma^{\text{eff}} > 0]$$

## Proof & Conclusion

$$\mathcal{L}(x(t)) - \vartheta \leq (\mathcal{L}(x_0) - \vartheta) e^{-\alpha t} \Rightarrow \text{dist}(x(t), \Omega \vartheta) \rightarrow 0$$

$$x \in \Omega \vartheta, \text{ all terms } \geq 0: \quad V_i \leq \vartheta \Rightarrow x_i \in \text{TCZ}_i \quad (\text{definition of TCZ})$$

$$\text{LaSalle invariant set: } \gamma_{ij} S_{ij} \rightarrow 0 \Rightarrow \|x_i - x_j\| \rightarrow 0 \text{ for all } i, j$$

$$\text{Therefore: } x \rightarrow \cap_i \text{TCZ}_i = \text{TCZ}^{\text{low\_shared}} \subseteq \text{TCZ}_{\text{shared}} \quad \blacksquare$$

## Lattice Setup & Abstraction Potential

Joint state space  $X = \prod_i X_i$ . Subsumption lattice  $(L, \leq)$  with top  $T = \text{LUB}(W_1, \dots, W_N)$

Abstraction map  $\varphi : X \rightarrow L$ . Potential  $A : X \rightarrow \mathbb{R}_{\geq 0}$  satisfies:

(A1)  $A(x) = 0 \iff \varphi(x) = T$     (A2)  $A(x) > 0$  when  $\varphi(x) \neq T$

(A3)  $A$  strictly decreases along upward lattice moves:  $\varphi(x') > \varphi(x) \implies A(x') < A(x)$

## Lyapunov Function & Statement

$\mathcal{L}_A(x) = \sum_i V_i(x_i) + \sum_{i,j} \gamma_{ij} S_{ij}(x_i, x_j) + \eta A(x)$ ,  $\gamma_{ij} > 0$ ,  $\eta > 0$

Condition:  $\nabla \mathcal{L}_A \cdot F \leq -\alpha (\mathcal{L}_A - \vartheta A)$  for all  $x \notin \Omega_A$

Conclusion:  $x^*(t) \rightarrow \text{LUB}(W_1, \dots, W_N)$

Note: (A1) is a design condition —  $A$  must faithfully encode distance from LUB.

## Steps 1 & 2

Step 1 (Descent):  $y_A = \mathcal{L}_A(x(t)) - \vartheta_A \leq (\mathcal{L}_A(x_0) - \vartheta_A) e^{-\alpha t}$

$\Rightarrow \text{dist}(x(t), \Omega_A) \rightarrow 0$  as  $t \rightarrow \infty$

Step 2 (Invariance):  $\nabla \mathcal{L}_A \cdot F \leq 0$  on  $\partial\Omega_A \Rightarrow \Omega_A$  is forward invariant

## Steps 3 & 4

Step 3 (LaSalle):  $x \in \Omega_A$ , all terms  $\geq 0 \Rightarrow V_i \leq \vartheta_A \Rightarrow x_i \in \text{TCZ}_i$

Invariant set:  $\gamma_{ij} S_{ij} \rightarrow 0 \Rightarrow \|x_i - x_j\| \rightarrow 0$  and  $A(x) \rightarrow 0$

Step 4 (Conclusion):  $A(x) = 0 \Leftrightarrow \varphi(x) = T = \text{LUB}(W_1, \dots, W_N)$

$\Rightarrow x^*(t) \rightarrow \text{LUB}(W_1, \dots, W_N) \blacksquare$

***“The future battlespace is not physical terrain,  
but the structure of cognitive potential  
landscapes.”***

---

Hideto Tomabeche | tomabeche@crl.co.jp

CyLab CMU · C5I Center GMU · Cognitive Research Laboratories

*Paper: “A Unified Theory of Latent Potentials: Homeostasis and Cognitive Warfare” (April 4, 2026)*

**Q & A**