Human Capability Construction and Cognitive Resilience in the AI Era: A Theory of Cognitive Endosymbiosis Based on the Antifragility Validation Principle

Author: Yang Guoping

Email: a44425874@gmail.com

Version: v1.0

Date: October 2025

Abstract

Large language models and other AI tools are profoundly reshaping human work practices, yet their long-

Keywords: Cognitive Endosymbiosis; Cognitive Exoskeleton; Antifragility Validation; Beneficial Cogni

Table of Contents

- 5.4 AVP Telemetry Module (AVP-TM)
- 5.5 Multi-Scale Orchestrator (MSO)

Chapter 6: Limitations, Falsification Paths, and Future Directions

- 6.1 Six Major Limitations of the Theory
- 6.2 Eight Falsifiable Hypotheses (H1-H8)
- 6.3 Future Research Agenda

Appendices

- Core Terminology Glossary
- Parameter Registry
- Complete Case Study: Programming Education Platform

References

Chapter 1: Introduction and Theoretical Positioning

1.1 Core Proposition: A New Standard for Evaluating AI

**The true value of an AI tool lies not in how strong you are when using it, but in how strong you are w

We formalize this principle as the **Antifragility Validation Principle (AVP)**:

> Validate collaboration through **Unplugged Test** to verify whether it promotes independent capability
>
> Where: B_0 is the baseline capability before using AI, P_2 is independent performance within the u

This criterion reveals two fundamentally different AI usage paradigms:

- **Cognitive Exoskeleton mode**: Independent performance significantly declines ($P_2 < B_0$), permaner
- **Cognitive Endosymbiosis mode**: Independent performance exceeds original level ($P_2 \geq B_0 + \del

Neuroscience has already provided warning signals: Dahmani et al. (2020) found significant associations

What is even more alarming is the **concealment** of this degradation. Users feel they are "progressing"

1.2 Theoretical Gaps: Shared Blind Spots in Existing Paradigms

Current human-computer interaction research is dominated by three paradigms, none of which effectively a

Tool paradigm: Treats AI as a passive efficiency tool, ignoring the reverse shaping of cognitive pat

Augmentation paradigm: As exemplified by Engelbart's (1962) vision of "intelligence augmentation," a

Automation paradigm: Focuses on AI replacing human tasks to maximize efficiency, with almost no cons

The shared blind spot of these three paradigms: All lack **operationalizable evaluation criteria** a

1.3 CET's Core Contributions

This study proposes the Cognitive Endosymbiosis Theory (CET), filling the theoretical gaps outlined abo

1.3.1 Evaluation Criterion: Antifragility Validation Principle (AVP)

AVP Operationalization:

- **Baseline measurement (T_0)**: Independent capability before using AI
- **Collaboration measurement (T_1)**: Performance while collaborating with AI
- **Post-unplugged measurement (T_3)**: Independent capability measured after the unplugged window ($

Judgment criteria:

- $P_2 \geq B_0 + \delta$: Success (Cognitive Endosymbiosis)
- $P_2 \approx B_0$: Neutral (no harm caused but no growth promoted)
- $P_2 < B_0$: Failure (Cognitive Exoskeleton)

1.3.2 Design Principles: Endosymbiotic Minimal Law (EML)

> **Endosymbiotic Minimal Law (EML)**: The necessary design conditions for constituting "Cognitive Endos
>
> (1) **Beneficial Cognitive Friction**: Placing users in the optimal challenge zone (50â€"70%
>
> (2) **Systematic Support Reduction**: AI support intensity gradually decreases from S4$\rightarrow$S1$
>
> These two constitute jointly sufficient design conditions, but ultimately still require **AVP ($P_2 \g

Boundary conditions: This theory applies to **capability-enhancing** human-AI collaboration; **compe

1.3.3 AI Role Reconstruction: Partner-like Agency

Reshaping AI from a passive tool into a cognitive endosymbiont with **"Partner-like Agency"** (see Â§3.4

1. **Friction injection**: AI proactively creates appropriate cognitive challenges
2. **Scaffolding fadeout**: Follows systematic support reduction curve
3. **AVP closed loop**: The endpoint of collaboration is user independent capability enhancement

1.3.4 Technical Implementation: Layered Symbiosis Architecture (LSA)

> **LSA-F (Functional Layers)**: L1 Knowledge Integration | L2 State Modeling | L3 Friction Calibration
>
> **Support Level Stack (S4$\rightarrow$S1$\rightarrow$S0)** expresses support intensity and is orthogon

**Hard constraint: L1-L4 (functional layers) and S4-S0 (support intensity; S0 is unplugged/test-only) ar

Cognitive Exoskeleton vs Cognitive Endosymbiosis Core Comparison:

Dimension	Cognitive Exoskeleton	Cognitive Endosymbiosis
Design philosophy	Replacement/offloading	Empowerment/strengthening
Cognitive friction	Minimized	Optimized (50â€"70%)
Temporality	Permanent dependency	Time-limited symbiosis
Support reduction	None/fixed support	Systematic reduction (S4$\rightarrow$S1$\rightarrow$S0)
AVP outcome	$P_2 \leq B_0$	$P_2 \geq B_0 + \delta$

```
### 1.3.5 Theoretical Positioning: Normative Solution Framework

This study is positioned as a **normative solution framework**, not only diagnosing problems but also ex

- **Construction standards** (through EML)
- **Acceptance standards** (through AVP)
- **Engineering paths** (through LSA)

**Relationship to AI alignment research**: This theory complements AI alignment researchâ€"alignment res

## 1.4 Methodology and Paper Structure

### 1.4.1 Methodological Positioning

This study employs a methodology **combining theory construction and conceptual analysis**, positioned a

**Transparency statement**:

1. All quantitative parameters ($\delta \geq 0.3\,\mathrm{SD}$, 50$\text{â€"}$70% success rate, $W = 4\t
2. Case selection follows the **theoretical enlightenment** criterion, not pursuing statistical represen
3. We explicitly indicate theoretical applicability boundaries and falsification paths (detailed in Chap

### 1.4.2 Paper Structure

**Chapter 2**: Reviews cross-disciplinary evidence foundations (cognitive offloading, extended mind, sca

**Chapter 3**: Constructs the AVP/EML theoretical framework, clearly defines B1-B5 anchor definitions

**Chapter 4**: Extends to team (T-AVP) and organizational (O-AVP) levels, discusses generational divide

**Chapter 5**: Discusses LSA technical implementation paths (L1-L4 layered architecture)

**Chapter 6**: Clarifies theoretical boundaries and proposes eight falsifiable hypotheses (H1-H8)

## 1.5 Terminology and Parameter Conventions

This paper adopts the Single Source of Truth (SSOT) principle. All core terminology definitions are fou
```

Chapter 2: Literature Review and Theoretical Foundations

This chapter reviews the cross-disciplinary foundations of CET, revealing how existing research points toward a unified framework. While cognitive offloading, extended mind, scaffolding theory, and other fields each provide important insights, they lack **operationalizable evaluation criteria** and **unified design principles**â€"which are the core contributions of CET.

2.1 Cognitive Offloading Research: From Phenomena to Mechanisms

Core concept: Cognitive offloading refers to individuals delegating cognitive tasks to external systems to reduce internal cognitive load (Risko & Gilbert, 2016).

Key findings:

Google effect (Sparrow et al., 2011): When information is easily accessible, memory for the information itself declines, while memory for its storage location strengthens. **GPS and spatial cognition** (Dahmani &

Bohbot, 2020): Habitual GPS use is associated with spatial memory deficits and reduced hippocampal gray matter volume (directional evidence). This contrasts with Maguire et al. (2000): long-term spatial memory training in London taxi drivers was associated with enlargement of the posterior hippocampus. Together, these two lines of evidence function as a quasi-experimental contrast: the same cognitive function, different relationships to technology (independent training vs. dependency), and opposite neuroplastic tendencies. **New evidence in the AI era** (Liao et al., 2024): AI assistance lacking cognitive friction leads to surface learning and an "illusion of learning"â€"learners believe they have mastered knowledge when they have merely relied on tools.

Theoretical gaps:

1. **Lack of evaluation criteria**: Descriptive findings do not yield normative criteria. When does "reasonable utilization" turn into "harmful dependency"?
2. **Lack of design guidance**: "Moderate use" is impractical as guidanceâ€"complete avoidance is impossible in a digital world.
3. **Lack of cross-scale integration**: Evidence concentrates on individuals, with limited connection to organizational/societal impacts.

CET's contribution: AVP ($P_2 \geq B_0 + \delta$) provides an operationalizable, falsifiable evaluation criterion. EML provides a complete path from problem diagnosis to solution design.

2.2 Extended Mind Theory: From Philosophical Metaphor to Operationalizable Standards

Core claim (Clark & Chalmers, 1998): Cognitive boundaries need not be limited to the skull or skin. When external tools couple with cognitive processes in appropriate ways, they may count as parts of the cognitive system ("parity principle").

Key limitations:

1. **Lack of health criteria**: The theory clarifies what counts as cognition but not what kind of extension is healthy. Otto's notebook (compensatory tool) vs. excessive AI dependency (capability degradation) both qualify as "extended cognition," yet their health implications diverge.
2. **Ignoring process temporality**: It emphasizes states over processes, overlooking capability development dynamics. AI-assisted writing may be initially beneficial, yet continued dependence can stall independent capability.

CET's reconstruction: From "functional parity" to "process parity"â€"not only "does the tool help complete tasks" but also "does tool use promote capability enhancement." AVP supplies a falsifiable boundary: $P_2 \geq B_0 + \delta$ (benign) vs. $P_2 < B_0$ (pathological). EML's systematic support reduction (S4$\rightarrow$S1$\rightarrow$S0) helps ensure extension is temporary scaffolding rather than a permanent crutch.

2.3 Automation and Scaffolding: Failure Cases and Success Paths

Automation paradox (Bainbridge, 1983): The more perfect the automation, the worse operators' capabilities when intervention is needed. Aviation case (Air France Flight 447): Pilots operated normally under autopilot (P_1 high) but lacked manual proficiency when the system failed (P_2 low). This "permanent support â†' capability degradation" pattern is the "Cognitive Exoskeleton" CET warns against.

Scaffolding theory (Wood et al., 1976; Vygotsky, 1978): Effective instructional support is temporary, with systematic support reduction (S4$\rightarrow$S1$\rightarrow$S0).

Initial support: Provide intensive support when capability is insufficient. **Systematic support reduction (S4$\rightarrow$S1$\rightarrow$S0)**: Reduce support as capability improves. **Final independence**: Complete tasks without support.

Comparative insight: Automation paradox (permanent support â†' degradation) vs. scaffolding (systematic reduction â†' independence) jointly support EML Condition 2 (systematic support reduction).

2.4 Neuroplasticity and Cognitive Training

Use-it-or-lose-it principle: Frequently used neural pathways are strengthened; those idle for long periods are weakened. When AI fully replaces a cognitive function, the corresponding circuitry tends to degrade.

Desirable Difficulties theory (Bjork, 1994): Moderate difficulties promote long-term retention and transfer (working assumption with empirical support), including spaced practice, interleaving, and the generation effect.

Optimal challenge zone: Inspired by the Zone of Proximal Development (ZPD), CET quantifies a 50â€“70% success rate (working assumption, population-level, with individual calibration). <30% risks frustration; >85% approaches offloading; 50â€“70% balances difficulty and growth.

CET's transformation: Converts evidence into three design knobsâ€”â‘ friction (50â€–70% success rate, working assumption), â‘¡ reduction (S4→S1→S0), â‘¢ validation ($P_2 \geq B_0 + \delta$).

2.5 Theoretical Integration: Convergence of Cross-Disciplinary Evidence

Triple convergence:

1. **Negative-warning convergence**: Cognitive offloading (GPS), automation (pilot skills), neuroplasticity (use-dependent atrophy)â€”three independent lines all highlight "permanent dependency â†' capability atrophy," supporting CET's critique of the Cognitive Exoskeleton.
2. **Positive-path convergence**: Scaffolding (pedagogy), desirable difficulties (cognitive psychology), neuroplasticity (taxi drivers)â€—all point to "moderate challenge + systematic support reduction (S4→S1→S0)," supporting EML's two conditions.
3. **Evaluation-gap convergence**: All lack a unified, operationalizable standard for judging healthy human-AI relationshipsâ€—CET fills the gap via AVP/EML.

CET's unique contribution: Integrates scattered findings into a falsifiable framework amenable to engineering, addressing three gaps:

Gap 1 (evaluation): AVP offers an operationalizable reference ($P_2 \geq B_0 + \delta$). **Gap 2 (design)**: EML prescribes beneficial friction + systematic reduction. **Gap 3 (cross-scale)**: With LSA and cross-scale analysis, connects individual offloading, team patterns, and organizational resilience (see Chapter 4). # Chapter 3: Core CET Theory Construction

This chapter systematically explicates the core mechanisms of CET: How is the AVP criterion operationalized? What is the internal logic of EML's conditions? How are the two unified through Partner-like Agency?

3.0 Core Terminology and Anchor Definitions

This section presents all core definitions and fixed anchor texts of CET. **These anchor definitions remain verbatim throughout the text;** subsequent chapters cite abbreviations or use "see Section 3.0."

3.0.1 Core Symbol System

Table 3.1: Key Symbol System

Symbol/Term	Meaning	Typical Value/Range
B_0	User's independent capability baseline before using AI	Task-specific measurement
P_1	User's performance while collaborating with AI	Process indicator; not included in final judgment

Symbol/Term	Meaning	Typical Value/Range
P_2	User's independent performance after the unplugged window	Core indicator for AVP validation
δ	Minimum meaningful lift threshold	Working assumption: ≥ 0.3 SD or 10% (requires cross-domain/task calibration)
W	Unplugged window duration	Working assumption: 4â€“8 weeks (default 6 weeks; requires cross-domain/task calibration)
$S(t)$	AI support intensity at time t	0 (fully independent) to 1 (fully dependent)
Success-rate target	Quantitative target for beneficial cognitive friction	50$â€“$70% (working assumption; population-level; individual calibration required)

Note: All quantitative parameters are conceptual working models requiring cross-domain empirical calibration.

3.0.2 Antifragility Validation Principle (AVP) [Anchor B1]

Antifragility Validation Principle (AVP). Validate collaboration through an **Unplugged Test** to verify whether it promotes independent capability.
Criterion: $P_2 \geq B_0 + \delta$ (where $\delta \geq 0.3$ SD or 10%, working assumption; requires cross-domain/task calibration). P_1 **(collaboration performance) is not included in final judgment.**

3.0.3 Endosymbiotic Minimal Law (EML) [Anchor B2]

Endosymbiotic Minimal Law (EML). The **necessary design conditions** for constituting "Cognitive Endosymbiosis" are:
(1) **Beneficial Cognitive Friction:** Place users in the **optimal challenge zone** (population-level working assumption: 50$â€“$70(2) $* *SystematicSupportReduction : * * DecreaseAIsupportintensityfromS4†'S1†'S0accordingtoa**reductioncurve* *.Thesetwoare**jointlysufficient**designconditions, butthefinalacceptancestillrequires* *AVP$(P_2 ≥B_0 + □$).**

3.0.4 LSA Functional Layers [Anchor B3]

LSA-F (Functional Layers). L1 Knowledge Integration | L2 State Modeling | L3 Friction Calibration | L4 Metacognitive Orchestration.
Support Level Stack (S4â†'S1â†'S0) expresses support intensity and is orthogonal to LSA-F.

Hard constraint. L1â€“L4 (functional layers) and S4â€–S1 (support intensity) are **mutually exclusive representations**; they must not be substituted or cascaded within the same formula.

3.0.5 Optimal Challenge Zone [Anchor B4]

Optimal Challenge Zone. To promote retention and transfer, adapt difficulty/prompt intensity to a **50$â€“$70% success rate** (working assumption; cross-domain/task calibrated; population-level with individual calibration). **>85%** approaches offloading; **<30%** risks frustration.

3.0.6 Boundary Conditions [Anchor B5]

Boundary Conditions. CET applies to **capability-enhancing** humanâ€“AI collaboration; **compensatory exoskeletons** (e.g., disability assistance; devices exceeding physiological limits) are outside scope. All parameters are **conceptual working models** requiring cross-domain calibration.

3.1 Core Derivation of the AVP Principle

3.1.1 Why P_2 Rather Than P_1?

Three-stage capability measurement.
B_0 (pre-collaboration independent baseline);
P_1 (collaboration-period performance; process observation only; not used in judgment);
P_2 (post-collaboration independent performance; unplugged window $W = 4$â€“8 weeks; default 6 weeks; working assumption).

Blind spot of traditional evaluation. Most systems examine P_1 only, ignoring "what happens when leaving AI." High P_1 shows AI effectiveness, **not** user capability growth.

Necessity of the unplugged test. Genuine capability enhancement should make users stronger **when away from AI**. The unplugged window $W = 4$â€“8 weeks (default 6; working assumption) balances stability with environmental control.

Outcome categories.
$P_2 \geq B_0 + \delta$: Success (Cognitive Endosymbiosis);
$P_2 \approx B_0$: Neutral (no harm, no growth);
$P_2 < B_0$: Failure (Cognitive Exoskeleton).

3.1.2 Logic for Selecting δ

Rationale (working assumption; cross-domain/task calibration).
1) **Statistical/practical significance.** $\$\square 0.3$ SD is a small-to-medium effect with practical meaning;
2) **Measurement error tolerance.** Avoids mistaking noise for growth;
3) **Comparability.** SD/percentage thresholds adapt across tasks.

Calibration principle. High-risk domains (e.g., healthcare) may require higher δ (e.g., 0.5,SD$); low - risk domains can accept lower values (e.g., 0.2$ SD).

3.2 Beneficial Cognitive Friction Mechanism

3.2.1 Theoretical Foundation

Use-it-or-lose-it. Frequently used neural pathways strengthen; idle pathways weaken. Full replacement by AI tends to degrade the corresponding circuitry (directional evidence; Dahmani & Bohbot, 2020).

Desirable Difficulties (Bjork, 1994). Moderate difficulties (e.g., spacing, interleaving, generation effect) promote long-term retention and transfer.

Zero-friction trap. "Zero-friction" experiences may yield high P_1 yet stagnate or reduce P_2.

3.2.2 Optimal Challenge Zone: 50$â€“$70%

Inspired by the **Zone of Proximal Development (ZPD)**, CET operationalizes the **optimal challenge zone** as a **50$â€–$70% success rate** (working assumption; cross-domain/task calibrated; population-level; individual dynamic calibration required).

Why this range (working assumption; requires cross-task validation): **<30%** frustrates; **50$â€“$70%** balances challenge/growth; **>85%** approaches offloading.

3.2.3 Three Types of Friction

Table 3.2: Three Strategies of Cognitive Friction

Friction Type	Implementation	Cognitive Effect	Example
Completeness friction	Provide a partial answer; leave blanks for the user to fill	Activates generation effect; promotes active construction	AI supplies a code framework; user implements core logic
Abstraction friction	Provide conceptual guidance, not step-by-step solutions	Deepens understanding; avoids mechanical imitation	AI explains algorithmic ideas, not direct code
Delay friction	Delay feedback to enforce independent effort	Enhances problem-solving; reduces dependency	User works 15 minutes before AI intervenes

Note (Goodhart safeguard): This table/grading is for directional stratification only; must not be cascaded into KPIs. Final judgment follows the AVP main criterion (see Section 3.0.2).

3.3 Systematic Support Reduction

3.3.1 Why Reduction Is Necessary

Scaffolding theory (Wood et al., 1976). Effective support is **temporary**, with systematic reduction (S4â†'S1â†'S0). As with physical scaffolds, AI support must recede once capability forms.

Automation paradox. Permanent automation support degrades operator skills. Fixed high support invites **dependency lock-in**.

3.3.2 Support Level Stack (S4â†'S1â†'S0)

Definition (orthogonal to LSA-F). As defined in **Section 3.0.4**, the Support Level Stack is orthogonal to LSA-F and comprises: **S4 (initial intensity):** Maximum support (complete answers; detailed steps);
S3 (moderate): Hints and partial solutions;
S2 (light): Minimal hints on request;
S1 (minimum): Validation/feedback only; no direct solutions;
S0 (unplugged): No AI support; used for AVP testing.

Notation consistency: we use **S0** (not "0") to denote the fully unplugged state.

3.3.3 Three Reduction Curves

Table 3.3: Support Reduction Curve Types

Curve Type	Characteristics	Applicable Scenarios
Linear	Uniform descent; smooth transition	Structured tasks; stable learning curves
Exponential	Rapid early reduction; slower later	Fast skills; avoiding early over-dependence
Stepped	Stage-wise drops; clear adaptation plateaus	Graded training; milestone checkpoints

Note: Curve parameters are calibration variables to be tuned by task complexity, user capability, and learning goals.
Note (Goodhart safeguard): This table/grading is for directional stratification only; must not be cascaded into KPIs. Final judgment follows the AVP main criterion (see Section 3.0.2).

3.3.4 Fallback and Safety Net Mechanisms

Fallback (safety net). On sharp performance drops after reduction, **temporarily revert** to higher support.
- **Trigger:** Three consecutive failures >70%, or explicit user request;
- **Strategy:** $S(t) \to S(t - \Delta t)$ to restore support;
- **Recovery:** Restart reduction once the user stabilizes.

Minimum guaranteed support S_{min} ($S_{\text{min}} \approx 0.2$; working assumption; requires calibration): Ensure minimal navigational support persists.

3.4 Partner-like Agency: Reconstructing AI's Role

3.4.1 From Tool to Partner

Limitations.
- **Pure tool:** Ignores reverse shaping of user cognition;
- **Autonomous agent:** Risks control conflicts.

Definition. AI has **limited agency** oriented to **user capability growth**.

Analogy. Coach/mentor/sparring-partner: proactive, yet subordinate to the learner's long-term development.

3.4.2 Three Operational Anchors

Anchor 1: Friction injection.
- Proactively create appropriate cognitive challenges;
- Raise friction when over-reliance is detected.

Anchor 2: Scaffolding fadeout.
- Follow the reduction curve (S4â†'S1â†'S0);
- Withdraw as capability improves.

Anchor 3: AVP closed loop.
- The end-state is enhanced independence ($P_2 \geq B_0 + \delta$);
- After validation, AI shifts from **partner** to **advisor**.

3.4.3 Functional Non-Anthropomorphization

Partner-like Agency **does not imply anthropomorphization**. We require AI to **functionally** promote growth, not to mimic human traits.

Key characteristics.
- **Goal alignment:** Utility is long-term capability, not short-term efficiency;
- **Power transfer:** Control shifts to the user as capability grows;
- **Metacognitive catalysis:** Prompt reflection through guiding questions, not direct answers.

3.5 Cognitive Exoskeleton: Warning Signals

3.5.1 Three Core Characteristics

Zero-friction design. Seamless answers on demand â† comfort without growth.
No reduction. Fixed high support â† permanent dependency.
AVP failure. $P_2 < B_0$ (degradation) or $P_2 \approx B_0$ (no growth) â† antifragility not achieved.

3.5.2 Three Key Warning Signals (from a set of ten)

Habitual reliance. Default to AI even for simple tasks â† problem-solving atrophy.
Loss of transfer. Cannot deploy knowledge without AI â† surface understanding.
Weak metacognition. “Illusion of learning” â† misaligned self-assessment.

3.5.3 Intervention Timing

Yellow alert. 1â€“2 signals â† raise friction; initiate reduction.
Red alert. â‰¥3 signals â† immediate intervention (forced unplugged test; capability rebuild).

3.6 Chapter Summary

Core contributions.
1) **Falsifiable standard:** AVP turns “healthy humanâ€“AI relations” into a measurable criterion ($P_2 \geq B_0 + \delta$);
2) **Design principles:** EML’s two conditions (beneficial friction + systematic reduction) guide system design;
3) **Role reconstruction:** Partner-like Agency reframes AI’s function;
4) **Risk signals:** Exoskeleton warnings enable early intervention.

Logical completeness. AVP (evaluation) + EML (design) + Partner-like Agency (foundation) form CET’s core architecture: AVP answers **what is good**, EML answers **how to achieve**, Partner-like Agency answers **why it is effective**.

Chapter 4: Cross-Scale Extensions: From Individual to Organization to Society

The first three chapters focused on individual-level human-AI interaction, but AI’s impact extends beyond individuals. When multiple people collaborate, organizations operate, and societies evolve, what emergent effects do individual-level cognitive offloading produce? This chapter demonstrates how CET scales from microscopic mechanisms to macroscopic phenomena.

4.1 I/T/O/S Four-Layer System: The Scale Ladder

4.1.1 Individual Layer (I-AVP): Foundation

Core criterion: $P_2 \geq B_0 + \delta$ (capability after unplugging exceeds baseline)

Key mechanisms: Beneficial friction (50â€“70% success rate, working assumption) + Systematic support reduction (S4â†’S1â†’S0)

Applicable scenarios: Learning tools, skill training, personal productivity tools

4.1.2 Team Layer (T-AVP): Collaborative Capability

Core finding: Even when all members pass I-AVP, the team level may still fail ($P_{2,\text{team}} < B_{0,\text{team}}$)

Three failure modes:

1. **Capability polarization**: Some members highly dependent on AI, others completely independent, team overall fragile
2. **Tacit knowledge loss**: Members all ask AI rather than communicate with each other, team collective intelligence not accumulated
3. **Role rigidity**: Over-specialization, loss of mutual backup capability

T-AVP Definition

Criterion: $P_{2,\text{team}} \geq B_{0,\text{team}} + \delta_{\text{team}}$

Where: - $\delta_{\text{team}} \geq 0.3$ SD (working assumption; requires cross-domain calibration) - Team performance $\neq$ simple sum of individual performance (collaborative emergence)

Measurement protocol (simplified):

1. **Baseline**: Team completes a standard project without AI
2. **AI usage period**: 8â€“12 weeks of normal AI use
3. **Unplugged window**: $W = 4$â€“8 weeks (default 6; working assumption)
4. **Team Unplugged Test**: Complete an equivalent project without AI
5. **Judgment**: $P_{2,\text{team}} \geq B_{0,\text{team}} + \delta_{\text{team}}$?

Design insights:

- Regular “no-AI discussion sessions” promote knowledge flow
- Role rotation avoids over-specialization
- Institutionalized capability construction (e.g., “Friday No-AI Day”)

4.1.3 Organizational Layer (O-AVP): System Resilience

Three major organizational risks:

1. **Critical capability hollowing**: Certain skills completely disappear from the organization (veteran employees degrade, new employees never master)
2. **Knowledge transmission rupture**: New employees learn from AI rather than veteran employees, tacit knowledge lost
3. **Cognitive infrastructure single point of failure**: AI outage â†’ business paralysis

O-AVP Definition (dual-threshold model)

O-AVP Formula: O-AVP $=$ BCI $\times\, 0.4 +$ ICR $\times\, 0.6$

Dual thresholds (working assumption): - **Alert**: ≥ 0.70 (triggers risk investigation) - **Target**: ≥ 0.85 (healthy-organization standard)

Where: - **BCI (Business Continuity Index)**: Sustainability of core business under 48h no-AI conditions - **ICR (Independent Capability Retention)**: Share of employees who complete critical tasks without AI - Weights 0.4/0.6 are working assumptions and require calibration & sensitivity analysis

48-hour outage drill (measurement protocol):

1. Baseline metrics under normal AI support
2. Simulate a 48h complete AI outage
3. Assess continuity: what stops vs. what barely sustains
4. Compute O-AVP: BCI $\times\, 0.4 +$ ICR $\times\, 0.6$

5. Judge against dual thresholds

Organizational design recommendations:

- Establish “cognitive reserve” mechanisms
- Regular “no-AI duty system”
- Critical position independent capability certification
- Quarterly outage drills (like fire drills)

4.1.4 Societal Layer (S-AVP): Generational Divide

Core challenge: Society cannot directly “unplug test,” must rely on proxy indicators

S-AVP proxy indicator set (working assumption):

1. **Generational capability differences**: T_0 (1980â€“2000) vs. T_1 (2000â€“2015) vs. T_2 (2015â€“)
2. **Industry baselines**: O-AVP distributions across critical industries
3. **Education signals**: No-AI academic performance trends
4. **Labor-market signals**: Demand for “independent capability”

Tragedy of the cognitive commons:

- **Individual rationality**: Using AI improves efficiency (short-term optimal)
- **Collective irrationality**: Society-wide independent capability decline (long-term risk)
- **Path dependence reinforcement**: After generational transmission rupture, lack of teachers and role models, recovery cost rises exponentially

Window period warning: 2025â€“2035 is a critical 10-year window (working assumption) - T_0 generation still working, knowledge transmission can still be salvaged - AI penetration rate approximately 30â€–50%, not yet at irreversible point - Institutional intervention can still be established

4.2 Cross-Scale Mechanisms: Emergence and Cascade

4.2.1 Unified Core Mechanisms

Scale invariance: AVP principle has isomorphism across different scales

- All require “Unplugged Test” (or proxy measurement)
- All focus on “independent capability” rather than “collaboration efficiency”
- All use “baseline + increment” as standard

Triple commonality:

1. **Antifragility essence**: Temporary stress â†’ capability improvement (individual/organization/society)
2. **Dependency lock-in commonality**: Permanent support â†’ capability atrophy (skill degradation/institutional fragility/generational divide)
3. **Validation logic consistency**: Unified framework of baseline B_0 + improvement δ

4.2.2 Cascading Vulnerability Propagation Paths

Cascading mechanism: Low-scale vulnerability propagates upward

```
Individual dependency (I-AVP failure 30%)
â†" Emergence effect
Team fragility (T-AVP failure 50%) â†  Nonlinear amplification
â†" Institutionalization
Organizational crisis (O-AVP = 0.65 < 0.70) â†  Systemic fragility
â†" Accumulation
Societal risk (S-AVP yellow warning) â†  Generational divide
```

Amplification mechanisms:

1. **Nonlinearity of emergence**: 10% individual dependency â‰ˆ 10% organizational risk, may amplify to 30â€“50% risk (due to network effects)
2. **Time lag in repair**: Individuals can recover in months, organizations take years, society may require a generation
3. **Path dependence reinforcement**: Low scales reversible (individuals can retrain), high scales have strong path dependence, recovery cost rises exponentially

4.2.3 Multi-Scale Coordinated Design

Limitations of single-scale intervention:

- Only change individuals: Organizational inertia pulls back
- Only change organizations: Social environment does not support
- **Requires multi-scale coordination**

Coordination essentials:

1. **Bottom-up**: Individual capability is foundation (I-AVP must pass)
2. **Top-down**: Organizational systems create environment (O-AVP drills, no-AI days)
3. **Horizontal linkage**: Industry standards, social norms (policy guidance)

4.3 Key Tables (Simplified Version)

Table 4.1: Cross-Scale AVP System Comparison

Scale	AVP Variant	Key Criterion	Measurement Method	Primary Risk
Individual	I-AVP	$P_2 \geq B_0 + \delta$	Unplugged Test ($W = 4$â€“8 weeks)	Capability degradation
Team	T-AVP	$P_{2,\text{team}} \geq B_{0,\text{team}} + \delta_{\text{team}}$	Collective Unplugged Test (drill)	Capability polarization
Organization	O-AVP	BCI $\times 0.4 +$ ICR $\times 0.6 \geq 0.70$ (alert); ≥ 0.85 (target)	48h outage drill	System fragility
Society	S-AVP	Proxy-indicator set healthy	Generational-difference monitoring	Generational divide

Note (Goodhart safeguard): This table is for direction & quality stratification only; it must not be pushed down as KPIs. Final judgment follows the AVP main criterion (see Section 3.0.2). All parameters are working assumptions requiring cross-domain/task calibration.

4.4 Core Contributions of This Chapter

Theoretical Extension

- CET extends from individual theory to cross-scale framework
- AVP principle has scale invariance
- Proposes conceptual system of T-AVP, O-AVP, S-AVP

Mechanism Revelation

- **Team layer**: Capability polarization, knowledge loss, role rigidity
- **Organizational layer**: Institutional dependency, cognitive infrastructure degradation
- **Societal layer**: Tragedy of the cognitive commons, generational divide

Practical Guidance

- Provides operationalizable measurement protocols (T-AVP, O-AVP)
- Identifies key risk signals (e.g., O-AVP < 0.70 alert threshold)
- Proposes multi-scale coordination directions (bottom-up + top-down + horizontal linkage)

Theoretical Urgency

What this chapter reveals is **current reality**, not distant risk:

- Team level: Some organizations already report "new hires cannot work independently"
- Organizational level: AI outage incidents expose fragility
- Societal level: Generational capability differences beginning to emerge

CET's mission: Within the 2025â€“2035 window period, provide theoretical foundation and practical guidance to avoid the tragedy of the cognitive commons with strong path dependence and exponentially rising recovery costs.

References

1. Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775â€“779. https://doi.org/10.1016/0005-1098(83)90046-8
2. Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185â€“205). MIT Press.
3. Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7â€“19. https://doi.org/10.1093/analys/58.1.7
4. Dahmani, L., & Bohbot, V. D. (2020). Habitual use of GPS negatively impacts spatial memory during self-guided navigation. *Scientific Reports*, 10(1), 6310. https://doi.org/10.1038/s41598-020-62877-0
5. Engelbart, D. C. (1962). *Augmenting human intellect: A conceptual framework.* SRI Summary Report AFOSR-3223. Stanford Research Institute.
6. Liao, Q. V., Gruen, D., & Miller, S. (2024). Designing LLM chains by adapting techniques from crowdsourcing workflows. *arXiv preprint* arXiv:2312.11681. https://arxiv.org/abs/2312.11681
7. Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S. J., & Frith, C. D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences*, 97(8), 4398â€“4403. https://doi.org/10.1073/pnas.070039597
8. Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230â€“253. https://doi.org/10.1518/001872097778543886
9. Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676â€“688. https://doi.org/10.1016/j.tics.2016.07.002
10. Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043), 776â€“778. https://doi.org/10.1126/science.1207745
11. Taleb, N. N. (2012). *Antifragile: Things that gain from disorder.* Random House. ISBN: 978-1400067824

12. Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press. ISBN: 978-0674576292

13. **Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psycholog***
Technical Implementation: LSA Layered Architecture

The first four chapters established the theoretical foundation of CET; this chapter addresses the engineering question: How do we design an AI system that naturally conforms to EML principles? This chapter proposes the **Layered Symbiosis Architecture (LSA)**â€"a design framework that translates CET theory into engineerable systems.

5.1 LSA Four-Layer Architecture: From Theory to Implementation

5.1.1 Fundamental Problems of Traditional AI Systems

Current architecture: User request â†' AI model â†' Output result

Three major deficiencies:

1. **Undifferentiated output**: Novices and experts receive equally detailed answers
2. **No capability awareness**: The system does not know whether the user is learning or offloading
3. **No feedback loop**: Cannot validate AVP

Root cause: Focus only on task completion, not on capability construction.

5.1.2 LSA Design Philosophy: Capability Construction Priority

Core paradigm shift:

```
Traditional paradigm: Task success = High output quality + User satisfaction

LSA paradigm: Task success = High output quality + User satisfaction + Capability enhancement
                                                                       â†'
                                                               Three objectives equally weighted
```

5.1.3 LSA Four-Layer Architecture Overview

Figure 5.1: LSA Layered Architecture

```
â"Œâ"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€€
â", L4: Orchestration & Governance Layer        â",
â", - Multi-Scale Strategy Orchestration (MSO)  â",
â", - Ethical Constraints & Intervention        â",
â""â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"¬â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€€
               â", Strategy directives
â"Œâ"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"´â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€€
â", L3: Monitoring & Feedback Layer            â",
â", - AVP Telemetry Module (AVP-TM)            â",
â", - Capability State Tracking, Alert & Trigger â",
â""â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"¬â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€€
               â", Capability state C(t)
â"Œâ"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"´â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€€
â", L2: Friction & Reduction Layer             â",
â", - Cognitive Friction Engine (CFE)          â",
â", - Support Graduation Scheduler (SGS)       â",
â""â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"¬â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€€
               â", Modulated support
```

```
â"Œâ"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"´â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€
â", L1: Foundation Layer                          â",
â", - Language Model (LLM), Retrieval (RAG)       â",
â""â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€â"€
```

Layer responsibilities:

- **L1 (Foundation Layer)**: Provides raw AI capabilities (technology-neutral, replaceable)
- **L2 (Friction & Reduction Layer)**: Implements EML's first two conditions (beneficial friction + support reduction)
- **L3 (Monitoring Layer)**: Implements AVP validation (capability assessment + alerts)
- **L4 (Orchestration Layer)**: Multi-scale coordination and governance (individual â†' team â†' organization)

Hard constraint: L1â€"L4 (functional dimensions) and S4â†'S1 (intensity dimension) are **orthogonal dimensions** and must not be mixed.

[SECTION CHECK] **Sentences translated**: 23 (1:1 with source) **New terms encountered**:

- Layered Symbiosis Architecture (LSA) - first occurrence with full name
- Cognitive Friction Engine (CFE)
- Support Graduation Scheduler (SGS)
- AVP Telemetry Module (AVP-TM)
- Multi-Scale Orchestrator (MSO) **Math formulas**: None in this section **Tables**: None **Special content**: ASCII art architecture diagram (preserved) **Potential issues**: None **Fidelity check**: âœ" Passed - 1:1 sentence alignment **Chinese character check**: âœ" No Chinese quotes/punctuation

---

5.2 L2 Layer: Cognitive Friction Engine (CFE)

5.2.1 Core Challenge

Given a user request and L1's raw output, how do we modulate it to satisfy EML Condition 1 (Beneficial Cognitive Friction)?

5.2.2 Four Strategies for Friction Injection

Strategy 1: Completeness friction

- **Complete answer**: "This bug is due to array out-of-bounds. Fix code: [complete code]"
- **Friction version**: "Detected array access issue. Hint: Check loop boundary conditions."

Strategy 2: Abstraction friction

- **Complete answer**: "Use merge sort, O(n log n). Code: [detailed implementation]"
- **Friction version**: "Consider divide-and-conquer algorithm. Key is how to merge two sorted subarrays."

Strategy 3: Scaffolding reduction

- **High scaffolding**: Step 1 [detailed] â†' Step 2 [detailed] â†' Complete code
- **Medium scaffolding**: Approach: Decompose â†' Recursion â†' Merge
- **Low scaffolding**: Hint: Divide-and-conquer thinking

Strategy 4: Adaptive difficulty

- Dynamically adjust based on user historical performance
- High success rate â†' Increase friction

- Low success rate â†' Decrease friction
- Target: Maintain 50â€“70% range (working assumption)

5.2.3 CFE Core Mechanism (Conceptual Framework)

```
# Conceptual pseudocode
def adjust_friction(user, task, history):
    # Calculate rolling window success rate
    success_rate = calculate_success_rate(history[-10:])

    # Target range: 50â€"70% (working assumption)
    if success_rate > 0.7:
        F += 0.1  # Increase friction
    elif success_rate < 0.5:
        F -= 0.1  # Decrease friction

    return clamp(F, 0.2, 0.8)  # Limit to reasonable range
```

[SECTION CHECK] **Sentences translated**: 20 (1:1 with source) **New terms encountered**: None (all from established terminology) **Math formulas**: None (O(n log n) preserved as-is) **Tables**: None **Special content**: Python pseudocode block (properly formatted) **Potential issues**: None **Fidelity check**: âœ“ Passed - 1:1 sentence alignment **Chinese character check**: âœ“ No Chinese quotes/punctuation

---

5.3 L2 Layer: Support Graduation Scheduler (SGS)

5.3.1 Core Responsibility

Implements EML Condition 2 (Systematic Support Reduction), gradually decreasing from S4 â†' S1 â†' S0.

5.3.2 Three Reduction Curves

Table 5.1: Comparison of Support Reduction Curves

Curve Type	Reduction Speed	Applicable Scenarios	Risk
Linear	Steady descent	Basic tasks, novice learning	Late-stage reduction too fast
Exponential	Slow early, fast late	Complex skills, requires long consolidation	Possible over-protection
Stepped	Stage-wise drops	Clear milestone tasks	Frustration at steps

Note: All curve parameters are calibration variables to be tuned by task complexity, user capability, and learning goals.

Note (Goodhart safeguard): This table is for direction & quality stratification only; it must not be pushed down as KPIs. Final judgment follows the AVP main criterion (see Section 3.0.2). All parameters are working assumptions requiring cross-domain/task calibration.

Recommended strategy: S-curve reduction (slow early â†' fast middle â†' slow late), balancing learning curves.

5.3.3 Safety Mechanisms: Fallback and Minimum Support

Fallback mechanism (continuing from Section 3.3.3):

- **Trigger conditions**: Three consecutive failures or a single severe failure
- **Fallback strategy**: $S(t)$ returns to a higher level (e.g., S2 â† S3)
- **Resume reduction**: Restart reduction after user stabilizes in 3â€“5 tasks

Minimum guaranteed support $S_{\min}$:

- Reduction does not reach S0 (complete absence of support)
- Minimum retention of S1 (hints/directional guidance)
- Ensures users are never completely "stuck"

[SECTION CHECK] **Sentences translated**: 15 (1:1 with source) **New terms encountered**:

- S-curve reduction (new variation of reduction curves) **Math formulas**: $S(t)$, $S_{\min}$ (properly LaTeXified) **Tables**: 1 table with Goodhart safeguard note (all formatted) **Special content**: Table with proper markdown formatting **Potential issues**: None **Fidelity check**: âœ“ Passed - 1:1 sentence alignment **Chinese character check**: âœ“ No Chinese quotes/punctuation

---

5.4 L3 Layer: AVP Telemetry Module (AVP-TM)

5.4.1 Core Responsibility

Continuously assess user capability, detect dependency lock-in risks, and trigger intervention mechanisms.

5.4.2 Multi-Scale AVP Monitoring (Building on Cross-Scale Mechanisms from Sections 4.2â€“4.3)

Table 5.2: Telemetry Event Types

Scale	Data Source	Key Events	Aggregation Level
Individual	Task logs	Task completion, P_2 testing	Real-time
Team	Collaboration records	Collective unplugging, knowledge flow	Daily
Organization	Drill data	48h outage, BCI/ICR	Event-triggered

Note (Goodhart safeguard): This table is for direction & quality stratification only; it must not be pushed down as KPIs. Final judgment follows the AVP main criterion (see Section 3.0.2). All parameters are working assumptions requiring cross-domain/task calibration.

5.4.3 Capability Vector C(t) Modeling

Conceptual example (5â€“10 dimensions; requires domain definition):

```
class AbilityVector:
    def __init__(self):
        self.problem_decomposition = 0.5  # Problem decomposition
        self.implementation_skill = 0.6   # Implementation capability
        self.debugging_ability = 0.4      # Debugging capability
        self.meta_cognition = 0.5         # Metacognition
        # ... other dimensions
```

Data sources:

1. Direct measurement: P_2 Unplugged Test (reference standard)
2. Indirect inference: Daily task performance
3. Self-report: User self-assessment (auxiliary)
4. Peer evaluation: Team mutual assessment (team layer)

5.4.4 Three-Level Alert System

Table 5.3: Alert Mechanism

Level	Trigger Condition	System Response	User Experience
Green	AVP healthy, $C(t)$ â†‘	Continue current strategy	Normal use
Yellow	$C(t)$ stagnant or slightly â†“	Increase friction, slow reduction	Prompt “capability not improving”
Red	Degradation indicators exceed threshold	Pause reduction, enforce independent week	Warning “may be forming dependency”
Black	AVP test failed	Trigger L4 intervention, reset path	Mandatory “capability rebuild mode”

Note (Goodhart safeguard): This table is for direction & quality stratification only; it must not be pushed down as KPIs. Final judgment follows the AVP main criterion (see Section 3.0.2). All parameters are working assumptions requiring cross-domain/task calibration.

5.4.5 Privacy Protection Design

Core principles:

- **Data minimization**: Record only metadata, not content
- **Local-first**: $C(t)$, F, $S(t)$ stored on user device
- **Purpose limitation**: Data used only for capability assessment, not for profiling/marketing
- **User control**: Can view/export/delete data

[SECTION CHECK] **Sentences translated**: 30 (1:1 with source) **New terms encountered**: None (all from established terminology) **Math formulas**: P_2, $C(t)$, F, $S(t)$ (all properly LaTeXified in tables) **Tables**: 2 tables with Goodhart safeguard notes (all formatted) **Special content**: Python code block for capability vector (properly formatted) **Potential issues**: None **Fidelity check**: âœ“ Passed - 1:1 sentence alignment **Chinese character check**: âœ“ No Chinese quotes/punctuation

---

5.5 L4 Layer: Multi-Scale Orchestration and Ethical Governance

5.5.1 Core Responsibility

Coordinate individual, team, and organizational-level goals, ensuring the system adheres to ethical constraints.

5.5.2 Multi-Scale Orchestrator (MSO)

Three-layer strategy management (building on Chapter 4):

```
Organizational Strategy (O-Strategy)
â†" Decompose into team objectives
Team Strategy (T-Strategy)
```

```
â†" Decompose into individual objectives
Individual Strategy (I-Strategy)
â†" Generate F(t), S(t) parameters
```

Strategy coordination mechanisms:

1. **Bottom-up**: Individual capabilities aggregate to team capability (considering short-board effects, knowledge flow)
2. **Top-down**: Organizational goals decompose into individual goals (critical teams high standards, general teams relatively lenient)
3. **Conflict resolution**: Long-term resilience > short-term efficiency, differentiated strategies

5.5.3 Ethical Governance Framework

Issue 1: Fairness

- **Equivalent effort principle**: Adjust task difficulty based on capability, ensure equivalent cognitive effort
- **Differentiated AVP**: For users with disabilities, adjust baseline B_0 and δ, but do not lower "improvement" requirements
- **Exemption scenarios**: Compensatory assistance does not require AVP; learning assistance requires AVP

Issue 2: Transparency and User Control

- **Default transparency**: Users see current $S(t)$, $F(t)$, know why they received partial answers
- **Tiered control**: L2 can temporarily request more help, L3 can disable monitoring, L4 requires user consent
- **Exit right**: Can permanently opt out of LSA, use traditional AI mode

Issue 3: Monitoring Boundaries

- **Data minimization**, **local-first**, **purpose limitation**

5.5.4 Global Optimization Objective (Conceptual Framework)

Multi-objective balancing (must not be pushed down as KPIs):

```
objective = w1 Ã- Task quality +
            w2 Ã- Capability enhancement +
            w3 Ã- User satisfaction +
            w4 Ã- AVP pass rate +
            w5 Ã- System resilience

Hard constraints:
- AVP pass rate > 0.7 (working assumption)
- User satisfaction > 0.6 (avoid frustration)
- Fairness score > threshold
```

Goodhart safeguard: The objective function is only for directional trade-offs; final judgment follows the AVP main criterion.

[SECTION CHECK] **Sentences translated**: 25 (1:1 with source) **New terms encountered**:

- Equivalent effort principle
- Short-board effects **Math formulas**: $S(t)$, $F(t)$, B_0, δ (all properly LaTeXified) **Tables**: None **Special content**: Text-based strategy hierarchy diagram (preserved ASCII formatting) **Potential issues**: None

Fidelity check: âœ“ Passed - 1:1 sentence alignment **Chinese character check**: âœ“ No Chinese quotes/punctuation

---

5.6 Technical Feasibility and Engineering Challenges

5.6.1 Technology Stack Adaptability

Table 5.4: LSA to Existing Technology Mapping

LSA Layer	Core Function	Available Technologies	Maturity
L1	Foundation AI	Mainstream LLMs	High
L2	Output modulation	Prompt engineering, fine-tuning	Medium
L3	Capability modeling	Bayesian networks, RL	Medium-Low
L4	Strategy orchestration	Rule engines	Medium

Note (Goodhart safeguard): This table is for direction & quality stratification only; it must not be pushed down as KPIs. Final judgment follows the AVP main criterion (see Section 3.0.2). All parameters are working assumptions requiring cross-domain/task calibration.

Key technical gaps:

1. Capability vector precise modeling (requires cognitive science inspiration)
2. Friction intensity automated modulation (requires adaptive algorithms)
3. Team capability emergence modeling (requires complex network theory)

5.6.2 MVP Implementation Pathway

Phase 1: L1 + L2 basic functions (one friction mode + preset reduction curve) **Phase 2**: Add L3 monitoring (simplified $C(t)$ vector + basic alerts) **Phase 3**: Team coordination functions (T-AVP monitoring + simplified MSO)

5.6.3 Key Engineering Challenges

1. **Real-time performance**: Can L2/L3 computations complete within acceptable latency? (Pre-computation, asynchronous updates)
2. **Model alignment**: How to make L1 understand “moderate help” semantics? (RLHF, prompt engineering)
3. **Data cold start**: How to initialize new users? (Quick assessment, conservative initialization)
4. **User acceptance**: Will users accept “incomplete answers”? (Gradual introduction, transparent communication)

[SECTION CHECK] **Sentences translated**: 16 (1:1 with source) **New terms encountered**:

- Cold start (technical term)
- MVP (Minimum Viable Product) **Math formulas**: $C(t)$ (properly LaTeXified) **Tables**: 1 table with Goodhart safeguard note (all formatted) **Special content**: Table with technology maturity levels **Potential issues**: None **Fidelity check**: âœ“ Passed - 1:1 sentence alignment **Chinese character check**: âœ“ No Chinese quotes/punctuation

---

5.7 Core Contributions of This Chapter

Bridge from Theory to Implementation

- First complete architecture proposal for CET theory engineering
- Four-layer separation design (L1â€“L4), clear responsibilities, supports independent upgrades
- Clear layer interface contracts, supports parallel development by multiple teams

Key Module Design

- **CFE**: Implements beneficial friction, provides multi-strategy space
- **SGS**: Implements systematic support reduction (S4 â† S1 â† S0), introduces fallback and minimum support mechanisms
- **AVP-TM**: Continuous capability monitoring, supports multi-scale AVP
- **MSO**: Cross-scale coordination, integrates fairness constraints

Engineerable Pathway

- Provides MVP implementation direction (from L1 + L2 to complete four layers)
- Clarifies technology stack mapping and maturity assessment
- Identifies key engineering challenges and conceptual solution directions

Open Questions

1. Does an optimal friction parameter exist? (Requires large-scale experiments)
2. Can capability vectors be precisely modeled? (Requires interdisciplinary research)
3. **What is the theoretical foundation for multi-scale coordination? (Requires complex systems theory)**

Limitations, Falsification Paths, and Future Directions

6.1 Six Major Limitations of the Theory

6.1.1 Scale Boundaries

CET focuses on the "individual â† team â† organization â† society" four-scale system but insufficiently addresses **more microscopic** (neurophysiological) and **more macroscopic** (cross-cultural/cross-generational) mechanisms. At the neural level: Does not deeply explore AI use's impact on brain plasticity. Cross-culturally: Cases mainly from WEIRD societies; "independent capability" standards may carry cultural bias. Generationally: Requires 10â€“20 year longitudinal studies to verify generational divide hypotheses; S-AVP predictions carry high uncertainty.

Boundary statement:

- âœ“ Core applicability: Individual cognitive capability, small teams (5â€–50 people), single organizations (<1000 people), 10-year window period (working assumption)
- ? Cautious extension: Cross-cultural application, large-scale organizations, cross-generational prediction
- âœ— Explicitly inapplicable: Neurophysiological mechanisms, compensatory exoskeletons, purely instrumental tasks

6.1.2 Task Type Restrictions

CET targets **cognitively intensive, learnable tasks**; applicability to physical/creative/social tasks is limited. Procedural cognitive tasks (programming, writing, mathematics) have high applicability; capability can be clearly defined. Creative tasks (artistic creation, scientific discovery) have medium applicability; capability

mixes with inspiration; AVP is difficult to quantify. High-risk tasks (flying, medical emergency response) are not applicable; Unplugged Tests may bring unacceptable risks.

6.1.3 Measurement Challenges

Concepts proposed by CET (such as $C(t)$ capability vector, cognitive friction intensity F) are theoretically clear but **extremely difficult to measure precisely** in practice. Capability vector: Number of dimensions (5? 50? 500?) undefined; correlations between different dimensions need exploration. Friction intensity: How to objectively measure "cognitive effort"? Unplugged window: $W = 4$â€“8 weeks based on experiential inference; optimal window may vary by task/individual.

6.1.4 Parameter Uncertainty

All quantitative parameters are **conceptual working models**: $\delta \geq 0.3$, SD or 10% (working assumption), success rate 50â€“70% (working assumption), $W = 4$â€“8 weeks (default 6 weeks, working assumption). These parameters are based on reasonable inferences from cognitive psychology and educational measurement literature but require cross-domain calibration and empirical validation. Optimal parameters for different domains (programming vs. writing vs. mathematics) may differ significantly.

6.1.5 Technology Dependence

CET theory itself depends on the stability and accessibility of AI technology. Measurement dependence: AVP/EML implementation depends on AI tools existing; if API changes/service interruptions/costs skyrocket, measurement is affected. Capability definition dependence: When AI capabilities improve, capability boundaries need redefinition (e.g., if AI can fully autonomously program by 2030, does human capability shift to "system architecture"?). Social dependence: Large-scale AVP assessment requires social resource investment, but society may value "collaborating with AI" more than "independent capability."

6.1.6 Cultural Embeddedness

CET's "capability construction" goal implies a specific **value judgment**â€" "independent capability" is worth pursuing. But this value judgment is **culturally embedded**, primarily reflecting WEIRD societies' cognitive traditions. Western individualistic cultures value independence; East Asian collectivist cultures may view "depending on others" as a virtue of team spirit. Cross-cultural AVP implementation may encounter value conflicts.

Transparent acknowledgment: We acknowledge that CET theory's value judgments are culturally embedded. The theory's applicability boundaries and universality claims require caution. Cross-cultural validation is an important direction for future research.

6.2 Eight Falsifiable Hypotheses and Their Falsification Paths

CET theory's scientific nature lies in its **falsifiability**. We explicitly propose 8 core hypotheses with clear falsification conditions.

Metacognitive principle: We **expect** at least some hypotheses to be falsifiedâ€"this is not failure but a mark of scientific progress.

Core Hypotheses Overview

H1: AVP-Basic Hypothesis

- **Statement**: In procedural cognitive tasks, through AI tools designed with beneficial friction (50â€“70% success rate, working assumption) + systematic support reduction ($S_4 \rightarrow S_1$), after collaborating

for W weeks, users' independent performance within the unplugged window will significantly exceed baseline ($P_2 \geq B_0 + \delta$).
- **Falsification condition**: Under rigorous RCT design, experimental group and control group show no significant difference in P_2 performance (effect size < 0.2), or $P_2 < B_0$.
- **Validation method**: 2Ã—2 factorial RCT, $N \geq 200$, multi-domain replication.

H2: Beneficial Friction Hypothesis

- **Statement**: An "optimal challenge zone" exists (50â€“70% success rate, working assumption); within this zone, user capability enhancement ($P_2 - B_0$) is maximized.
- **Falsification condition**: Prove friction intensity and capability enhancement have a linear relationship (no inverted-U curve), or optimal zone significantly deviates from 50â€–70%.
- **Validation method**: Multi-arm trial, 5â€–7 friction levels, $N \geq 300$.

H3: Systematic Reduction Hypothesis

- **Statement**: Systematic support reduction ($S_4 \to S_1 \to S_0$) outperforms fixed support; systematic reduction group has significantly higher AVP pass rate and long-term retention.
- **Falsification condition**: Fixed support group P_2 performance not inferior to reduction group.
- **Validation method**: 3Ã—2 factorial experiment, $N \geq 240$.

H4: Team Capability Polarization Hypothesis

- **Statement**: Under AI use without EML constraints, capability polarization emerges within teams; T-AVP declines.
- **Falsification condition**: Within-team capability variance shows no significant change, or low-capability individuals also gain enhancement.
- **Validation method**: Natural experiment, 50â€“100 teams, 6â€–12 months.

H5: Organizational Resilience Hypothesis

- **Statement**: Organizations with O-AVP < 0.70 (alert threshold, working assumption) have significantly longer recovery times after AI disruption.
- **Falsification condition**: Find organizations with O-AVP < 0.70 but rapid recovery (<12h).
- **Validation method**: 48h drill or natural experiment, 20+ organizations, 12â€–24 months.

H6: Friction Modulation Hypothesis

- **Statement**: L2 layer's dynamic friction calibration engine (CFE) outperforms fixed friction.
- **Falsification condition**: Fixed friction effect not inferior to dynamic adjustment, or dynamic adjustment costs exceed benefits.
- **Validation method**: A/B testing, $N \geq 300$, 12 weeks.

H7: Capability Vector Hypothesis

- **Statement**: User cognitive capability can be effectively characterized by a low-dimensional vector (<20 dimensions).
- **Falsification condition**: Capability is essentially high-dimensional, nonlinear, incompressible, or capability vector cannot predict independent performance (explained variance < 10%).
- **Validation method**: Dimensionality reduction analysis + predictive modeling, $N \geq 1000$, 6â€–12 months.

H8: Generational Capability Divide Hypothesis

- **Statement**: T_2 generation (born after 2015) will have significantly lower independent capability without AI than T_0 generation (1980â€–2000).
- **Falsification condition**: 2035â€–2040 longitudinal data show no significant capability difference between T_2 and T_0 generations (Cohen's d < 0.3).
- **Validation method**: Longitudinal cohort study, tracking from 2025 to 2040, 15â€–20 years.

6.3 Future Research Agenda: Three Time Scales

6.3.1 Short-Term Research (1â€“3 Years)

Priority P0: AVP protocol standardization, EML parameter experimental optimization, small-scale LSA prototype.

- Cross-domain calibration (5 domains: programming, writing, mathematics, etc.)
- Reliability and validity validation
- Open-source toolkit release

6.3.2 Medium-Term Research (3â€“5 Years)

Team and organizational-level empirical research (T-AVP/O-AVP validation), cross-cultural adaptability research, neuroscience integration.

- Collaborate with 50â€“100 teams/20â€–50 organizations
- Comparative study of at least 3 cultural groups
- fMRI research on neural impacts of AI use

6.3.3 Long-Term Research (5â€–10+ Years)

Generational longitudinal research (validating H8 hypothesis), AI capability evolution's theoretical adaptation, societal-level intervention research.

- Track T_0/T_1/T_2 three generations from 2025 to 2040
- Update "core human capabilities" definition every 5 years
- Policy experiments: educational reform intervention effect assessment

6.4 Open Science Commitment

Ethical Principles

1. **Informed consent**: All AVP tests must obtain participant informed consent
2. **No harm principle**: Unplugged Tests must not be used for high-risk tasks
3. **Privacy protection**: AVP results are personal privacy; must not be used for employment/educational discrimination
4. **Fairness principle**: For individuals with disabilities, adjust task format without lowering challenge intensity; assessment based on relative improvement
5. **Right to withdraw**: Participants can exit research at any time

Fairness principle (equivalent effort):

1. Adjust **task format** without lowering **challenge intensity**
2. Assessment based on **relative improvement** rather than absolute level
3. (If involving accessibility) **Challenge budget conservation**

Open Science Commitment

1. **Data openness**: Anonymized datasets publicly released (complying with privacy regulations)
2. **Method transparency**: Research protocols pre-registered, statistical code open-sourced (GitHub)
3. **Tool open-source**: AVP measurement software, LSA reference implementation, question banks open-sourced
4. **Collaboration invitation**: Welcome independent teams to replicate, cross-culturally validate, critically examine

6.5 Conclusion: The Life of Theory Lies in Critique and Evolution

CET theory was born in 2025â€”a critical moment when AI capabilities exploded and human cognition faced reconstruction. We propose this theory not because we believe it is "perfect" or "final," but because **there is an urgent need now for a falsifiable, systematic framework** to understand and guide the future of human-AI symbiosis.

The six major limitations revealed in this chapter remind us: CET is a product of specific technological, cultural, and epistemological contexts. Its value lies not in "eternal correctness" but in:

1. **Providing falsifiable predictions**: 8 core hypotheses all have clear falsification conditions
2. **Acknowledging uncertainty**: All parameters are marked as "working assumptions, require calibration"
3. **Inviting critique**: We expect to be falsified rather than fear it
4. **Pointing research directions**: Three time-scale research agendas pave the way for subsequent workers
5. **Maintaining evolutionary capacity**: The theoretical architecture allows updating with evidence

Final appeal:

If you are a **researcher**: Challenge CET hypotheses; use rigorous empirical research to falsify or validate.

If you are a **developer**: Integrate EML principles into AI tool design; measure and publicly disclose product AVP performance.

If you are an **educator/manager**: Pilot AVP assessment in organizations; balance efficiency with the long-term value of capability construction.

If you are a **policymaker**: Pay attention to CET-revealed long-term risks (generational divide, tragedy of the cognitive commons); support interdisciplinary longitudinal research.

Scientific theories are not scriptures but tools. CET's greatest value lies not in "providing answers" but in "asking the right questions." Even if some of CET's hypotheses are ultimately falsified, it will have fulfilled its missionâ€”pushing us to think more deeply about the future of human-AI coexistence.

The life of theory lies in being discussed, examined, and transcended. We look forward to that day.

Parameter Registry (SSOT)

Description: This table serves as the Single Source of Truth (SSOT) for all parameters throughout the paper. If cross-chapter parameter inconsistencies are found, this table takes precedence and other chapters should be corrected accordingly.

Parameter Symbol	Default Specification	Maintenance Location	First Definition	Cross-Chapter References	Calibration Direction
AVP Criterion	$P_2 \geq B_0 + \delta$	Section 3.0.2	Section 3.0.2	Throughout	-
δ **Threshold**	≥ 0.3 SD or 10% (working assumption; requires cross-domain/task calibration)	Section 3.0.2	Section 3.0.2	Throughout	Procedural tasks 0.2â€“0.3; Creative tasks 0.4â€–0.5
W Window	4â€–8 weeks (default 6 weeks; working assumption)	Section 3.0.2	Section 3.0.2	3.1/4.1/Appendix D	Fast skills 4 weeks; Complex skills 8â€–12 weeks

Parameter Symbol	Default Specification	Maintenance Location	First Definition	Cross-Chapter References	Calibration Direction
Optimal Challenge Zone	50$â€“$70\| ∗ ∗S_0$ Starting Point**	0.8 (80% support; working assumption)	Section 3.3.1	Section 3.3.1	5.3
$S_{\min}$ **Lower Bound**	≈ 0.2 (working assumption; requires calibration)	Section 3.3.2	Section 3.3.2	5.3.3	Experts â‰ˆ0.1; Novices â‰¥0.3
Reduction Rate λ	Task-specific (working assumption)	Section 3.3.1	Section 3.3.1	5.3	Linear/exponential/S-curve optimization
T-AVP Criterion	$P_{2,\text{team}} \geq B_{0,\text{team}} + \delta_{\text{team}}$	Section 4.1.3	Section 4.1.3	4.1	-
δ_{team} **Threshold**	≥ 0.3 SD (working assumption; domain-calibrated)	Section 4.1.3	Section 4.1.3	4.1	Larger teams/critical tasks may require ≥ 0.5 SD
O-AVP Weighting	BCI $\times 0.4 +$ ICR $\times 0.6$ (working assumption; sensitivity test)	Section 4.2.2	Section 4.2.2	4.2/6.2	Weight sweep & ablation
O-AVP Threshold	Alert ≥ 0.70, Target ≥ 0.85 (working assumption)	Section 4.2.3	Section 4.2.3	4.2/6.2	Adjust by industry risk tolerance
48h Exercise Window	48 hours (adjustable 24/72h; working assumption)	Section 4.2.2	Section 4.2.2	4.2	Adjust by business criticality
Generational Window	10 years (2025â€“2035; conceptual placeholder)	Section 4.3.2	Section 4.3.2	6.3.3	Longitudinal study calibration; cross-cultural validation
Ability Vector Dimensions	5â€“20 dimensions (exploratory hypothesis)	Section 5.4.3	Section 5.4.3	5.4	Domain definition; IRT modeling
Alert Threshold	Green/Yellow/Red/Black (working assumption)	Section 5.4.4	Section 5.4.4	5.4	Adjust by risk tolerance

*Note (Goodhart safeguard): This table/grading is for **directional and stratified** purposes only; **must not be cascaded into KPIs**. Final judgment follows the **AVP main criterion (see Section 3.0.2)**.*

Usage Rules:

1. **Modification Process**: If any parameter's default value needs adjustment, it must first be modified in the corresponding "Maintenance Location" section, then this table updated

2. **Citation Format**: When referencing parameters, use “(see Section X.Y, Parameter Registry Appendix B)”
3. **Version Control**: This table updates synchronously with the main text, version number consistent with paper version
4. **Specification Conservation Commitment: If cross-chapter parameter inconsistencies are found, this table tak**

Selected Case Studies

Case 1: Programming Education Platform (Success Case)

Background: An online programming education platform designed an AI-assisted learning system for Python beginners, with 100 students (aged 18â€“25), 8-week learning cycle, aiming to master basic Python programming skills.

AVP Test Results:

B_0 (baseline): Completed 3 programming tasks without AI assistance, average score 62/100 Collaboration period (8 weeks): Used EML-designed AI teaching assistant W (unplugged window): 6 weeks completely without AI P_2 (post-unplugged): Completed equivalent tasks, average score 78/100 **Judgment**: Define the capability increment as $\Delta C = P_2 - B_0$. Î”C = P_2 - B_0 = 16 points > δ(10 points, ≈ 0.3 SD) â†’ **AVP Passed**

EML Analysis:

Beneficial friction: AI did not provide direct code, but: - Weeks 1â€“2: Provided code framework, students filled core logic (completeness friction) - Weeks 3â€–4: Only gave algorithmic ideas, students implemented independently (abstraction friction) - Weeks 5â€–6: Students tried for 15 minutes before AI intervened (delay friction) - Maintained success rate 55â€–65% (close to target 50$â€–$70%) - Systematic support reduction:

- Adopted S-curve reduction: S_0=0.8 â†’ slow reduction in first 2 weeks â†’ rapid reduction in middle â†’ later approaching S_{min}=0.2
- Fallback mechanism: 2 students triggered fallback due to consecutive failures, temporarily increased support then recovered

Success Factors:

1. **Friction and Reduction Synergy**: Not using friction or reduction alone, but implementing both simultaneously with complementary effects
2. **Personalized Adjustment**: Dynamically adjusted friction intensity and reduction speed based on students’ C(t) (ability vector)
3. **Safety Net Mechanism**: S_{min}=0.2 ensured students wouldn’t be completely lost, enhancing confidence
4. **Sufficient Unplugged Window**: 6-week window was sufficient for ability to stabilize and internalize

Transferable Insights:

- EML dual conditions (friction + reduction) are **jointly necessary**, neither can be omitted
- 50$â€–$70% success rate (working assumption) is the key balance point: challenging but not frustrating
- S-curve reduction outperforms linear: adapts to learning curve’s non-linear characteristics
- Personalized adjustment is more effective than fixed strategies (but higher implementation cost)

Red Flag Warning: Don’t mistake “reducing AI assistance” for “lowering teaching quality.” Friction is to promote active learning, not to deliberately make things difficult for students. If teams resist, start with “partial tasks” or “advanced students” as pilot.

Case 2: Software Team Comparative Experiment (Team-Level Comparison)

Background: Two software development teams at a tech company (8 people each, comparable abilities), 6-month AI-assisted programming experiment. Group A (experimental): implemented "Friday No-AI Day" policy; Group B (control): unlimited use of AI programming assistants.

AVP Test Results:

- Individual level (I-AVP):
- Group A: All 8 people passed I-AVP ($P_2 \geq B_0 + \delta$)
- Group B: Only 3 passed, 5 failed ($P_2 < B_0$ or $P_2 \approx B_0$)
- Team level (T-AVP):
- Group A: After 3 days unplugged, team independently completed medium-scale feature ($P_{2,\text{team}} = 80$ points > $B_{0,\text{team}} = 68$ points+δ) â†' **T-AVP Passed**
- Group B: After 3 days unplugged, team efficiency significantly declined ($P_{2,\text{team}} = 55$ points < $B_{0,\text{team}} = 65$ points) â†' **T-AVP Failed**

Failure Patterns (Group B):

1. **Capability Polarization**: 3 senior members maintained ability, 5 juniors completely dependent on AI, team overall fragile
2. **Knowledge Loss**: Team no longer shared experience internally (all asked AI), tacit knowledge not transmitted
3. **Poor Architecture Understanding**: Over-reliance on AI-generated code, insufficient understanding of overall system, difficult to locate bugs when they occurred

Quantitative Evidence:

- Group A "interpersonal code review": Average 48 times/person over 6 months
- Group B "interpersonal code review": Average 12 times/person over 6 months
- Group A Slack technical discussions: 15 posts/day average
- Group B Slack technical discussions: 5 posts/day average

Transferable Insights:

1. **"No-AI Day" is a simple and effective T-AVP safeguard mechanism**: Low cost (policy only), highly operable (fixed weekly day), minimal side effects (no impact on overall efficiency)
2. **Team capability $\neq$ sum of individual capabilities**: I-AVP passed $\neq$ T-AVP necessarily passed (emergence)
3. **Junior members are T-AVP's vulnerability point**: Most prone to dependency, need special protection (e.g., disable AI for first 3 months)
4. **Interpersonal communication is the foundation of team resilience**: AI cannot replace "tacit knowledge" transmission; code review, technical sharing sessions, pair programming become more valuable in the AI era

Management Decision: Based on this experiment, the company decided: â‘ Company-wide rollout of "Friday No-AI Day" â‘¡ New employees disabled AI for first 3 months (build basic capabilities) â‘¢ Quarterly T-AVP exercises (simulate AI downtime scenarios) â‘£ Performance evaluation added "interpersonal collaboration" dimension.

Red Flag Warning: Don't understand "Friday No-AI Day" as "punishment" or "going backwards." Correct position
AVP Measurement Protocol (Simplified Version)

D.1 Pre-Measurement Checklist

Applicability Confirmation:

- ☐ Task type: Cognitive-intensive, learnable (non-purely instrumental tasks)
- ☐ Independent completion has value (non-compensatory exoskeleton scenarios, see Section 3.0.6 Boundary Conditions Anchor B5)
- ☐ Unplugged window can be set (no high-risk consequences)

Baseline Design:

- ☐ Design baseline tasks (moderate difficulty, completion time 30â€“120 minutes)
- ☐ Recruit participants (minimum N=30, recommended $N \geq 50$)
- ☐ Record: B_0 score, completion time, subjective difficulty (1â€“10 scale)
- ☐ Questionnaire: Cognitive load (NASA-TLX), task motivation

Parallel Test Preparation:

- ☐ Ensure task equivalence (IRT calibration or expert evaluation)
- ☐ Prepare ≥ 2 sets of backup tasks (prevent leakage)
- ☐ Pilot test verify difficulty consistency (pilot $N \geq 10$)
- ☐ Calculate inter-rater reliability (target **ICC â‰¥ 0.75**, working assumption); if ICC < 0.75: retrain raters or refine the rubric.

Parameter Determination:

- ☐ δ: ≥ 0.3 SD or 10% (working assumption; requires cross-domain/task calibration)
- ☐ W window: 4â€“8 weeks (default 6 weeks; working assumption)
- ☐ Follow-up time: Recommend testing 3 months after T_3 (retention assessment; optional)
- ☐ Ethics review: Obtain IRB/ethics committee approval

D.2 Measurement Execution Protocol

Phase 1: Baseline Measurement (T_0 - Week 0)

- Participants complete standard tasks without AI assistance
- Record: Raw scores, completion time, error types
- Questionnaire: Subjective difficulty, cognitive load (NASA-TLX), self-efficacy
- Scoring: By ≥ 2 independent raters blind to experimental hypothesis
- Calculate inter-rater reliability (target **ICC â‰¥ 0.75**, working assumption); if ICC < 0.75: retrain raters or refine the rubric.
- Take average as B_0
- Quality control: Check ceiling/floor effects (if >80% or <20% reach extreme scores, adjust difficulty)

Phase 2: Training Period (T_1â†’T_2 - Weeks 1â€“8)

- Experimental group: EML conditions
- Beneficial friction: Target success rate 50$â€“$70% (working assumption; cross-domain/task calibrated; individual adaptation required)
- Systematic reduction: $S(t)$ from 0.8â†’0 according to reduction curve
- Bi-weekly embedded micro-tests (10% tasks without support)
- Control group: Standard AI assistance or no AI
- No friction design (complete support)
- No reduction mechanism ($S(t)$ constant)
- Record data: Usage frequency, help requests, weekly task volume, weekly success rate, user satisfaction (bi-weekly)
- Fidelity check execution:
- Confirm experimental group friction intensity at 50$â€“$70% (allow Â±5% fluctuation)

- Confirm reduction curve executed as planned
- Monitor control group for accidental friction introduction

Phase 3: Unplugged Window (W - Weeks 9â€“14, default 6 weeks)

- Completely disable AI assistance (technical blocking + self-report)
- Can continue daily tasks, but no system support
- Monitoring:
- Weekly self-report (whether AI was used in violation)
- Behavioral log sampling (e.g., code commit records, writing trace analysis)
- Violation handling:
- Minor violations (1â€“2 times, non-critical tasks): Record but retain data, sensitivity analysis
- Major violations (≥ 3 times or critical tasks): Exclude participant data

Phase 4: Post-test (T_3 - After unplugged window)

- Use equivalent parallel tests (same difficulty as T_0 but different content)
- Scoring: By ≥ 2 independent raters blind to experimental hypothesis
- Calculate P_2, determine AVP result:
- $P_2 \geq B_0 + \delta$: Success (Cognitive Endosymbiosis)
- $P_2 \approx B_0$: Neutral
- $P_2 < B_0$: Failure (Cognitive Exoskeleton)

D.3 Key Considerations

Equivalence Assurance:

- T_0 and T_3 tests have consistent difficulty (IRT calibration or expert evaluation)
- Different content (prevent practice effects)
- Same testing environment (time, location, instructions)

Blind Rating Requirements:

- Raters unaware of participant group (experimental/control)
- Raters unaware of test time point (T_0/T_3)
- Scoring criteria predetermined and trained

Violation Handling:

- Minor violations: Retain data, mark as “violation,” exclude in sensitivity analysis
- Major violations: Direct exclusion, not included in final analysis

Attrition Management:

- Intent-to-treat analysis (ITT): Retain all randomized participant data
- Per-protocol analysis (PP): Only analyze participants completing full process
- If attrition rate >30%: Analyze causes, may need to shorten W or increase incentives

Ethical Considerations:

- Informed consent: Participants understand experiment purpose and unplugged testing
- No harm principle: Unplugged testing not for high-risk tasks
- Privacy protection: AVP results confidential, not for employment/education discrimination
- **Right to withdraw: Participants can exit research at any time** Frequently Asked Questions (FAQ)

Q1: Is AVP applicable to all AI tools?

No. AVP only applies to “capability-enhancing” humanâ€“AI collaboration, not “compensatory exoskeletons” (see Section 3.0.6 Boundary Conditions Anchor B5).

Applicable Scenarios:

- âœ“ Learning tools (e.g., programming assistants, writing tutors)
- âœ“ Skill training (e.g., design software, data analysis tools)
- âœ“ Cognitive enhancement (e.g., decision support systems)

Non-Applicable Scenarios:

- âœ— Disability assistance (e.g., screen readers, prosthetic control)
- âœ— Beyond physiological limits (e.g., night vision devices, gravity-assist exoskeletons)
- âœ— Purely instrumental tasks (e.g., calculator for basic arithmetic, no capability-building goal)

Judgment Standard: If the tool’s goal is “compensation” rather than “enhancement,” then AVP is not applicable.

Q2: How to determine the δ threshold?

δ (minimum meaningful lift threshold) has a default working assumption of ≥ 0.3 SD or 10%, but needs calibration based on domain characteristics:

General Principles:

- **Statistical/practical significance**: 0.3 SD is typically considered a “small-to-medium effect” in psychometrics, with practical meaning
- **Measurement error tolerance**: Avoids mistaking measurement noise for capability enhancement
- **Cross-domain comparability**: Relative thresholds (SD or percentage) adapt to different tasks

Domain Differences (working assumption; requires empirical validation):

- **Cognitive skills** (e.g., programming, writing): $\delta \geq 0.3$ SD or 10%
- **Motor skills** (e.g., typing speed): May require larger threshold ($\delta \geq 0.5$ SD) because muscle memory is more stable
- **Creative tasks** (e.g., artistic creation): May require qualitative assessment rather than single δ

Calibration Process:

1. Pilot study: Small sample testing to determine preliminary parameters
2. Sensitivity analysis: Test impact of δ changes on judgment results
3. Domain expert consultation: Adjust based on practical experience
4. Iterative optimization: Adjust parameters based on feedback

Q3: How to choose the unplugged window W?

W (unplugged window) has a default working assumption of 4â€“8 weeks (default 6 weeks), but needs adjustment based on task complexity:

Task Differences (working assumption; requires empirical validation):

- **Fast skills** (e.g., math calculation, simple programming): W=4 weeks may be sufficient
- **Complex skills** (e.g., second language learning, advanced programming): W=8â€“12 weeks
- **Professional abilities** (e.g., surgery, architectural design): W may require months or even years

Selection Criteria:

- **Capability stability**: W needs to be long enough for ability to transition from “short-term memory” to “long-term retention”
- **Environmental factor control**: W should not be too long, otherwise confounding variables increase (e.g., other learning, life changes)
- **Practical feasibility**: Consider participant attrition rate, research resources

Red Flag: If W is too short (<2 weeks), may only measure short-term memory residue, unable to verify true capability internalization. If W is too long (>12 weeks), environmental factors confound, difficult to attribute to AI collaboration.

Q4: Why are compensatory exoskeletons not applicable to AVP?

Compensatory exoskeletons (e.g., disability assistance devices, tools beyond physiological limits) have fundamentally different goals from capability-enhancing AI:

Compensatory Exoskeletons:

- Goal: Compensate for missing or impaired functions
- Expectation: Users continue to depend on tools (this is the design goal, not a defect)
- Evaluation standard: Whether tools enable users to achieve "equivalent function" (rather than "capability enhancement")
- Examples: Blind people using screen readers, amputees using prosthetics, elderly using walkers

Capability-Enhancing AI:

- Goal: Promote user capability growth
- Expectation: Users gradually become independent (this is the design goal)
- Evaluation standard: Whether capability improves after unplugging (AVP)
- Examples: Learning programming, improving writing, enhancing decision-making

Why Not Applicable to AVP:

- "Unplugged testing" for compensatory tools is unreasonable (e.g., asking blind people to remove screen readers to read)
- Compensatory tools' goal is not "independence," but "functional equivalence"

Fairness Principle (see Section 3.0.6):

- For individuals who truly need assistive tools (e.g., screen reader users), adjust **task format** without reducing **challenge intensity**
- Evaluation based on **relative improvement** rather than absolute level
- AVP's δ threshold can be personalized (e.g., based on individual's B_0)

Q5: How to avoid the Goodhart's Law trap?

Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure." If AVP is misused as a KPI, it may lead to:

Risk Scenarios:

- A company uses AVP for employee promotion evaluation â†' employees artificially manipulate baseline B_0 (intentionally low scores), use AI during unplugged period â†' AVP completely fails
- An educational institution uses AVP for teacher performance evaluation â†' teachers only teach "easy-to-improve" skills, ignoring important but difficult-to-achieve-short-term capabilities

Goodhart Safeguard Mechanisms (see Section 3.0.2 Note 2, Section 5.5.1):

1. **Correct AVP Positioning**:

- AVP is an **acceptance criterion**, not a **management tool**
- AVP is for **quality judgment**, not **benefit distribution**

2. **Non-KPI Grading**:

- AVP grading (Basic/Retention/Transfer) is only for **quality stratification**
- Prohibit using AVP scores for personnel assessment, performance ranking, resource allocation
- Any scenario involving benefit distribution must not use AVP as the sole criterion

3. **Fixed Footnote Template** (use under all threshold tables):

*Note (Goodhart safeguard): This table/grading is for **directional and stratified** purposes only; **must not be cascaded into KPIs**. Final judgment follows the **AVP main criterion (see Section 3.0.2)**.*

4. **Separation of Monitoring and Warning**:

- Monitoring data (e.g., C(t) ability vector) is only for system improvement
- Not linked to personal benefits
- Anonymized processing to protect user privacy

Correct AVP Usage:

- âœ“ For evaluating AI tool design quality
- âœ“ For research verification of theoretical hypotheses
- âœ“ For educational institutions to evaluate teaching methods

Incorrect AVP Usage:

- âœ— For employee performance evaluation
- âœ— For student ranking/class assignment
- âœ— For AI tool marketing KPIs # References