AI时代的人类能力建构与认知韧性理论

副标题：基于反脆弱性验证原则的认知内共生范式

作者：杨国平 邮箱：a44425874@gmail.com 版本：V1.0 日期：2025年9月

摘要

大语言模型等AI工具正深刻改变人类工作方式，但其对独立能力的长期影响尚不明确。本研究提出认知能力迁移理论(CET)，
B5)测量框架，提出”等效努力匹配学习”(EML)设计原则，通过有益摩擦和渐进撤出促进能力内化。理论提出8个可证伪假说

关键词：认知内共生；认知外骨骼；反脆弱性验证；有益认知摩擦；AI人机关系；认知可持续性

注：本文遵循单一事实源（SSOT）原则，所有核心参数定义以第7章第6节参数登记簿为准，其他章节仅作引用。

术语与参数说明：本文涉及的所有核心参数（δ阈值、W窗口、成功率范围等）以单一事实源(SSOT)原则管理，统一定义于第
B5锚点定义见第3章第3.0.2-3.0.6节。

第一章 引言与理论定位

1.1 核心命题：评估AI的新标准

当我们评估一个AI工具的价值时，我们真正应该关注什么？

主流的评价标准聚焦于使用时的表现：任务完成速度、输出质量、用户满意度。然而，本文提出一个根本性的反驳：AI工具

我们将这一原则形式化为反脆弱性验证原则（Antifragility Validation Principle, AVP）：

> 以拔线测试（Unplugged Test）检验协作是否促进独立能力。核心判据：$P \geq B + \delta$。
>
> 详细定义见3.0.2节（AVP定义锚点B1）。

一句话原则：设计靠EML（摩擦+削减）| 验收靠AVP（拔线+对比）

这个判据揭示了两种根本不同的AI使用范式：

场景对比：想象一位医生、程序员或学生，在使用AI辅助工作6个月后，突然失去AI访问权限：

- 认知外骨骼模式：其独立表现显著下降（$P < B$），甚至无法完成原本能够胜任的任务
- 认知内共生模式：其独立表现不仅保持，甚至超越原有水平（$P \geq B + \delta$），因为AI的使用过程强化了其底层能力

这不是假设场景。神经科学研究已经提供了警示信号：Dahmani等人（2020）的研究发现，习惯性使用GPS导航的人群，其空
offloading），其负面效应已有充分的实证基础（Risko & Gilbert, 2016; Sparrow et al., 2011）。

更令人警醒的是这种退化的隐蔽性。每个用户都感觉自己在”进步”——任务完成更快，输出质量更高。但这种表面的能力
—不是因为AI背叛我们，而是因为我们主动放弃了独立思考的能力。

1.1.5 术语澄清与理论命名

关于”认知外骨骼”与”认知内共生”：

本文使用”认知外骨骼”指代AI使用导致人类能力退化的病理模式，“认知内共生”指代AI使用促进人类能力提升的健康模
理论的核心目标是避免外骨骼、建立内共生。

术语层级：
- AVP：评估是否达成内共生的验收标准
- EML：设计内共生系统的构造法则
- 内共生：健康的人机关系（理论目标）
- 外骨骼：病理的依赖状态（理论警示）

后续章节将始终以”促进内共生、避免外骨骼”为理论立场。

1.2 理论缺口：现有范式的共同盲点

当前的人机交互研究和实践，主要由三种范式主导，但它们都未能有效应对上述挑战：

1.2.1 工具范式（Tool Paradigm） 将AI视为被动的效率提升工具，强调人类的完全主导地位。这一范式（如传统HCI研究）

1.2.2 增强范式（Augmentation Paradigm） 如Engelbart（1962）的”智力增强”愿景，关注如何通过技术放大人类能力。
Extended Mind理论（Clark & Chalmers, 1998）虽然承认认知的延展性，但未能区分良性延展（促进成长）与病理延展（导致

1.2.3 自动化范式（Automation Paradigm） 聚焦用AI替代人类完成任务，以实现效率最大化。这一范式对人类长期认知健
& Riley, 1997）和”自动化自满”（automation complacency）研究已经警示了过度依赖自动化系统的风险，但这些洞察尚
这三种范式的共同盲点在于：它们都将AI视为外在于人类认知系统的工具或替代品，而未能将人机关系置于一个动态、共生、

1.3 本理论的核心贡献：从诊断到解决方案

本研究旨在填补上述理论缺口，为AI时代的人机协作提供一个可验证的范式框架。理论贡献体现在四个层面：

1.3.1 建立判别标准：反脆弱性验证原则（AVP） 本研究的首要贡献是将AVP确立为评估人机交互健康性的参考标准。这不仅

AVP的可操作化包括：

- 基线测量（T ）：用户使用AI前的独立能力
- 协作测量（T ）：用户与AI协作时的表现
- 独立测量（T ）：用户在”拔线窗口”（W=4-8周，默认6周，工作假设，需跨领域/任务校准）后的独立能力

测量要求：T 与T 应采用等值平行测验以控制重测与熟悉化效应。

判定标准：

- P ≥ B + δ：成功（认知内共生）
- P ≈ B ：中性（未造成损害但也未促进成长）
- P < B ：失败（认知外骨骼，造成依赖性退化）

1.3.2 确立设计原则：内共生最小法则（EML） 本研究提出内共生最小法则（Endosymbiotic Minimal Law）作为健康人机协作的判定框架：

内共生最小法则（EML）：构成”认知内共生”的设计必要条件为：

（1）有益认知摩擦：使用户处于最优挑战区（成功率50–70%，工作假设，需跨领域/任务校准，个体自适应）

（2）系统性支持削减：AI支持强度按既定削减曲线从S4→S1→S0

二者为联合充分的设计条件，但最终仍需AVP（P ≥ B + δ）作为验收必要条件。

逻辑关系：条件（1）和（2）构成设计层面的联合充分条件；AVP验证是结果验收的必要条件。三者联合构成内共生的充分必

边界条件声明：

边界条件：本理论适用于能力增强型人机协作；补偿性外骨骼（如残障辅助、超越生理极限的设备）不适用此判据。所

1.3.3 重新定义AI角色：从工具到认知伙伴 本研究提出将AI从被动工具重塑为具有“伙伴式主体性”（Partnership Agency）的认知共生体。这种转变的核心不是拟人化，而是功能性的角色重构：

可操作化的三个锚点：

1. 摩擦注入：AI主动创造适度认知挑战（而非总是提供最简单路径）
2. 脚手架消退：遵循系统性支持削减曲线

3. 以AVP闭环为交互终点：协作的最终目标是用户独立能力的提升

这种”伙伴性”与传统HCI追求的”无缝”、“零摩擦”体验形成根本性对立：一个真正的认知伙伴，不应只是顺从的助手，

1.3.4 技术实现框架：分层共生架构（LSA） 本研究提出分层共生架构（Layered Symbiosis Architecture, LSA）作为技术承载。第一章仅概述核心结构，详细实现见第五章：

LSA-F（功能分层）：L1 知识整合 | L2 状态建模 | L3 摩擦校准 | L4 元认知协调

支持档位栈（S4→S1）用于表达支持强度，与LSA-F为正交维度。

表1：认知外骨骼 vs 认知内共生的核心对照

维度	认知外骨骼	认知内共生
设计哲学	替代/卸载	赋能/强化
认知摩擦	最小化	优化（50 - 70%，工作假设）
时间性	永久依赖	临时共生
支持削减	无/固定支持	系统性递减（S4→S1）
AVP结果	P ≤ B	P ≥ B + δ
神经效应*	能力退化趋势	能力增强趋势

*注：神经效应一栏为基于现有认知神经科学研究（如Dahmani et al., 2020; Maguire et al., 2000）的趋势性推断，需要进一步针对性实证研究验证。

1.3.5 理论定位：规范性解决方案框架 本研究定位为一个规范性解决方案框架（normative solution framework）。我们不仅诊断问题，更明确应当如何构造人机协作以避免依赖退化，并用AVP作为参考标准验证是否达标。

与AI对齐研究的关系：

本理论与AI对齐（AI Alignment）研究是互补而非对立的关系：

- 对齐研究关注”AI的意图是否与人类价值一致”
- 本理论关注”人机协作是否促进人类能力的可持续发展”

我们提供的是在假设AI已对齐的前提下，如何设计健康的人机交互模式的完整方案，包括：

- 构造标准（通过EML）
- 验收标准（通过AVP）
- 工程路径（通过LSA）

1.4 研究方法与论文结构

1.4.1 方法论定位 本研究采用理论构建与概念分析相结合的方法。本理论定位为一个可证伪的理论框架：我们提出了一套可

透明性声明：

1. 本文使用的所有量化参数（如Cohen’s d ≥ 0.3 或 ≥10%（working assumption），成功率50 - 70%，拔线窗口W=4-8周，默认6周）均为概念工作模型，基于认知心理学和教育测量学文献的合理推断，但需要跨领域
2. 案例选择遵循理论启发性标准，不追求统计代表性，而是用于阐释机制和边界条件。
3. 我们明确指出理论的适用边界和证伪路径（详见第六章），坚持开放科学原则。

1.4.2 论文结构 本文共六章：第二章梳理跨学科证据基础，第三章构建AVP/EML理论框架， 第四章扩展至团队和组织层面，第六章明确理论边界并提出证伪路径。

CET论文摘要与参考文献模板

中文摘要（结构式摘要，300-500字）

研究目的：随着生成式AI的广泛应用，人类独立认知能力面临退化风险。本研究旨在建立一个可证伪的理论框架，用于评估和

研究方法：本文提出认知能力建构与评估理论（简称”本理论”），核心包括三大机制：（1）反脆弱性验证原则（AVP）——通过”拔线测试”检验协作是否促进独立能力（判据：P ≥B +δ；≥0.3 SD 或 10%（工作假设，需跨领域/任务校准）SD 或 10%，工作假设，需跨领域/任务校准）；（2）内共生最小法则（EML）——设计有益认知摩擦（50–70%成功率，工作假设，需跨领域/任务校准）与系统性支持削减（S4→S1）；（3）伙伴式主体性——AI作为认知伙伴的角色定位。理论采用跨学科综合方法，整合认知心理学、教育技术、组织行为学和AI伦理的研究成果，并通

主要结果：本文建立了四层跨尺度验证体系（个体I-AVP、团队T-AVP、组织O-AVP双阈值模型：告警≥0.70，目标≥0.85，工AVP），识别了”认知外骨骼”（过度依赖导致能力退化）与”认知内共生”（能力持续增强）两种模式，并设计了分层共生H8），为后续实证研究提供清晰路径。

研究结论：本理论为AI时代的能力建构提供了首个系统性、可操作的评估标准。其价值在于：提出了超越表面效率（P ）、团队-组织-社会）的一致性框架；设计了可工程化的技术实现路径（LSA）。理论的所有参数均为工作假设，需通过跨领域实

关键词：人工智能；认知能力建构；反脆弱性验证；人机协作；能力评估；认知韧性

英文摘要（Structured Abstract, 300-500 words）

Objective: With the widespread adoption of generative AI, human independent cognitive capabilities face risks of degradation. This study aims to establish a falsifiable theoretical framework for assessing and optimizing human capability building in the AI era.

Methods: This paper proposes a Cognitive Capability Building and Assessment Theory (abbreviated as “this theory”), comprising three core mechanisms: (1) Antifragility Validation Principle (AVP)—verifying whether collaboration enhances independent capability through “unplugged tests” (criterion: P ≥B +δ, δ≥0.3 SD or 10%, working hypothesis, requires cross-domain/task calibration); (2) Endosymbiotic Minimal Law (EML)—designing beneficial cognitive friction (50–70% success rate, working hypothesis, requires cross-domain/task calibration) and systematic support reduction (S4→S1); (3) Partner-like Agency—positioning AI as a cognitive partner. The theory adopts an interdisciplinary synthesis approach, integrating research from cognitive psychology, educational technology, organizational behavior, and AI ethics, validated through literature analysis, case studies, and conceptual modeling.

Results: This paper establishes a four-layer cross-scale validation system (Individual I-AVP, Team T-AVP, Organizational O-AVP dual-threshold model: alert ≥0.70, target ≥0.85, working hypothesis, requires cross-domain/task calibration, Societal S-AVP), identifies two contrasting patterns—“cognitive exoskeleton” (over-reliance leading to capability degradation) versus “cognitive endosymbiosis” (continuous capability enhancement)—and designs a Layered Symbiosis Architecture (LSA) as an engineering implementation pathway. Through case analyses across three domains (writing, programming, team collaboration), the theory’s explanatory and predictive power is validated. The study explicitly articulates six major limitations and eight falsifiable hypotheses (H1-H8), providing clear pathways for subsequent empirical research.

Conclusions: This theory provides the first systematic, operational assessment standard for capability building in the AI era. Its value lies in: proposing an evaluation paradigm that transcends surface efficiency (P) to focus on independent capability (P); establishing a cross-scale consistency framework (individual-team-organization-society); designing an engineerable technical implementation pathway (LSA). All theoretical parameters are working hypotheses requiring calibration through cross-domain empirical research. We proactively acknowledge the theory’s limitations and invite critique, validation, and transcendence from the academic community.

Keywords: Artificial Intelligence; Cognitive Capability Building; Antifragility Validation; Human-AI Collaboration; Capability Assessment; Cognitive Resilience; Cross-Scale Mechanisms

---

参考文献

[1] Goodhart C A E. Problems of monetary management: The UK experience[C]//Papers in Monetary Economics. Sydney: Reserve Bank of Australia, 1975 DOI: https://doi.org/10.2307/2232004.

[2] Taleb N N. Antifragile: Things that gain from disorder[M]. Random House, 2012.

[3] Vygotsky L S. Mind in society: The development of higher psychological processes[M]. Harvard University Press, 1978.

[4] Bandura A. Social learning theory[M]. Prentice Hall, 1977.

[5] Ericsson K A, Krampe R T, Tesch-Römer C. The role of deliberate practice in the acquisition of expert performance[J]. Psychological Review, 1993, 100(3): 363-406.

[6] Collins A, Brown J S, Newman S E. Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics[M]//Knowing, learning, and instruction. Routledge, 2018: 453-494.

[7] Sweller J. Cognitive load theory[M]//Psychology of learning and motivation. Academic Press, 2011: 37-76.

[8] Mayer R E. Multimedia learning[M]. Cambridge University Press, 2009.

[9] Paas F, Renkl A, Sweller J. Cognitive load theory and instructional design: Recent developments[J]. Educational Psychologist, 2003, 38(1): 1-4.

[10] Van Merriënboer J J, Kirschner P A. Ten steps to complex learning: A systematic approach to four-component instructional design[M]. Routledge, 2017.

注：以上参考文献为示例，实际使用时请根据正文引用情况调整。

参考文献

[1] Goodhart C A E. Problems of monetary management: The UK experience[C]//Papers in Monetary Economics. Sydney: Reserve Bank of Australia, 1975 DOI: https://doi.org/10.2307/2232004.

[2] Taleb N N. Antifragile: Things that gain from disorder[M]. Random House, 2012.

[3] Vygotsky L S. Mind in society: The development of higher psychological processes[M]. Harvard University Press, 1978.

[4] Bandura A. Social learning theory[M]. Prentice Hall, 1977.

[5] Ericsson K A, Krampe R T, Tesch-Römer C. The role of deliberate practice in the acquisition of expert performance[J]. Psychological Review, 1993, 100(3): 363-406.

[6] Collins A, Brown J S, Newman S E. Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics[M]//Knowing, learning, and instruction. Routledge, 2018: 453-494.

[7] Sweller J. Cognitive load theory[M]//Psychology of learning and motivation. Academic Press, 2011: 37-76.

[8] Mayer R E. Multimedia learning[M]. Cambridge University Press, 2009.

[9] Paas F, Renkl A, Sweller J. Cognitive load theory and instructional design: Recent developments[J]. Educational Psychologist, 2003, 38(1): 1-4.

[10] Van Merriënboer J J, Kirschner P A. Ten steps to complex learning: A systematic approach to four-component instructional design[M]. Routledge, 2017.

注：以上参考文献为示例，实际使用时请根据正文引用情况调整。