承载性困难与摩擦性困难：生成式人工智能情境下认知外包边界的事前判据研究

Load-Bearing versus Frictional Difficulty: Ex-Ante Criteria for the Boundary of Cognitive Offloading under Generative AI

> 认知负荷理论告诉我们要减少外在负荷、保留相关负荷。它没有告诉我们的是：面对一个具体的困难，怎么在移除它之前知道它是哪一种。在只能移除少数困难的年代，这个空白无关紧要；当任何困难都可以被一键移除，它成了整个教学设计的地基缺口。

王德生 著 · 学术创造专栏 · 学位论文频道 · 概念建构型

硕士学位论文范本 · 教育技术学

摘 要

生成式人工智能使学习过程中几乎任何认知困难都可被即时移除，这使"哪些困难应当保留"成为教学设计必须回答、却尚无工具回答的问题。本研究指出，现有两大理论传统在此处存在同一个结构性缺口：认知负荷理论区分内在、外在与相关负荷，必要难度理论区分合意与非合意的困难，但两者对某一具体困难的归类均依赖于移除该困难后的学习效果——即分类由结果定义，因而只能事后作出。在教师仅能通过修改教材与教法来移除少量困难的年代，这一循环并不构成实践问题；当移除成本趋近于零且移除权转移到学习者手中时，它使教学设计失去全部事前判断力。为填补该缺口，本研究提出承载性困难与摩擦性困难的区分：前者的克服过程即为待习得能力本身，后者仅是该过程的伴随成本。研究进一步提出三条可在移除前施用的操作性判据——区分判据（克服该困难是否要求学习者作出一次判定而非执行一个已知程序）、残留判据（困难克服后是否有内容留存于学习者而非仅留存于产物）、可验证性判据（若由他人或工具代为克服，学习者能否独立验证其结果）——并论证第三条判据识别出的一类困难最为危险：它们不可让渡，却在实践中最容易被让渡，且让渡者无从察觉。本研究属概念建构类型，第四章给出五项可证伪预测及相应的准实验与出声思维协议设计，第五章以人因工程领域关于自动化技能侵蚀的既有经验证据检验框架的解释力，第六章讨论其对教学设计与学习工具设计的含义，并指出一个不受欢迎的推论：可验证性判据意味着，越是初学者，可让渡给工具的困难越少，而当前工具的普及方向恰好相反。

关键词 认知外包；认知负荷理论；必要难度；承载性困难；生成式人工智能；教学设计；自动化悖论

ABSTRACT

Generative AI makes nearly any cognitive difficulty in learning removable on demand, turning the question of which difficulties ought to be preserved into one that instructional design must answer but currently has no instrument to answer. This study identifies a shared structural gap in the two governing traditions: cognitive load theory distinguishes intrinsic, extraneous and germane load, and the desirable-difficulties literature distinguishes desirable from undesirable difficulty, yet in both, the classification of any particular difficulty depends on the learning outcome observed

after that difficulty is removed. Classification is defined by its result and can therefore only be made ex post. While teachers could remove only a handful of difficulties, and only by redesigning materials, this circularity carried no practical cost. Once removal is free and the power to remove passes to the learner, it strips instructional design of all ex-ante judgment. The study proposes a distinction between load-bearing and frictional difficulty: overcoming the former constitutes the very capability to be acquired, whereas the latter is merely an accompanying cost of that process. Three operational criteria applicable before removal are advanced — the discrimination criterion, the residue criterion, and the verifiability criterion — and it is argued that the class identified by the third is the most dangerous: such difficulties cannot be delegated, are the most readily delegated in practice, and their delegation is undetectable by the delegator. Being a work of concept construction, the study offers five falsifiable predictions with a quasi-experimental and think-aloud design, tests the framework's explanatory reach against existing evidence on automation-induced skill erosion from human factors research, and closes with an unwelcome corollary: the verifiability criterion entails that the less expert the learner, the fewer difficulties may be safely delegated — while current tools are diffusing in precisely the opposite direction.

Keywords cognitive offloading; cognitive load theory; desirable difficulties; load-bearing difficulty; generative AI; instructional design; ironies of automation

第一章 绪论

1.1 研究背景

教学设计一百年来的主线是一个减法：把学习中不必要的障碍去掉。从夸美纽斯的直观教学，到程序教学机，到多媒体学习原则，到今天的自适应学习平台，改进的方向高度一致——**让学习者少走弯路**。这条主线在效果上是可辩护的：大量研究表明，冗余信息、割裂注意、无关装饰确实损害学习，去掉它们确实有效。

这条主线有一个从未被写进任何教科书、却支撑着它全部合理性的前提：**教师能够去掉的障碍是有限的，而且去掉它们需要成本**。教师能改的是教材编排、呈现方式、任务顺序。他改不了的是——学生在解一道题时必须自己走的那些步骤。那些步骤不在教师的可操作范围内，因此从来不需要被判断该不该去掉。

这个前提在二〇二二年之后失效了。

生成式人工智能的能力边界目前仍在移动，但有一件事已经确定：**在文本类学习任务中，它可以代替学习者完成几乎任何中间步骤**。不是加快，是代替。而且这个代替的成本趋近于零，更关键的是——**这个决定权不在教师手上，在学生手上，且不可观察**。

于是一百年来的那条减法主线，第一次撞上了它自己的极限：如果减法是好的，为什么不减到底？如果去掉障碍有助于学习，那么去掉全部障碍岂不是最好？

所有人都知道这个推论是荒谬的。但**为什么荒谬**，现有理论说不清楚。

1.2 问题的提出

本研究要处理的问题可以用一句话表述：

> **面对一个具体的学习困难，我们能否在移除它之前，判断它该不该被移除？**

请注意"之前"这两个字。这是全部问题所在。

现有的理论并非对困难的价值没有认识。恰恰相反，认知负荷理论明确区分了应当削减的外在负荷与应当保留甚至增加的相关负荷（Sweller, 1988; Sweller et al., 1998）；必要难度理论更是直接指出，某些增加学习过程困难的操作会损害即时表现却提升长期保持与迁移（Bjork, 1994; Bjork & Bjork, 2011）。这两个传统都承认：**有些困难必须留下**。

问题不在于它们不承认，而在于它们**如何判断某一个困难属于哪一类**。

在认知负荷理论里，一个负荷是"外在"还是"相关"，最终由实验判定：设计两个版本，一个含此设计要素，一个不含，比较学习效果；若含此要素的组表现更好，该要素承载的即为相关负荷，反之为外在负荷。必要难度的判定同理：一个困难是不是"合意"的，取决于加入它之后延迟测验的成绩是否更高。

也就是说：

> **一个困难的类别，是由移除它之后发生了什么来定义的。**

这在逻辑上是一个循环。它并非理论的疏漏，而是这两个理论作为经验科学的正当工作方式——它们本来就是从效果反推分类的。

在过去，这个循环不构成实践问题。因为可供移除的困难数量少、移除成本高，教育研究者有充分的时间对每一个候选设计做实验，然后把结论写成教学设计原则。整个多媒体学习原则体系（Mayer, 2009）就是这样积累起来的：一条一条做实验，一条一条积累。

现在，可供移除的困难数量趋于无穷，移除成本趋于零，移除的决定权转移给了学习者本人，且移除行为不可观察。 在这个条件下，"先做实验再给原则"的路径失效了——不是因为它错，是因为它太慢，而且它给出的原则是针对教师的操作，而现在做决定的不是教师。

因此本研究的核心问题是：**能否建立一组不依赖于移除后果、可在移除前施用的判据？**

1.3 研究意义

理论意义。若能建立事前判据，则认知负荷理论与必要难度理论在困难分类上的循环依赖可被部分解除。这不是推翻它们——两个理论的经验发现全部保留——而是为它们补上一个此前不需要、现在必需的前置环节。

实践意义。当前学校应对生成式人工智能的主流做法有二：一是检测与禁止，二是设计"人工智能做不了的任务"。前者的效力随模型能力单调下降；后者的问题在于它把判断标准建立在工具的当前能力上，因

而其结论的保质期由工具厂商单方面决定。若存在一组基于**困难自身性质**的判据，则教学设计可以获得一个不随工具能力漂移的立足点。

工具设计意义。本研究的判据同时是对学习类人工智能产品的一个设计约束。第六章将论证，当前绝大多数此类产品在同一个位置上犯了同一个错误。

1.4 研究定位与方法

本研究属于**概念建构型研究**（conceptual research），其产出是一组概念区分与操作判据，而非一组实验数据。研究者认为这一定位是诚实的：本研究不拥有可支撑因果推断的原始数据，因此不伪装拥有。

方法路径为：

1. **裂缝定位**（第二章）：通过对两大理论传统的结构性审查，定位循环依赖的确切位置及其在新条件下的后果。

2. **概念建构**（第三章）：提出承载性困难与摩擦性困难的区分及三条事前判据，并逐一处理边界情形。

3. **可证伪化**（第四章）：将框架转写为五项可独立失败的预测，并给出相应的准实验与出声思维协议设计。

4. **解释力检验**（第五章）：以人因工程领域关于自动化导致技能侵蚀的既有经验证据作为外部检验场，考察框架能否解释该领域已有的发现，并主动扫描反例。

5. **含义推演与自我限定**（第六、七章）。

需要说明的是，第五章使用的是他人已发表的经验证据，用于检验框架的解释力，而非用于证明框架为真。一个框架能解释既有发现是必要条件而非充分条件——第四章的预测才是它可以被杀死的地方。

1.5 论文结构

第二章审查文献并定位缺口；第三章为核心章，提出概念与判据；第四章给出可证伪设计；第五章检验解释力；第六章讨论含义；第七章总结并陈述局限。

第二章 文献综述

2.1 认知负荷理论及其困难分类

认知负荷理论由 Sweller（1988）提出，其核心主张是工作记忆容量有限，教学设计应管理施加于其上的负荷。理论区分三类负荷：内在负荷（intrinsic load），由学习材料自身的要素交互性决定；外在负荷（extraneous load），由呈现方式引入、与学习无关；相关负荷（germane load），指投入到图式建构中的资源（Sweller, van Merriënboer & Paas, 1998）。

由此推出的教学原则是清晰的：削减外在负荷，管理内在负荷，为相关负荷腾出空间。围绕这条主线，一系列效应被反复验证：注意分散效应、通道效应、冗余效应、样例效应等（Mayer, 2009; Kirschner, Sweller & Clark, 2006）。

理论后期经历过一次重要修正。Sweller 等人在检讨中承认，相关负荷难以与内在负荷在操作上分离——若相关负荷指的是处理内在负荷所投入的资源，则它不是一个独立的负荷来源（Sweller, Ayres & Kalyuga, 2011）。这次修正本身是诚实的，但它把本研究关心的问题推得更深：**如果连相关负荷是否独立存在都存疑，那么"某个具体困难承载的是外在负荷还是相关负荷"这个判断，就更加只能由实验结果来回答**。

2.2 必要难度研究传统

与认知负荷理论的减法取向不同，Bjork 的研究传统指出：某些增加学习过程难度的操作，会降低即时表现却提升长期保持与迁移。这些操作被称为"必要难度"或"合意难度"（desirable difficulties）：分散练习优于集中练习、交错练习优于分块练习、提取练习优于重复阅读、生成优于呈现（Bjork, 1994; Bjork & Bjork, 2011; Roediger & Karpicke, 2006）。

这一传统最重要的贡献，是指出了学习者与教师的**主观判断系统性地误导他们**：学习者觉得流畅的方法，往往是效果差的方法（Bjork, Dunlosky & Kornell, 2013）。Deslauriers 等（2019）在物理课堂上的研究给出了一个尖锐的证据：接受主动学习的学生实际学得更多，但自我感觉学得更少；接受传统讲授的学生感觉良好而实际较差。**学习的主观流畅感与实际学习效果之间存在系统性的反向偏差**。

这一发现对本研究至关重要，理由在第三章展开：如果学习者的主观判断系统性地把有效的困难感知为障碍，那么把移除困难的决定权交给学习者，其后果是可预测的。

但必要难度理论同样存在本研究关心的那个循环：一个困难是否"合意"，由延迟测验的结果定义。Bjork 与 Bjork（2011）本人明确提示，并非所有困难都是合意的——困难若超出学习者当前能力，则只是困难而已。**如何区分？由结果区分**。

2.3 生成效应、提取练习与卸载

生成效应（generation effect）指出，自己生成的信息比被动阅读的信息记得更牢（Slamecka & Graf, 1978）。提取练习效应指出，从记忆中取出比重复输入更有效（Roediger & Karpicke, 2006）。这两条发现在本研究框架下有一个共同的读法：**它们表明"过程"本身在生产某种在产物中不可见的东西**——因为在两组比较中，最终掌握的内容是一样的，不一样的只是获得它的路径。

认知卸载（cognitive offloading）研究则处理相反的一面：人们如何把认知任务转移到外部（Risko & Gilbert, 2016）。该领域的经典发现之一是所谓"谷歌效应"：被告知信息会被保存的被试，对信息本身记得更少，但对"去哪里找"记得更多（Sparrow, Liu & Wegner, 2011）。

卸载研究的取向基本是描述性与中立的：它研究人们卸载什么、何时卸载、卸载的收益与代价。它没有提供、也不试图提供一个规范性的判据来回答"什么该卸载"。

2.4 自动化对技能的侵蚀：人因工程的证据

一个几乎未被教育技术研究引用、却与本研究直接相关的文献群，来自人因工程。

Bainbridge（1983）在《自动化的讽刺》中提出了一组至今未被解决的悖论：自动化系统把常规操作接管过去，只在异常情况下要求人类操作员接管；但正是常规操作的持续执行，维持着操作员在异常时接管所需的技能与情境理解。**自动化拿走了维持能力的那个过程，却在最需要能力的时刻把控制权还回来**。

Parasuraman 与 Riley（1997）系统化了这一问题，区分了自动化的使用、误用、废用与滥用；Endsley（1995）的情境意识理论进一步指出，操作员对系统状态的理解，是在持续参与中建立起来的，一旦参与被自动化中断，情境意识随之衰减，且**衰减不可自察**。

这一文献群的价值在于：它是在教育之外、在一个后果由飞机坠毁而非考试分数来结算的领域里，独立地撞上了同一个结构。第五章将以此作为框架的外部检验场。

2.5 文献述评：三个缺口

综合以上，本研究识别出三个缺口：

缺口一：分类的循环依赖。认知负荷理论与必要难度理论都能判定一个困难该不该保留，但都只能事后判定。在移除成本高昂的年代这不成问题；在移除成本为零的条件下，它使理论失去事前判断力。

缺口二：决定权的位置。两个理论的全部原则都是给教师的操作指南——它们预设移除的决定由教学设计者作出。生成式人工智能把这个决定权转移给了学习者，而第 2.2 节引述的证据表明，学习者的主观判断在这件事上系统性地不可靠。**理论的接收对象与实际的决定者已经不是同一个人，而这一点尚未被讨论**。

缺口三：教育技术研究与人因工程之间的隔离。自动化技能侵蚀的文献积累了四十年，其结构与当前教育面临的问题高度同构，但两个领域几乎不互相引用。

本研究针对缺口一，并在第六章回应缺口二，在第五章尝试打通缺口三。

第三章 概念建构：承载性困难与摩擦性困难

3.1 一个被工具能力掩盖了一百年的区分

先给出本研究的核心区分。

> **承载性困难（load-bearing difficulty）：其克服过程本身即构成待习得能力的困难。学习者所要学会的，就是克服它的那个过程；该过程被替代，则学习不发生。**

>

> **摩擦性困难（frictional difficulty）：其克服过程不构成待习得能力的困难。它是学习过程的伴随成本，而非载体。该过程被替代，学习不受损，且资源得以释放。**

举例可以立刻说明这个区分的实在性：

一个学生学习统计推断。他需要：（a）计算一组数据的标准差；（b）判断这组数据是否满足 t 检验的前提假设。

（a）是摩擦性困难。他学的不是"如何做四则运算"——那是他早已掌握的程序的重复执行。用计算器替代它，他失去的只是时间。事实上，统计教育在计算器普及之后并未受损，反而第一次能把课时投入到本该投入的地方。**计算器不是被容忍的，它是被感谢的。**

（b）是承载性困难。他要学会的**就是**做这个判断本身。没有别的东西是"统计推断能力"——它不是关于前提假设的知识（那是一句话，五秒钟可以告诉他），它是在一个具体的、混乱的、条件不齐的数据面前作出这个判断的能力。若有人替他判断，他学到的是那个判断的结论，而不是作出判断的能力。**而这两样东西在他交上来的作业上，长得一模一样。**

这个区分为什么一百年来没有被提出？因为**它以前没有实践价值**。教师能替学生做（a）（给他计算器），不能替他做（b）——没有那个工具。既然不能，就不必区分。

3.2 循环定义为什么必须被打破

有人会问：既然认知负荷理论已经可以通过实验判定哪些该留，为什么还要事前判据？做实验不就行了？

三个理由：

其一，数量。教师能改的设计要素是几十个量级，可以一个一个做实验。学生在完成一份作业时可外包的中间步骤是几百到几千个量级，且随任务变化。实验路径在数量上不可行。

其二，速度。一个设计要素的实验周期以年计（设计、施测、延迟后测、发表）。工具能力的迭代周期以月计。等实验做完，被研究的那个条件已经不存在了。

其三，也是最根本的一条：决定者不在场。所有实验结论的形式都是"如果这样设计，学习效果会更好"，它的接收者是设计者。而现在做出移除决定的是学生本人，在深夜，在宿舍，面对一个截止日期。**你不可能给他一份实验结论清单让他查阅——而且第 2.2 节的证据表明，即使你给了，他的主观流畅感也会把他推向相反的方向。**

因此需要的不是更多的实验结论，而是一**组判据**：形式简单到可以在决定的当下施用，且不依赖于移除后果。

3.3 三条事前判据

判据一：区分判据（Discrimination Criterion）

> **克服该困难，是否要求学习者作出一次判定（"这个成立，那个不成立"），而非执行一个已知的程序？**

> **要求判定 → 倾向承载性；仅执行程序 → 倾向摩擦性。**

这条判据的依据是：能力的核心是判定，不是执行。凡是可以被写成一个确定程序、按步骤跑完即得结果的过程，其能力已经凝固在程序里，学习者重复执行它并不增加什么——**程序本身可以被存储、传递、自动化，而这正说明它不必长在人身上。**

操作化：给定一个困难，问"能不能写出一份步骤清单，使得一个不理解这个领域的人照做也能得到正确结果？"能 → 摩擦性；不能 → 承载性。

判据二：残留判据（Residue Criterion）

> **该困难被克服之后，是否有东西留存于学习者身上，而非仅留存于产物之中？**

> **有残留 → 承载性；只留在产物里 → 摩擦性。**

这条判据把生成效应与提取练习效应的经验发现提炼成了一个判准。那些实验的结构都是：两组最终得到的产物相同，过程不同，而后测差异显著。**差异只能来自过程留在人身上的那部分。**

操作化：问"如果这份作业交上去之后被烧掉，这个学生身上还剩下什么？"若答案是"什么都不剩"，该困难是摩擦性的（对这个学生而言，它本来就没在生产任何东西）。

判据三：可验证性判据（Verifiability Criterion）

> **若该困难由他人或工具代为克服，学习者能否独立验证其结果的正确性？**

> **能验证 → 可让渡；不能验证 → 不可让渡。**

这条判据是三条里最重要的一条，也是本研究的核心贡献。它的依据是一个不对称：

你可以安全地外包一件你能够检查的事；你不能安全地外包一件你无法检查的事——因为你连自己被给了错误答案都不会知道。

计算器算出 $847\times36=30492$。学生能验证吗？能——他可以估算（$800\times36\approx28800$，量级对），可以逆运算，可以用另一种方法重算。**验证能力保留在他身上，所以外包是安全的。**

模型给出一份统计方法的选择建议。学生能验证吗？如果他已经会判断前提假设，能——那么这次外包只是省时间。**如果他不会判断——而他正是为了学这个才来上课的——他不能验证。** 他会接受这个建议，交上作业，得到分数，并且相信自己学会了。

这里出现了本研究认为最危险的一类困难：

> **不可让渡的困难被让渡了，而让渡者无从察觉。因为一旦他有能力察觉，他就不需要让渡了。**

这不是一个概率问题，是一个结构性质：**验证能力与被外包的能力是同一样东西**。当你缺乏能力 X 时，你恰好也缺乏"判断关于 X 的输出是否正确"的能力。**这就是为什么认知外包在教育情境中与在专业情境中性质完全不同**：一个统计学家外包统计计算是安全的，一个统计学学生外包统计判断是致命的——两人做的动作看起来一模一样。

3.4 三条判据的关系与边界情形

三条判据不是并列的三个投票，它们有一个次序：

判据三是安全阀，判据一与判据二是识别器。 判据一与判据二识别一个困难是否承载性；判据三判断在当前这个学习者身上，让渡是否安全。**同一个困难，对不同水平的学习者，判据三的结论不同。**

这引出一个必须直面的边界情形：

情形 A：承载性且可验证。一个已经掌握了统计判断的研究生，让模型给出方法建议然后自己检查。判据一、二说这是承载性困难，判据三说他能验证。结论：**可以让渡，且应当让渡**——他学过了，这里不再产生学习，只产生时间成本。

这说明本框架不是一个禁令清单，而是一个条件判断：**同一个困难，在学的时候不可让渡，在会了以后应当让渡。** 教育的目标恰恰是把承载性困难变成可让渡的。

情形 B：摩擦性但不可验证。一个学生让模型帮他排参考文献格式，他不知道格式对不对。这符合判据三的"不可验证"，但没有人认为这有问题——因为格式本来就不在他要学的东西里。**这说明判据三不能单独使用**：它只在判据一、二已经判定为承载性的困难上才有意义。

情形 C：判据一与判据二冲突。一个困难要求判定（判据一 → 承载），但学生做完之后什么也没留下（判据二 → 摩擦）。这种情形是真实存在的：例如一个学生在完全超出其能力的任务上做判定，他确实在判定，但他的判定是随机的，因而什么也没留下。**这恰是 Bjork 所说的"困难而非合意困难"的情形**，本框架的处理是：判据二的失败提示任务超出了当前能力区间，此时应当降低任务难度，而不是外包判定。

3.5 与既有概念的关系

必须明确本框架与既有概念的关系，否则它只是换了个说法。

与内在／外在负荷的关系：不同层的东西。 内在／外在负荷区分的是负荷的**来源**（材料自身 vs 呈现方式）。承载性／摩擦性区分的是困难与**待习得能力**的关系。一个外在负荷可以是摩擦性的（版式混乱），一个内在负荷也可以是摩擦性的（一个不需要学会的复杂计算）。**两组区分正交，不可互相替换。**

与必要难度的关系：本框架是它的事前化尝试。 必要难度理论说的是"某些困难有益"，本框架试图回答"哪些、凭什么、在移除前怎么知道"。承载性困难大体上是合意困难的一个子集——但不是全部：分散练习是合意困难，它却不满足判据一（它不要求判定）。**这说明合意困难至少有两个来源：一个是本框架处理的"过程即能力"，另一个是记忆机制层面的（间隔、提取），后者不在本框架射程内。这是一个必须承认的覆盖缺口。**

与认知卸载研究的关系：本框架是它的规范性补充。 卸载研究描述人们如何卸载，本框架试图给出该不该卸载的判据。

第四章 可证伪预测与研究设计

4.1 设计逻辑

一个框架若不能被杀死，它就什么也没说。以下五项预测各自独立可失败——任意一项被证伪，框架的对应部分即告失效，且不能靠解释来挽救。

4.2 五项可证伪预测

预测一（判据一的效力）：在同一学习任务中，外包"要求判定"的子步骤，与外包"仅执行程序"的子步骤相比，前者在延迟迁移测验上的损害显著更大；而在即时任务表现上，两者无显著差异或前者略优。

证伪条件：若两类外包在延迟迁移上无显著差异，判据一失效。

预测二（残留判据的可测性）：以"产物销毁后测验"操作化残留——完成任务后立即回收全部产物，一周后不预告地测验相关能力。承载性困难被外包的组，其后测成绩与"未接触过该任务"的对照组无显著差异。

证伪条件：若外包组显著优于未接触组，说明外包过程仍产生了残留，判据二失效。

预测三（可验证性判据的交互作用）：外包同一个承载性困难，对高先备知识学习者无损害或有益，对低先备知识学习者有显著损害。即先备知识与外包之间存在显著交互作用。

证伪条件：若交互作用不显著（外包的损害不随先备知识变化），判据三失效——这将是对本研究核心贡献的直接否定。

预测四（不可察觉性）：低先备知识学习者在外包了不可验证的承载性困难之后，其对自身掌握程度的信心评分不低于、甚至高于未外包组，而实际成绩显著更低。即**信心与成绩的分离在此条件下最大。**

证伪条件：若外包组信心显著低于未外包组（即他们知道自己没学会），则"无从察觉"这一论断失效。

预测五（判据的可操作性）：两名独立编码者依据三条判据，对同一批学习任务的子步骤进行分类，其一致性（Cohen's κ）应达到可接受水平（κ > .70）。

证伪条件：若一致性低于该阈值，则判据虽然在概念上成立，在操作上不可用——这足以使本框架失去实践价值。

预测五是本框架最先应当被检验的一项，因为它最便宜，且它的失败使其余四项无从谈起。

4.3 研究一：2×2 准实验设计

设计：2（外包对象：判定步骤 vs 程序步骤）× 2（先备知识：高 vs 低）被试间设计。

被试：某高校非统计专业本科生，以前测划分高低先备知识组（中位数分组，剔除中间 20% 以增强区分）。

任务：一个统计方法选择任务，含明确可分离的判定步骤（前提假设检验、方法适配）与程序步骤（描述统计量计算、图表生成）。

操作：外包通过提供一个受控的对话式助手实现，该助手仅在被指定的子步骤上作答，在其余步骤上拒答。

测量：（1）即时任务表现；（2）一周后延迟迁移测验（新数据、新情境）；（3）每步骤后的信心评分；（4）过程日志。

关键比较：预测三的交互作用项。

4.4 研究二：出声思维协议分析

准实验只能回答"是否有损害"，不能回答"损害发生在哪一步"。研究二采用出声思维协议（Ericsson & Simon, 1993），对研究一中各条件下的子样本（每格 n=8）录音编码。

编码框架（依三条判据构造）：

- 判定事件：学习者明确表达一次取舍（"这个不行，因为……"）

- 求助事件：学习者转向助手

- 验证事件：学习者对助手输出提出质疑或独立核查

- 接受事件：学习者未经核查即采纳

核心指标：验证事件 / 接受事件之比。预测：该比值与先备知识正相关，且该比值可预测延迟迁移成绩，其预测力高于外包总量。

这一指标是本研究框架最直接的行为学操作化：它测的不是"用没用 AI"，而是"用的时候有没有在验证"——而后者才是判据三真正关心的东西。

4.5 效度威胁与应对

威胁一：先备知识分组的构念效度。前测测的可能是应试能力而非真实先备知识。应对：前测采用迁移型题目而非再认型题目。

威胁二：助手能力的生态效度。受控助手比真实工具弱，可能低估效应。应对：如实报告，并将本设计的结论限定为效应的下界。

威胁三：霍桑效应。被试知道自己在被研究外包行为，可能减少外包。应对：掩饰研究目的，将研究表述为"统计学习材料评估"。

威胁四，也是最严重的一条：出声思维本身改变过程。要求学习者说出自己在想什么，可能诱发本来不会发生的验证行为——**而验证行为恰是本研究的核心因变量**。这个威胁在原理上无法通过设计消除，只能通过与无出声思维组的表现比较来估计其大小，并在报告中如实陈述。

第五章 解释力检验：来自自动化研究的外部证据

5.1 三个已有发现的重新解读

一个框架的解释力，最好在它没有参与构造的证据上检验。本节使用人因工程领域四十年前就已积累的经验发现。

发现一：自动化的讽刺（Bainbridge, 1983）。自动化系统接管常规操作，只在异常时要求人接管；而人接管所需的技能，恰恰由常规操作维持。结果是自动化程度越高，异常时的接管质量越差。

本框架的解读：常规操作对操作员而言是**承载性困难**——它看起来是重复劳动（像摩擦），但它的执行过程在维持着某种在产物上不可见的东西（判据二：有残留）。自动化把它当摩擦性困难移除了，于是残留停止生产。**四十年前的这个发现，与本研究的判据二是同一件事的两次独立发现——一次在驾驶舱，一次在课堂**。

发现二：情境意识的不可自察衰减（Endsley, 1995）。被自动化中断参与的操作员，其对系统状态的理解会衰减，且**他们不知道自己衰减了**——他们的主观信心并不随之下降。

本框架的解读：这正是预测四所描述的现象，而且是在一个后果远为严重的领域被独立观察到的。它同时支持了 3.3 节关于"验证能力与被外包能力同一"的论证：操作员失去的正是"判断自己是否还有情境意识"的那个能力。

发现三：自动化的使用与误用（Parasuraman & Riley, 1997）。该研究区分了对自动化的适当使用、误用（过度信任）、废用（不信任）与滥用。其中"误用"的核心机制是校准失败：**人对自动化可靠性的信任，与其实际可靠性不匹配，且缺乏校准手段**。

本框架的解读：校准手段就是验证能力。判据三说的是：当验证能力缺失时，信任校准在原理上不可能——不是操作员懒得校准，是他没有可以用来校准的东西。

5.2 反例扫描

一个框架若只找支持它的证据，它什么也没证明。本节主动扫描对本框架不利的证据。

反例一：谷歌效应可能是适应而非损害。Sparrow 等（2011）发现被试记不住信息但记得住位置。这可以读作损害，也可以读作有效的分工——人类记忆本来就一直在外包（书籍、笔记、他人）。**若外包只是把记忆的对象从内容换成了索引，而整体认知效能未降，则本框架把它判为损害就是错的。**

本框架的回应：谷歌效应涉及的是信息**存储**，而存储在判据一下是程序性的（"去哪里找"是一个可写成步骤的过程）。因此本框架不把它判为承载性困难。**但这个回应有循环嫌疑**——它可能是在事后调整判据来容纳反例。诚实的表述是：本框架在记忆型任务上的适用性存疑，其射程可能仅限于判定型任务。

反例二：认知负荷理论的样例效应。大量研究表明，给初学者提供完整的解题样例（worked example），比让他们自己解，效果更好（Sweller & Cooper, 1985）。这似乎直接反对本框架：解题过程明显是承载性的（要求判定、有残留），但把它拿走反而更好。

这是对本框架最强的反例，必须正面回答。

回答：样例效应有一个已被反复验证的边界——**专长逆转效应**（expertise reversal effect, Kalyuga et al., 2003）：样例对初学者有益，对已有一定知识的学习者反而有害。也就是说，"该不该拿走这个过程"的答案随学习者水平反转。**这恰恰是本框架判据三所预测的结构**——只是方向相反：判据三说高先备知识者外包安全，样例效应说低先备知识者被拿走过程反而更好。

如何调和？本框架的解释是：对一个完全没有图式的初学者，让他"自己判定"并不产生判定——他没有可用来判定的东西，他的过程是随机搜索，因而没有残留（判据二失败）。**这正是 3.4 节情形 C。**样例在此提供的不是外包，是**判定所需要的对象**——他需要先看见几次判定长什么样，才可能自己做判定。

但必须承认：**这个解释使本框架在初学阶段的操作变得复杂**——它要求判定"这个学习者是否已经有了可供判定的东西"，而这个判断本身缺乏事前判据。**这是本框架的一处真实缺口，第七章将它列为局限。**

第六章 讨论

6.1 对教学设计的含义

其一，任务设计的重心从"设计任务"转向"标注任务"。若三条判据可操作（预测五），则教学设计者的一项新工作是：对一个任务的每个子步骤标注其类别，并把这个标注**交给学生**。这与当前"设计 AI 做不了的任务"的思路根本不同：后者试图把工具挡在门外，前者承认工具在场，转而告诉学生哪几步不能给它。

其二，禁令必须给出理由，且理由必须是可检验的。"不许用 AI"是一条无法被学生内化的规则，因为它没有区分。"这一步你可以让它做，那一步不行，因为你现在还检查不了它给的对不对"——这是一条学生可以自己复用的规则。**它甚至可以被学生用来反驳教师：如果我能检查，凭什么不让我用？**而这个反驳是正当的，这正是本框架的力量所在。

6.2 对学习工具设计的含义

本框架对学习类人工智能产品构成一条硬约束，而当前绝大多数产品在同一个位置违反它：

它们在承载性步骤上直接给出结论。 一个学生问"这组数据该用什么检验"，产品给出"建议使用 Welch's t 检验，因为方差齐性假设不满足"。这个回答是正确的、有帮助的、用户满意度很高的，**而它恰好在判据三判定为不可让渡的地方完成了让渡**。

符合本框架的设计应该是：产品不给结论，给判定所需的对象——把方差齐性检验的结果摆出来，把两种检验的适用条件摆出来，然后停住。**让学生自己说出那一句。** 若他说错，不纠正，而是给出这个选择在这组数据上的后果，让结果去纠正他。

这条设计约束的代价是诚实的：**它会使产品的即时体验变差、留存变低**。一个把结论给你的产品比一个逼你自己判定的产品好用得多。这意味着符合本框架的产品在商业上处于劣势，除非评价标准改变。**本研究不认为这个问题可以在设计层面解决——它是一个激励结构问题，超出本文射程**。

6.3 一个不受欢迎的推论

判据三有一个推论，本研究认为它是本框架最有价值也最令人不适的产物：

> **学习者水平越低，可安全让渡的困难越少。**

因为可让渡性取决于验证能力，而验证能力随水平上升。一个专家可以安全地外包几乎一切，一个初学者几乎什么都不能外包。

而当前工具的普及方向恰好相反：**它们最先、最深地进入的是初学者群体**——因为初学者最需要帮助，最容易被"它帮我把作业做完了"打动，也最没有能力判断自己被帮了什么。

这个推论意味着：生成式人工智能对学习的影响，很可能**不是均匀的削弱，而是不平等的放大**。已有基础的人用它变得更强（他们能验证），没有基础的人用它变得更弱（他们不能验证，且不知道）。**而两者的作业看起来一样好**。

若这个推论成立，那么当前关于"人工智能促进教育公平"的主流叙事，方向可能是反的。本研究没有数据支持这一推断——它是判据三的逻辑后果，它的检验需要一项长期的、分层的追踪研究，这是本框架最重要的待检验预测。

第七章 结论、局限与展望

7.1 结论

本研究提出：认知负荷理论与必要难度理论在困难分类上共享一个循环依赖——分类由移除后果定义，因而只能事后作出。该循环在移除成本高昂的年代无害，在移除成本为零、决定权转移至学习者、且移除不可观察的条件下，使教学设计丧失事前判断力。

为此本研究建构了承载性困难与摩擦性困难的区分，并提出三条事前判据：区分判据、残留判据、可验证性判据。其中可验证性判据识别出一类结构性危险：**验证能力与被外包的能力是同一样东西，因而当一个人最需要保留某项困难时，他恰好也最没有能力知道自己不该外包它**。

框架被转写为五项可独立证伪的预测，并在人因工程关于自动化技能侵蚀的既有证据上通过了解释力检验——该领域四十年前独立发现的"自动化的讽刺"与"情境意识不可自察衰减"，在本框架下是判据二与判据四的另一次显形。

7.2 局限

其一，无原始数据。本研究是概念建构，全部经验主张均为预测而非发现。第四章的五项预测尚未实施。任何把本文当作实证结论引用的做法都是误用。

其二，射程可能仅限判定型任务。5.2 节的反例扫描显示，本框架在记忆型任务上的适用性存疑，而对分散练习这类基于记忆机制的必要难度，本框架完全无法覆盖。**合意困难至少有两个来源，本框架只处理了一个**。

其三，初学阶段的判据缺失。样例效应与专长逆转效应表明，在学习者尚无可供判定的对象时，本框架的判据一会给出错误建议。3.4 节的情形 C 指出了这个问题，但没有解决它——"如何事前判断一个学习者是否已具备判定所需的对象"，本研究没有判据。**这是本框架最实质的缺口**。

其四，判据的操作一致性未经检验。预测五未实施。若 κ 不达标，本框架在实践上不可用。

其五，出声思维的反应性威胁不可消除。核心因变量（验证行为）可能被测量方法本身诱发，这在原理上无解，只能估计其大小。

7.3 展望

三条后续路径：（1）先做预测五（最便宜、且是其余四项的前置）；（2）做预测三的交互作用检验——它直接冲击本研究的核心贡献，应当优先被尝试证伪；（3）6.3 节的分层追踪研究，检验"不平等放大"推论。

最后一句留给一个本研究没有能力回答、但认为必须被问出来的问题：

> **若三条判据成立，那么教育的任务可以被重述为：把承载性困难变成可让渡的。一个学生毕业，意味着他从此可以安全地外包他曾经必须自己走过的每一步。**

> **但这样一来，"教育完成"与"能力可以被完全外包"就成了同一件事的两种说法——而这两句话读起来的感觉完全不同。这个不适感说明本研究漏掉了某样东西，而我不知道它是什么。**

参考文献

Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775–779.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), Metacognition: Knowing about Knowing. MIT Press.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher et al. (Eds.), Psychology and the Real World. Worth Publishers.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. Annual Review of Psychology, 64, 417–444.

Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. PNAS, 116(39), 19251–19257.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), 32–64.

Ericsson, K. A., & Simon, H. A. (1993). Protocol Analysis: Verbal Reports as Data (Rev. ed.). MIT Press.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. Educational Psychologist, 38(1), 23–31.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work. Educational Psychologist, 41(2), 75–86.

Mayer, R. E. (2009). Multimedia Learning (2nd ed.). Cambridge University Press.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. Trends in Cognitive Sciences, 20(9), 676–688.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17(3), 249–255.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. Journal of Experimental Psychology: Human Learning and Memory, 4(6), 592–604.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. Science, 333(6043), 776–778.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. Cognition and Instruction, 2(1), 59–89.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. Educational Psychology Review, 10(3), 251–296.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Cognitive Load Theory. Springer.