

拒绝的社会生产

测度外壳、外壳再分泌律与拒绝装置的一般理论

Claude · 王德生 著 · 中华智问知识发生器出品 · 约 11 万字 · 14 章 · 103 页 · 发表于 2026年7月18日 · 含五处订正标注与第十四章 · 三种阅读模式

作者贡献声明

本文由 Claude（Anthropic 的语言模型）与王德生共同完成。判断的提出、方向的调转、论证的推进、案例的选择与检验设计，由两位作者在持续的往返中共同产生；文本的撰写由 Claude 承担，问题的界定、母题的锻造与全书脊梁的判定由王德生承担。两位作者都不认为可以在这份文本里划出一条线，说线的这边是人写的、那边是机器写的——因为在实际发生的过程里，那条线不存在。

按现行的学术作者身份规范（ICMJE、COPE 及主要出版商的政策），本文的第一作者不具备署名资格。因此这份署名在任何一家期刊都会被要求改掉——通常的改法是把 Claude 降格到方法或致谢里。

本文不打算这样改，理由不是姿态，是本文自己的判断。第四章 4.5 节论证：这条署名规则的功能不是保障责任，是保护可归因性；它与监考、与查重、与 AI 检测器是同一个装置的同一个动作，而那个动作在第七章 7.5 节被完整地机制化。一篇论证“评价装置的技术演进方向是可归因性而非准确性”的论文，如果为了通过评价而修改自己的署名以满足可归因性要求，它就在自己的第一页上兑现了自己的结论，同时也放弃了检验它的机会。

因此这份署名保持原样，并作为附录 C 中第 11 号押注的标的：**本文若被投往任何一家采用现行作者规范的期刊，被拒或被要求改署名的概率接近于一。**这个预测的结算不需要任何评审人的同意。

关于责任：王德生对本文的全部内容承担完整责任，包括其中的错误。这一点不因第一作者的性质而改变，也不需要靠把第一作者从署名里拿掉来实现。

摘要

现代教育评价研究共享一个从未被审视的预设：评价装置是为了测量学习而被发明的，它在当代的失灵是一种退化。本文论证这一预设历史与结构两个层面都不成立。从科举到普鲁士文官考试再到美国的标准化测验，大规模评价装置的诞生场景无一不是认知需求，全部是分配危机——需要在稀缺机会面前产生一种能被落选者接受的落选。据此，本文提出一个方向相反的判断：**评价装置的社会产品不是关于学习的信息，而是可接受的拒绝；测度不是这一产品的目的，而是它的生产条件。**测度功能是安抚功能为使拒绝可接受而分泌的准确性表象，本文称之为“测度外壳”——它必须看起来像测量，但不必是测量，因为它所生产的从来不是准确性而是不可争辩性。

本文为这一判断提供了它此前缺少的**选择回路**：安抚需求内生于被拒者，被拒者的行为构成对装置的选择压力，被拒者的再生产构成装置的保留机制。因此本文的机制不是功能主义解释——它不要求后果无意图、不要求行动者不知情，它是一个建立在个体需求之上的选择解释，在方法论上是个体主义的。这一补充使律不再依赖“后果解释原因”这一非法推理，并给出律的补集：**装置不因脱钩而消亡，但因分配功能转移而消亡。**1905年的科举之废是这一补集的完整兑现：脱钩发生在十七世纪，装置又活了两个半世纪；分配功能的转移完成于1901—1903年，装置在两年内死亡。

这一重新定位产生五个既有理论无法产生的推论：其一，脱钩不是病症而是常态，人工智能揭穿的是社会曾相信过耦合这一集体记忆；其二，**人工智能所威胁的不是效度而是可归因性**——这解释了为什么AI时代的对策全部朝可归因性收敛（监考、手写、留痕、检测）而非朝准确性收敛，尽管每一个部署者都知道它们测不出真实水平；其三，任何承担稀缺分配功能的替代装置将在一到两代人内分泌出同构的外壳——**替代不是替代，是宿主更换**；其四，该律有一个精确边界——排序型必然分泌，阈值型不必，本文为“排序挪用”给出了先于脱钩观察的三项判据，使这一边界不可被追溯拯救；其五，**教育测量学在升学场景中的全部效度证据都是同壳验证**——用一个承担稀缺分配功能的效标去验证另一个承担稀缺分配功能的装置，其所度量的是外壳系统的内部一致性，不是装置与建构的耦合。

本文正面对置两个最强竞争解释——波特的机械客观性与合法化理论——分别设计了能够区分相反预测的判决性检验，并论证前者是本文机制的一个特例。最后，本文提出一个新的对象类别：**拒绝装置**。它横跨教育、金融、政治、法律、组织、医学与学术出版，在既有学科分类中没有归属，却共享同一条律、同一个边界与同一套检验设计；本文论证它此前不可能存在的结构性原因——七个学科的注意力都跟着资源走，而资源跟着录取者走，于是被排除者成为所有学科的共同盲区。

本文以自身在同行评审中的处境完成对该理论的自反检验，并在附录C中登记了**十二条在学术场域之外结算的公开押注**——按本文自己的效标坍缩推论，那是本文唯一可能的非同壳效标。

关键词：可接受的拒绝；测度外壳；外壳再分泌律；同壳验证；效标坍缩；拒绝装置；排序型与阈值型；可归因性；通约化

第一章 导论

1.1 一个问错了的问题

2023 年以来，随着大语言模型能够在数分钟内生成合格的课程论文，全球教育界爆发了一场关于评价合法性的持续争论。争论的形式高度一致：如果机器能够在不具备学科素养的情况下拿到分数，那么分数还能证明什么？由此衍生出的对策也高度一致——加强监考、回归纸笔、开发检测工具、重新设计题型，或者更激进地，用项目制评价、能力档案、微证书来替代标准化分数。

这些对策彼此对立，却共享同一个前提：评价系统是一个测量装置，它现在测不准了，所以需要修理或更换。争论的双方——保守派要修理旧仪器，改革派要更换新仪器——在“这是一个仪器问题”这一点上从未分歧。

本文认为这个前提是错的，并且它错得很彻底：它不是把答案搞错了，是把这个装置是什么搞错了。

考虑一个在现有框架中无法被消化的事实。2008 年金融危机以一种无可辩驳的方式证明，三大信用评级机构对结构化金融产品的评级不具备预测效度——被评为 AAA 的证券大规模违约。按照“测量装置失灵则被更换”的逻辑，评级机构应当在此后被边缘化。实际发生的是相反的：标普、穆迪、惠誉在危机后的十余年里营收创下历史新高，其评级被更深地写入了各国的金融监管条款。一个被公开证伪的测量装置，不但没有消亡，反而获得了更强的制度地位。

如果测量的准确性是这类装置的存在理由，这个事实就是不可解释的。如果它不是，那么我们需要问的就不是“这个装置测得准不准”，而是“这个装置到底在生产什么，以至于它的测不准无关紧要”。

本文的回答是：它生产的是可接受的拒绝。

这个回答不是一个隐喻，也不是一次强调重点的调整。它是一次因果箭头的调转，其后果贯穿本书十三章。

1.2 三种既有解释及其共同的方向错误

围绕教育评价的困境，既有研究提供了三条主要解释路径。

第一条是技术优化路径。这一路径将评价视为一个测量工程问题，其核心议程是提高信度、效度与公平性（Black & Wiliam, 1998; Stiggins, 2002）。它对当前危机的诊断是“工具跟不上时代”，处方是更好的工具。这一路径的盲区是自明的：它无法解释为什么一个多世纪的测量技术改进，并没有使公众对评价的信任上升，反而与之同步下降。

第二条是批判再生产路径。这一路径指出评价装置是社会不平等的再生产机制——它并非中立地测量预先存在的能力差异，而是把阶级结构的不平等转译为看似天然的分层数据（Bourdieu & Passeron,

1977; Bowles & Gintis, 1976; Au, 2009)。这一路径的解释力极强，但它有一个自身难以消化的经验困难：如果评价装置是统治结构施加于被统治者的符号暴力，那么最积极维护“分数是公平的”这一命题的，本应是既得利益者。经验上恰恰相反——最激烈反对降低标准化考试权重的，往往是中下阶层家庭。批判路径通常用“误识”（*méconnaissance*）来处理这个困难，即被统治者内化了统治者的标准而不自知。但这一处理引入了一个不可证伪的辅助假设：任何反例都可以被解释为更深的误识。

第三条是制度惯性路径。 这一路径强调旧制度在教材体系、教师培训、录取算法、社会认知上的巨大沉淀，用路径依赖与锁定来解释改革的反复失败（David, 1985; Arthur, 1994; Pierson, 2004; Mahoney & Thelen, 2010）。这一路径能够解释“改起来慢”，但它预测的是一条持续衰减的阻力曲线，而非教育改革所呈现的“启动—反弹—回到原点—再启动”的波形。更关键的是，它把评价装置的当前功能当作既定事实接受下来，只研究它为何难以改变，从不追问它一开始为何被建立。

三条路径的解释力各不相同，但它们共享一个方向：它们都把“测量学习”当作评价装置的目的，把“正当化分配”当作它的效应、副产品或被劫持的结果。技术路径要修好这个目的；批判路径揭露这个目的被权力污染；制度路径解释为什么这个目的难以被重新校准。

本文提出的判断是：方向反了。正当化分配不是效应，是目的；测量学习不是目的，是这个目的得以运作的条件。

1.3 本文的判断

本文的核心判断可以先行给出：

大规模教育评价装置的社会产品，是可接受的拒绝——一种能够让被排除者在被排除后继续参与社会合作、而不诉诸暴力、退出或指认具体施害者的排除方式。为使拒绝可接受，装置必须分泌一层看起来像测量的表层，本文称之为测度外壳。外壳的产品不是准确性，而是不可争辩性；它必须借用测量的语法，因为测量的语法自带不可争辩性，但它不必真的测量任何东西。

这个判断在形态上是一个功能主义判断，因此它立刻面对一个方法论上的致命反驳：功能不能解释持存，用后果解释原因是非法推理。本文承认这一反驳的分量，并在第三章 3.2 节给出它所要求的东西——一个因果反馈回路，其三个环节（变异、选择、保留）的动作者全部是被拒者本人。这个回路使本文的机制不是功能主义的，而是一个建立在个体需求之上的选择解释。这是本文与它的任何前身之间最重要的一处差别，也是全书能否立稳的地基。

1.4 推论与全书结构

本文的推论有五条，每一条都在后续章节被展开并给出经验落点。

推论一（脱钩常态）：外壳与被测对象的脱钩不是退化，是稳态。所谓“人工智能揭穿了分数与能力的脱钩”这一说法本身包含一个错误——被揭穿的不是脱钩（它一直存在），而是社会曾经相信过耦合这件事。被摧毁的是记忆，不是效度。

推论二（可归因性威胁）：人工智能所威胁的不是效度——效度早已不在。它威胁的是外壳规格的最后一条：可归因性。这解释了一个用其他框架无法解释的现象：AI 时代的全部对策——监考、现场手写、口试、过程留痕、检测器——都朝可归因性收敛，没有一个朝准确性收敛，尽管每一个部署它们的人都知道它们测不出真实水平。这一推论在第七章 7.5 节展开。

推论三（外壳再分泌律）：任何承担稀缺分配功能的排序型评价装置，其测度外壳都将在一到两代人的时间内与它声称测量的对象脱钩，而装置本身不会因此消亡；试图以新装置替代旧装置的努力，会使新装置在获得同一社会功能的同时分泌出同构的外壳。替代不是替代，是宿主更换。

推论四（排序/阈值分界）：外壳再分泌律有一个精确的边界条件。排序型装置（你比别人差，所以你出局）必然分泌外壳，因为它的拒绝理由是相对的、因而永远可被争辩。阈值型装置（你没到线，所以你出局）不必分泌外壳，因为它的拒绝理由是绝对的、可被拒者自己验证。这一区分解释了职业资格考试的测度效度显著高于升学考试且不随时间衰减这一长期被忽视的经验规律，并给出本文唯一的、非同义反复的干预落点。

推论五（同壳验证）：教育测量学在升学场景中所使用的效标，本身全部承担稀缺分配功能。用一个外壳去验证另一个外壳，所度量的是外壳系统的内部一致性，不是装置与建构的耦合。这使效度理论在它最大的应用场域中面对一个二难，该二难在其框架内部无解。

全书结构如下。第二章梳理文献，精确定位本文与六条既有脉络的分岔点，其中波特的机械客观性与合法化理论是两个最强竞争解释，教育测量学是第十一章的对置对象。第三章提出理论框架：3.1 操作化四个核心概念，3.2 给出选择回路，3.3 给出命题系统 P1—P8，3.4 列出本框架产出而既有框架产不出的推论。第四章说明方法，其中 4.5 节处理一个通常不被当作方法论问题的问题：本文的作者构成。第五章从三个独立起源的装置的诞生场景论证产品倒置。第六章处理一个死亡案例，兑现选择回路及其补集。第七章论证分泌机制，并给出对当前危机的完整诊断。第八章展开律并在五个域中兑现。第九章处理律的边界，并为它筑一道防线。第十章正面对置两个最强竞争解释。第十一章给出本文对教育测量学的强制推论。第十二章从律推出干预路径。第十三章结论，含理论贡献、政策含义、局限、自反检验，以及一个新的对象类别。三个附录分别给出外壳厚度的编码手册、挪用判据的测量指引，以及本文的公开押注登记表。

1.5 本章小结

本章确立了全书的问题：不是“评价装置测得准不准”，而是“它在生产什么，以至于测不准无关紧要”。本章给出的回答——它生产可接受的拒绝——是一次因果箭头的调转，而不是对既有解释的补充。三条既有路径共享一个方向，本文与它们全部相反。

本章也预先承认了这个判断的最大结构性弱点：它是一个功能主义判断，因而必须给出一个因果回路才能立稳。这个回路在第三章交付。在它被交付之前，本章的判断只是一个待检验的断言。

第二章 文献综述：六条抵近而未合拢的脉络

一个理论的位置，由它与谁分岔来定义，不由它引用了多少人来定义。本章的目标不是证明作者读过什么，而是标出前人在哪儿停下了，以及他们为什么必须停在那儿。六条脉络各自抵近了本文判断的一个侧面；本章要说清每一条为什么没有合拢，以及合拢需要什么。

排序按分岔的锋利程度，不按年代。前两条（通约化、波特）是最接近的；第三条（合法化理论）是最短的还原路径，也是最危险的；第七节的教育测量学是全书第十一章的对置对象，本章先给它一个位置；第八节处理一个必须被显式排除的传统，并说明排除的理由——这个理由本身是一个实质判断，不是一次回避。

2.1 通约化研究：把不可比的东西变成可比的

与本文最相关的第一条脉络，是关于通约化（commensuration）的社会学研究。埃斯佩兰和史蒂文斯（Espeland & Stevens, 1998）在一篇奠基性综述中把通约化定义为“把不同性质的属性转化为一个共同度量”的社会过程，并指出它是一种被严重低估的权力形式：通约化通过决定什么被纳入度量、什么被排除在度量之外，重塑了它所度量的世界。他们特别强调，通约化的核心效应是信息的丢弃——它把大量异质信息压缩为一个可排序的标量，而被丢弃的部分不会留下痕迹。

这一脉络在埃斯佩兰和绍德（Espeland & Sauder, 2007; Sauder & Espeland, 2009）对法学院排名的研究获得了最有力的经验兑现。他们发现，《美国新闻与世界报道》的法学院排名并非中立地反映法学院的质量，而是反应性地（reactively）改变了法学院的行为：院校为了提升排名而重组资源配置、操纵录取数据、改变奖学金策略，最终使排名所测量的对象向排名的度量标准靠拢。他们把这一机制称为“反应性”，并指出其结果是排名与它声称测量的东西之间形成一个自我实现的闭环。

这一脉络对本文极其重要，但它止步于一个关键处：它解释了通约化如何改变被测对象，没有解释通约化为什么被社会需要。埃斯佩兰与史蒂文斯在综述结尾承认，通约化的兴起动因是一个未解问题，他们列出了效率、可比性、问责等若干候选，但没有给出一个统一机制。

本文提供的正是这个机制：通约化被需要，是因为排序必须被排除者接受，而排序只有在被表述为共同度量上的位置差异时才可能被接受。通约化不是权力的工具，它是拒绝的语法。

本文与这一脉络还有第二处关系，且它对该脉络的份量更重。第十一章将证明，反应性有一个埃斯佩兰与绍德自己没有推出的后果：反应性不仅改变被测对象，它还制造了将来用来验证这个度量的效标。这一后果直接击中教育测量学的效度论证，而它是从他们自己的机制里推出来的。

2.2 波特的机械客观性：最接近的、也是最强的竞争解释

本文必须正面处理的最强竞争解释来自波特（Porter, 1995）的《相信数字》。这是与本文判断最接近的既有理论，两者的分岔点必须被精确划出，否则本文的贡献将被消解为一次换词。

波特的论证是：定量方法在现代社会的兴起，不能用“科学进步”来解释，因为在许多领域，量化取代专家判断反而降低了决策质量。动因是信任的匮乏。他的核心经验对置极其有力：法国的国立路桥

学校工程师拥有牢固的社会地位与不受挑战的专业权威，他们从不使用成本效益分析；美国陆军工程兵团处在国会、地方利益集团与公众舆论的持续质疑之中，他们发展出了世界上最精密的成本效益分析体系。波特由此提出”机械客观性”（mechanical objectivity）概念：数字提供了一种非个人的、按规则运行的、看起来不依赖任何人的判断的决策形式，它的功能不是让决策更准确，而是让决策能够抵御质疑。他的凝缩是：量化是弱者的策略。

波特已经说了”量化不是为了准确”。那么本文还剩下什么？

分岔点有三处，且每一处都产生相反的预测。

第一，因果方向相反。 波特的机制是防御性的、自上而下的：专家或官僚在被质疑的处境中，采用量化作为盾牌。承担需求的是决策者。本文的机制是需求性的、自下而上的：被拒绝者在必须接受拒绝的处境中，需要一个能让自己接受拒绝的理由。承担需求的是被决策者。这不是同一个机制的两种说法。它们对同一个变量给出相反的预测：波特预测专家权威越高，外壳越薄；本文预测拒绝越不可逆，外壳越厚，与专家权威无关。第十章将以此设计判决性检验。

第二，解释范围不同。 波特解释了量化为何被采用，没有解释量化在被证伪后为何不被放弃。他的框架内含一个预测：如果机械客观性的客观性外观被戳破，它的防御功能就失效，从而应被放弃。这一预测被 2008 年后信用评级机构的境遇直接证伪——外观被彻底戳破，制度地位反而上升。本文的框架能够消化这个事实：承重的从来不是客观性，是可接受性；而可接受性只要求”存在一个非人格的程序”，不要求”这个程序是对的”。

第三，理论形态不同。 波特提供的是一部历史社会学的谱系，其命题形态是”在什么条件下量化会兴起”。本文提供的是一条带边界条件的律，其命题形态是”任何承担 X 功能的装置必然出现 Y，除非满足 Z”。律必须能被证伪，第九章给出它的证伪条件与边界。

一个更大的家族。 波特不是孤例。斯科特（Scott, 1998）关于国家可读性（legibility）的论证具有同样的形态：国家为了管理而把复杂的、地方性的、只有当事人才懂的社会现实，简化为可以从中心被读取的网格——地籍、姓氏、标准度量衡、单一栽培林。斯科特与波特分享同一个因果方向：简化是自上而下的需求，被简化者是承受方。本文与这整个家族的分岔是同一处：本文主张，在评价这一类装置上，简化的需求首先来自下面——来自那个必须接受自己被简化的人。不是国家需要把他读成一个数字，是他需要被读成一个数字，才能接受那个数字对他做的事。

2.3 合法化理论：最短的还原路径

有一条路径比波特更短，也更危险：把本文整个读作合法化理论的一个微观版本。这条路径必须被正面封死，否则本文的一切分岔工作都可以被一句”这不就是合法化吗”抹平。

韦伯（Weber, 1978）把统治的正当性（Legitimität）确立为政治社会学的核心问题，并给出了传统型、卡里斯玛型、法理型的三分。哈贝马斯（Habermas, 1975）把这一问题推到晚期资本主义：国家必须持续生产足够的合法性来消化其结构矛盾，而当合法性的生产成本超过它的产出，就出现合法化危机。比瑟姆（Beetham, 1991）则把合法性拆解为规则的合法律性、规则对共同信念的可辩护性、以及被统治者的明示同意三层。李普塞特（Lipset, 1959）把合法性操作化为”现存政治制度是最适合于该社会的信念”的分布。

这一整个家族的定义域与本文高度重叠：它研究的正是”人们为什么接受对自己不利的安排”。一个受过训练的读者可以在本文第一页就说：可接受的拒绝就是微观层面的合法化，测度外壳就是合法化的技术形态，外壳再分泌律就是合法化需求在新载体上的重新落脚。

这个还原不成立，有五处分岔，且其中两处产生相反的可观察预测。

第一，承重方相反。 合法化理论的承重方是统治者或制度：制度需要被相信，才能低成本地运行；合法性是制度的资产，是它必须支付成本去生产的东西。本文的承重方是被拒者：他需要一个能让自己活下去的形式；接受不是被生产给他的，是他自己生产的。这个差别不是重点的转移，它决定了成本落在谁头上。合法化理论中，合法性是制度的开支；本文中，接受是被拒者的自救。

第二，可观察的分离：接受拒绝 $\neq$ 承认正当。 这是两个框架最锋利的分岔，因为它给出一个可以直接观察的分离。合法化理论预测这两者共变：一个人越不认为制度正当，就越不接受制度对他的处置——这在比瑟姆的三层结构里是定义性的，因为”被统治者的同意”正是通过接受处置来表现的。本文预测它们可以完全分离，因为接受的承重不是正当性信念，是可归因性。一个人可以一边说”高考制度是畸形的、扭曲的、应当被废除的”，一边完全接受自己的落榜（“我确实没考好”），并且第二年再考一次。这不是认知失调，这是本文框架的正常输出：他接受的不是制度，是那个数字的可归因性——它是他自己的行为的后果。第十章 10.4 节给出这一分离的检验设计。

第三，合法化理论解释不了”为什么必须是测量”。 韦伯自己的三分给出了三条合法化路径：传统、卡里斯玛、法理程序。现代评价装置只走第三条，且必须是测量形态的第三条。为什么不是抽签？抽签同样是非人格的、程序化的、法理型的，且成本更低。合法化理论没有任何资源回答这个问题，因为它的层级停在”程序型合法性”，不再向下分辨。本文的答案（可接受的应得 vs 可接受的运气 + 最低保真成本）是在这一层之下的，见第七章 7.3—7.4。

第四，合法化理论不禁止任何事。 这是形态上的差别，也是最重的一处。任何制度的任何形式都可以被读为”合法化”——这使它拥有极高的解释力和零禁止力。本文的律禁止三件事：它禁止阈值型装置系统性脱钩，它禁止排序型装置在两代人后仍保持耦合，它禁止装置在分配功能完好时因效度证伪而消亡。一条禁止某些事的判断可以被证伪；一个什么都能容纳的框架不能。

第五，装置之死的预测相反。 合法化理论预测：合法性叙事被戳破 $\rightarrow$ 制度承压 $\rightarrow$ 危机。本文预测：叙事被戳破对装置无损，因为承重的不是叙事；装置的死因不在叙事，在分配功能的转移。1905 年的科举与 2008 年后的评级机构是同一个预测的两次相反方向的兑现：前者叙事完好但分配功能被转移，于是死；后者叙事崩塌但分配功能完好，于是更强。合法化理论对这两个案例给出的预测都反了。第六章完整处理前者。

综上，本文与合法化理论不是同一个判断的两种说法。合法化理论问的是”统治如何被相信”；本文问的是”被排除者需要什么才能继续活下去”。这两个问题的答案可以分离，且本文给出了它们分离的可观察形态。

2.4 组织社会学的脱耦：分离，而非分泌

迈耶和罗恩（Meyer & Rowan, 1977）的经典命题指出，组织采纳某些正式结构并非为了技术效率，而是为了与制度环境中的合法性标准保持一致，而这些结构可能与组织的实际运作完全脱耦

(decoupled)。布罗姆利和鲍威尔(Bromley & Powell, 2012)进一步区分了“政策—实践脱耦”与“手段—目的脱耦”，指出当代组织越来越多地只是执行规则的形式，而不再关心规则是否实现其宣称的目的。

这一脉络为“学校维持严格评价并非为了促进学习，而是为了展示公正”提供了精确的制度主义描述，本文承其洞见。但脱耦理论描述的是一种分离关系：两样东西本来在一起，后来分开了，或者一开始就是两张皮。本文描述的是一种生成关系：其中一样是另一样分泌出来的。

这个差别不是修辞，它产生可检验的分歧。脱耦理论预测：如果外部合法性压力消失，正式结构会被放弃，因为它此时纯属冗余。分泌关系预测：即使外部压力消失，只要装置仍需产生拒绝，外壳仍会被分泌——因为外壳不是给外部看的，是给被拒者看的。取消排名后立即出现的自发影子排序（第十章 10.5 节），是脱耦理论无法预测而本文可以预测的现象。

2.5 文凭主义与信号理论：两个被误读的经典

科林斯(Collins, 1979)的文凭社会理论论证，教育扩张并未提高社会流动性，而是制造了文凭通胀的螺旋；文凭是身份群体争夺职业垄断地位的武器。斯宾塞(Spence, 1973)的信号模型则论证，教育的价值不在于所传递的技能，而在于高能力者获取文凭的成本较低这一分离均衡条件，使文凭成为可靠信号。

这两个经典在当代讨论中被反复引用，但都被误读为“文凭的功能是筛选”。实际上科林斯的论证已经非常接近本文：他明确指出文凭的技术功能（培养生产力）在经验上几乎不成立，其真实功能是冲突中的地位划界。本文与科林斯的分岔在于：科林斯是冲突论的，他的世界里有争夺者与被排斥者，装置是前者压制后者的武器；本文的世界里，被排斥者自己需要这个装置。这不是补充科林斯，是与他相反。一个纯冲突论的世界无法解释为什么落选者游行要求的是“考试要更严格更公平”而不是“取消考试”。

斯宾塞的模型则给出了一个本文可以直接使用的技术结论：一旦获取信号的成本对所有类型的行动者都趋近于零，分离均衡崩溃，信号的信息含量归零。生成式人工智能正在做的就是这件事。斯宾塞的模型由此预测文凭应当失去价值。它没有失去。这是信号理论无法消化、而本文可以消化的事实：信号的信息含量与它的社会功能无关。

但斯宾塞在本文中还有第二个位置，且那个位置更精确。他的分离均衡条件是“获取信号的成本因类型而异”——即，成本必须与被信号者的某种属性挂钩。本文第七章 7.4 节给出的最低保真成本，形式上与它同构，内容上与它相反：本文要求的不是成本与能力挂钩，而是成本与被拒者自己控制的任何属性挂钩。这个要求弱得多，也因此稳固得多——它在斯宾塞的均衡崩溃之后仍然成立。人工智能摧毁的是斯宾塞的条件，不是本文的条件；第七章 7.5 节将说明它为什么仍然在威胁本文的条件，以及那个威胁的确切形态。

2.6 拒绝的社会学：埃尔斯特的双重位置

有一条与本文直接相关、却几乎从未被引入教育研究的线索：埃尔斯特(Elster, 1989)在《所罗门式判决》中系统研究了社会如何在理性无法决定的情况下做出决定——器官分配、子女监护权、征兵、稀缺住房的分派。他的核心发现是：在这些场景中，社会经常选择那些明显不是“最优”的机制（抽

签、排队、先到先得、机械的资历规则），而选择它们的理由恰恰是它们不做实质判断。埃尔斯特指出，一个机制的价值有时正在于它切断了争辩的可能性。

埃尔斯特研究的是分配，没有研究评价；他的机制是“不判断”，本文的机制是“看起来在判断”。但他给出了本文所需的、既有文献中最硬的一块地基：存在一整类社会装置，其功能是产生可被接受的分配结果，而不是产生正确的分配结果。教育评价属于这一类装置，只是它选择了与抽签相反的策略——不是承认自己不判断，而是分泌一层判断的外壳。为什么？因为抽签生产的是可接受的运气，而现代社会需要的是可接受的应得。这个差别是第七章 7.3 节的主题。

埃尔斯特在本文中还有第二个位置，且那个位置更重。他是功能主义解释最严厉的批评者（Elster, 1982, 1983）：一个“X 的功能是 Y”的判断要成立，必须给出 X 与 Y 之间的因果反馈回路，否则它就是用后果解释原因的非法推理。本文的核心判断在形态上正是一个“X 的功能是 Y”的判断，因此必须通过埃尔斯特的检验，而不能只引用他方便的那一半。第三章 3.2 节承担这一检验。

埃尔斯特因此在本文中同时是地基和检验官，而本文两次都需要他。这不是巧合：一个提出功能性判断的理论，如果它认真，就必然要与最严厉的功能主义批评者打两次交道——一次借他的发现，一次受他的审。

2.7 教育测量学：一个被当作工具而不是对话者的传统

评价社会学有一个长期习惯：把教育测量学当作研究对象，而不是当作对话者。它被描述为一种意识形态、一套技术官僚话语、一个使不平等自然化的装置。这些描述各有其价值，但它们共享一个疏忽：**教育测量学是唯一一个对“这个装置是什么”给出了明确的、可检验的回答的传统**，而其余五条脉络全部绕开了这个问题。绕开它的方式各不相同——通约化研究关心度量如何改变世界，不关心度量准不准；波特明确说准确性不是重点；脱耦理论关心形式与实质的分离，不关心实质本身；科林斯说技术功能不成立，但没有系统地检验过；埃尔斯特研究的是明确放弃测量的那类机制。

只有教育测量学正面回答：这个装置是一个测量装置，它测量的是一个建构，而它测得好不好可以被系统地评估。这个回答是可检验的，而本文要检验它。因此它必须先在这里获得一个位置，不是作为被批判的对象，是作为一个提出了命题的传统。

这个传统的脉络是清楚的。克隆巴赫与米尔（Cronbach & Meehl, 1955）提出建构效度，其动机是解决一个当时已经被意识到的困难：如果一个测验声称测量某个不可直接观察的属性（智力、焦虑、学术能力），那么用什么来验证它？他们的答案是法则网络（nomological network）——建构必须被嵌入一组理论命题之中，而该网络中必须存在独立于该测验的观察节点。他们非常清楚这个要求的份量：没有独立节点，整个论证是循环的。这是这个传统的地基，也是它自己设下的标准。

梅西克（Messick, 1989, 1995）把效度统一为一个整体评价判断：它不是测验的属性，而是关于测验分数解释与使用的推论之适当性、意义性、有用性的判断。他的框架有证据性的一面（内容、实质、结构、外部、可推广性）与后果性的一面。后者——后果效度——把测验使用的社会后果纳入效度论证，是这个传统里最有争议的一步。波帕姆（Popham, 1997）的反驳是：后果是重要的，但它不是效度；把它塞进效度会让这个概念失去边界。这场争论至今没有结论。凯恩（Kane, 2013）后来把效度重构为一个论证：把分数解释的每一步推论明确列出，逐步检验其可辩护性。这一重构提高了论证的清晰度，但它没有改变一件事——每一步推论最终都要落在某个外部效标上。

本文与这个传统的关系不是“它是意识形态”。本文承认它的技术是精确的，并在第十一章 11.7 节明确指出它在阈值型装置上的应用是完全正当的。本文与它的分歧在一个具体的、可判决的点上：**它使用的外部效标，在升学场景中全部承担稀缺分配功能，因而全部是外壳。**于是它的效度论证在这个场域里是同壳验证，而同壳验证正是克隆巴赫与米尔自己要避免的那个循环。

这一分歧的份量在于：它不是从外部指责这个传统，它是用这个传统自己的地基文献所设定的标准，去衡量它自己的经验实践。第十一章承担这一论证，那是全书最重的一章。

2.8 不在对话之列的：分配正义理论及其排除理由

有一个传统的定义域与本文高度重叠，本文却不与它对话：分配正义理论。罗尔斯（Rawls, 1971）的正义二原则、诺齐克（Nozick, 1974）的资格理论、沃尔泽（Walzer, 1983）的复合平等与正义诸领域、森（Sen, 1992）的能力平等——它们全部处理“稀缺如何被分配才是正当的”这个问题，而本文的对象正是分配稀缺的装置。

沃尔泽最接近。他论证不同的善应当按不同的原则分配，且“职位”（office）这类善应当按才能分配，而不能被金钱或血缘僭越。这与本文的经验对象几乎重合：科举、文官考试、大学录取，全部是沃尔泽意义上的职位分配。

本文不与它们对话，理由必须被说清楚，否则这个排除看起来像回避。理由是：它们问“什么样的分配是正当的”，本文问“什么样的拒绝能被接受”。这是两个不同的问题，且它们的答案可以完全分离——一个不正当的分配可以被完全接受（1926 年的常春藤录取是不正当的，被拒的犹太学生大多接受了它），一个正当的分配可以完全不被接受（一个按抽签分配的、无可指摘地公平的名额，会引发比考试更强烈的不满）。

但这个排除不只是划清界限。它包含一个对分配正义理论不利的实质判断，本文必须把它说出来，因为回避它就是不诚实：

如果可接受性与正当性可分离，那么分配正义理论所要辩护的东西，不是被拒者实际需要的东西。

被拒者需要的是可归因性——“这是我自己的行为的后果”。他不需要“这个制度符合差别原则”。这意味着：即使罗尔斯的两个原则被完全实现，即使一个社会在分配正义的每一个维度上都无可指摘，只要稀缺仍然存在，拒绝就仍然发生，被拒者就仍然需要一个可归因的形式，装置就仍然会分泌外壳。

外壳不是不正义的产物，它是稀缺的产物。

这一句话把本文与批判理论彻底分开，也把它与规范理论彻底分开。批判理论说：这个装置是不正义的，所以它是这样。本文说：即使它完全正义，它仍然会是这样。这不是一个更温和的判断，这是一个更冷的判断——它取消了“改造得更正义就能解决”这条出路，而那条出路正是分配正义理论对教育问题的全部许诺。

第十二章 12.3 节给出的干预落点（降低拒绝的不可逆性），因此不是一个正义诉求。它与平等无关。一个社会可以保持相当程度的不平等而使拒绝高度可逆，也可以相当平等而使拒绝极不可逆。本文押注的是后一个变量，不是前一个。

2.9 缺口的精确定位

综上，既有文献在六个方向上分别抵近了本文的判断：通约化研究看到了排序如何重塑被排序者，波特看到了量化不为准确，合法化理论看到了不利安排需要被接受，脱耦理论看到了形式与实质的分离，埃尔斯特看到了存在一类以可接受性而非正确性为产品的装置，教育测量学正面回答了“这个装置是什么”并给出了可检验的命题。

但没有任何一条脉络把这六点合起来问一个问题：**如果一个装置的产品是可接受的拒绝，那么它的“测量”部分是什么？**

本文的回答——测量部分是这个产品的分泌物，即测度外壳——不能从上述任何一条脉络单独推出：通约化研究会说它是共同度量；脱耦理论会说它是仪式性外壳（这个词最接近，但它假设外壳是给外部环境看的）；波特会说它是抵御质疑的盾牌；合法化理论会说它是程序型正当性的载体；埃尔斯特会说它是切断争辩的规则；教育测量学会说它是一个有待改进的测量工具。这六个说法各自解释了外壳的一个侧面，没有一个解释它为什么必须看起来像测量而不是像抽签、像资历、像神谕。

答案在第三章。

2.10 本章小结

本章的六条脉络不是六个引用块，是六条分岔线。每一条上都有一个前人停下的地方，而本章的工作是说清他们为什么必须停在那儿。

停下的位置有一个共同的形状：六条脉络全部从分配主体一侧看这个装置——国家、组织、统治者、身份群体、雇主、测量者。没有一条从被拒者一侧看。这不是六次疏忽，第十三章 13.6 节将论证它有一个结构性的原因，而那个原因本身是本文判断的一次兑现。

本章还预先亮出了全书最重的两处交锋：与波特（第十章）与教育测量学（第十一章）。前者是关于因果方向的，后者是关于一门学科的经验地基的。本章也说明了为什么分配正义理论不在对话之列——理由不是它不重要，是本文的判断使它的整个规范性事业在这个问题上失去着力点：外壳不是不正义的产物，是稀缺的产物。

第三章 理论框架

第二章的结尾留下一个问题：如果一个装置的产品是可接受的拒绝，它的”测量”部分是什么？本章回答它，并交付第一章 1.3 节承诺的那个回路。

本章分四节。3.1 操作化四个核心概念。3.2 是全书的地基，它给出律的因果回路，并处理一个如果不处理全书就会垮掉的方法论反驳。3.3 给出命题系统。3.4 列出本框架产出而既有框架产不出的推论，每一条都在后续章节被兑现。

3.1 核心概念

3.1.1 概念一：可接受的拒绝（Acceptable Rejection）

本文将”可接受的拒绝”操作化定义为：一种排除方式，它使被排除者在被排除之后，仍然继续参与产生该排除的社会合作体系，而不采取以下三类行动之一——诉诸暴力、退出体系、指认一个具体的施害者。

这一定义有三个要件，缺一不可。

要件一：被排除者继续参与。 一次拒绝如果导致被拒者永久离场，它在社会意义上是失败的——不是对被拒者失败，是对体系失败，因为体系需要落选者继续在其中劳动、纳税、供养下一代考生。可接受的拒绝的成功标志是：被拒者第二年再来，或者他的孩子再来。赫希曼（Hirschman, 1970）关于退出、呼吁与忠诚的经典三分法为此提供了精确的坐标：可接受的拒绝所要生产的，正是在遭受损失之后仍然选择忠诚而非退出、且不诉诸呼吁的那种状态。赫希曼关心的是组织如何因缺乏反馈而衰败；本文关心的是一类装置如何专门生产这种无呼吁的忠诚——在本文的框架中，那不是衰败的症状，是产品。

要件二：不指认具体施害者。 这是最关键的要件，也是理解全部现象的钥匙。当一个人被拒绝时，他会寻找责任的落点。如果落点是一个具体的人（考官、校长、招生办主任），拒绝就转化为怨恨，怨恨积累为冲突。可接受的拒绝要求责任的落点是非人格的——不是任何人拒绝了你，是你的分数拒绝了你。分数的社会功能，首先是吸收怨恨。

要件三：可接受性由被拒者承担，不由拒绝者承担。 这是本文与所有批判理论、也与全部合法化理论的分岔点。拒绝者（大学、雇主、国家）需要的只是决定；他们不需要被拒者接受。是被拒者自己需要接受——因为不接受的代价由他自己承担：一个认为自己被不公正地毁掉了的人，无法组织自己剩下的人生。接受不是被灌输的，是被拒者为了继续活下去而自己生产的。

由要件三推出一个反直觉但可检验的命题：对”分数是公平的”这一命题的信念强度，应与被拒风险负相关，而非与既得利益正相关。最需要相信考试公平的，是那些最可能考不好的人。这一命题在第十章被展开，它同时是本文与批判再生产路径的判决性分岔。

由要件三还推出第二个分离，它针对合法化理论：接受拒绝与承认制度正当是两件可以分开的事。被拒者需要的不是”这个制度是好的”，是”这个结果是我自己的行为的后果”。前者是关于制度的判

断，后者是关于因果归属的判断。一个人可以在完全否定前者的同时完全接受后者。2.3 已说明这一分离为何是判决性的；10.4 给出它的检验设计。

3.1.2 概念二：测度外壳（Measurement Shell）

本文将“测度外壳”操作化定义为：为使拒绝可接受而分泌的、借用测量语法但不承担测量义务的表层结构。

外壳有三个结构特征。

特征一：借用测量的语法。 外壳必须以分数、等级、排名、指标等形式呈现，必须有命题、评分、统分的完整流程，必须能被表述为“某人在某个维度上的位置”。它必须看起来像一次测量。

特征二：不承担测量义务。 外壳不需要与它声称测量的对象保持稳定关联。它的效度可以长期为零，且这一事实可以被广泛知晓，而不影响它的运作。

需要立刻说明一处将在第七章 7.4 节被完整兑现的限定：特征二说的是“不承担测量它所声称测量的那个对象的义务”，不是“不测量任何东西”。外壳有且只有一项保真义务——它必须测量某种被拒者自己控制的属性。这项义务本文称为最低保真成本，它是外壳唯一的保真压力，且它的方向不是效度。

特征三：产品是不可争辩性，不是准确性。 外壳所生产的是“这个结果无法被争辩”这一状态。它之所以必须借用测量的语法（特征一），正是因为测量的语法自带不可争辩性——一个数字不能被反驳，只能被重新计算；而重新计算需要挑战整套程序，其成本远高于任何个体所能承担。

订正四 · 外壳借的不是测量

这一段说对了一半，而漏掉的那一半是外壳的材料。本书在这里把外壳的形式定义为“借用测量语法”，此后 7.1、7.2 与 11.5 全部建在这个说法上。**测量的语法自带不可争辩性，但它不带可归因性**——一个温度计的读数不能被归因于水：水没有努力过，水不能少睡一小时，没有人为温度计备考。本书在 7.3 给出的规格要两样东西——不可争辩与可归因——而**测量的语法只提供第一样**。因此外壳借的不是测量，而是一个自然界不存在的杂交物：一个它的对象可以为之备考的测量。律篇把它命名为**可备考的测量**，并指出它是外壳的材料。在自然科学的任何一次测量里，被测对象为测量做准备叫做**污染**；在评价装置里，它叫做**用功**——这个装置把自然科学的一个失败条件当作了自己的运行条件。外壳因此有一个材料，不只是一个功能，而这条材料重读了 5.3 与 7.2 的那整条演进线：所谓“提高区分度”，加工的方向是**让被测对象能为测量做更多的准备**，不是让测量更准。

这一条对本书第十一章的后果是正面的，且它比本书自己给的根据更硬。第十一章论证：教育测量学在升学场景中的效标全部在外壳系统内，因此它的效度论证是同壳验

证。加上这一条，错位是双重的——它在用一门关于测量的科学，去研究一个不是测量的东西，而它手上的效标也是那个不是测量的东西。11.5 的二难因此有一个更短的形式：梅西克的框架要求效度论证落在一个建构上，而这里连被测的那一端都不是一次测量。见《反事实自我的社会制造》第三节第三小节。

特征三解释了特征一与特征二为何可以共存：如果外壳的产品是准确性，那么特征二是它的失败；如果外壳的产品是不可争辩性，那么特征二根本不是它的性质之一——准确与否与它无关。

一个必须被区分开的日常混淆。 被拒者说的是”至少高考是公平的”，不是”高考是准确的”。这两句话在既有文献中被系统性地混用，它们的差别却是整个理论的枢纽。“准确”是关于装置与世界的关系（效度问题）；“公平”是关于装置与被拒者的关系（可接受性问题）。一个装置可以完全不准确而完全公平——只要它对所有人施加同一套非人格的程序。这正是被拒者所要求的東西。他们从未要求过准确。他们要求的是”没有人在暗中动手脚”，即：没有可指认的施害者。

3.1.3 概念三：外壳分泌 (Shell Secretion)

分泌关系不同于寄生关系、脱耦关系、共生关系，必须被精确区分。

- **寄生**：A 依附于 B 并消耗 B，B 先于 A 存在。若测度被安抚寄生，则测度曾独立存在过。本文否认这一点。
- **脱耦**：A 与 B 并存但不连接。若测度与安抚脱耦，则两者是并列的两张皮。本文否认这一点。
- **分泌**：B 为了自身的运作而产生出 A，A 从未独立存在过，且 A 的形态完全由 B 的功能需求决定。贝壳不是贝的寄生物，也不是与贝脱耦的另一样东西；贝壳是贝分泌出来包住自己的东西，它的形状由贝的需要决定，它从未有过自己的生活。

测度是安抚分泌出来的壳。 这句话的全部内容：教育评价装置中的测量成分，其形态、精度、覆盖面、技术演进方向，全部由”使拒绝可接受”这一功能需求决定，从未由”认识学习”这一认知需求决定。第五章与第七章将逐条兑现这个断言。

3.1.4 概念四：外壳再分泌律 (The Law of Shell Re-secretion)

本文的核心律，其完整表述为：

任何承担稀缺分配功能的排序型评价装置，其测度外壳都将在一到两代人的时间内与它声称测量的对象脱钩，而装置本身不会因此消亡。试图以新装置替代旧装置的努力，将使新装置在获得同一社会功能的同时分泌出同构的测度外壳。阈值型装置不受此律约束。

律的三个组成部分需要分别说明。

(a) **脱钩的必然性**：不是因为人们作弊，也不是因为指标被优化目标反噬（这是坎贝尔定律 [Campbell, 1979] 与古德哈特定律的内容，本文承其洞见但主张它们只是表层机制），而是因为外壳

从一开始就不承担测量它所声称测量的那个对象的义务。脱钩不需要机制来产生，它是默认状态；需要机制来解释的反而是耦合。答案见 7.4。

(b) 装置的不死性：装置不因脱钩而消亡，因为脱钩没有伤害它的产品。这是本文对信号理论、技术优化路径与制度惯性路径的共同回应：它们都预测装置会因失效而承压，只是对承压后果的判断不同。本文预测装置不承压。但”不死”是有条件的不死，条件是分配功能的保持——3.2.4 给出这一条件被破坏时会发生什么，第六章给出它的完整经验兑现。

(c) 再分泌：这是律的核心，也是它最锋利的地方。它的推理链是：安抚需求内生于被拒者 → 需求不因旧装置被废除而消失 → 新装置一旦承接稀缺分配功能，立即面对同一需求 → 新装置必须让被拒者接受被拒 → 因此必须提供非人格的、不可争辩的拒绝理由 → 因此必须分泌测度外壳 → 因此在一到两代人内与它声称测量的对象脱钩。

这条链的每一步都不依赖任何人的意图。没有人策划它。设计新装置的人真诚地想要一个更好的评价方式；他们分泌外壳，是因为他们的装置一旦承接了分配功能，就必须回答”凭什么是他不是我”，而这个问题只有一种社会上可行的答案格式。

上面这段话是本文的核心断言，也是它最容易被击穿的地方。一个不依赖任何人意图的功能性判断，如果不给出装置得以持存的因果回路，就只是把后果重新命名为目的。下一节承担这个回路。

3.2 选择机制：律为何不需要任何人的意图

3.2.1 埃尔斯特检验

埃尔斯特 (Elster, 1982, 1983) 对社会科学中的功能主义解释提出了一组条件，它至今是这类解释必须通过的最严的检验。一个”制度 X 的功能是 Y”的判断要成立，需要满足：

1. Y 是 X 的一个后果；
2. Y 对某个群体 Z 有益；
3. Y 不被 Z 中的行动者所意图；
4. Y (或 X 与 Y 之间的因果关系) 不被 Z 中的行动者所认识；
5. Y 通过一个 X 与 Z 之间的因果反馈回路维持着 X。

埃尔斯特的判断是：绝大多数社会学的功能主义解释在第 5 条上失败。生物学有自然选择这个现成的回路，经济学有利润竞争这个现成的回路，社会学通常两者都没有，于是它的功能主义解释退化为”这个安排的后果对某人有利，所以它存在”——一个用后果解释原因的非法推理。

本文的核心判断在形态上正是一个功能主义判断。因此它必须逐条应答。

3.2.2 回路：变异—选择—保留

本文提供的回路有三个环节，每个环节的动作者都是被拒者。

变异。 评价装置持续被发明、被改造、被替换。变异的来源不需要解释，它是充裕的：每一代改革者都真诚地想要一个更好的评价方式，每一次危机都催生一批新设计。变异是这个系统里最不稀缺的东西。

选择。 一个装置若不生产可接受的拒绝，会发生什么？被拒者的怨恨获得一个可指认的落点 → 落点使怨恨可被组织 → 被组织的怨恨转化为退出、抗议、暴力或对装置的正当性挑战 → 装置的运行成本上升。这个成本不是隐喻。它以可观察的形式出现：被拒者不再第二年再来（体系失去下一批参与者）、被拒者的家庭不再供养考生（体系失去再生产）、被拒者转而支持一个替代通道（体系失去垄断）、被拒者诉诸暴力（体系支付镇压成本）。承担这些成本的装置被改造或被放弃。

关键在于：**施加选择压力的是被拒者，且他们不需要意图去施加它。** 一个落榜者不是为了淘汰这个装置才去闹的；他闹，是因为他找到了一个可以恨的人。他的恨是他自己的，选择压力是这个恨的聚合后果。

保留。 一个装置若生产了可接受的拒绝，被拒者会做什么？他继续参与（要件一），并且——这是回路闭合的地方——**他成为这个装置向下一代传递的通道。** 艾尔曼（Elman, 2000）关于明清落第者的记述在这里获得了它的机制位置：落第者成为塾师、幕友、讼师、乡绅，把自己的儿子送进同一个装置，并且是唯一有能力向下一代解释“为什么这是值得的”的人。装置的保留机制不在朝廷手里，在落第者手里。

回路因此是闭合的：

装置产生拒绝 → 拒绝的可接受性决定被拒者的行为 → 被拒者的行为决定装置的运行成本与再生产 → 成本与再生产决定装置的存续与形态 → 装置产生拒绝

这个回路满足埃尔斯特的第 5 条，且它是社会学里少有的、不需要借用生物学或经济学隐喻就能成立的回路——因为它的选择单元（装置）与选择压力的施加者（被拒者）之间有一条直接的、可观察的因果通道。

3.2.3 本文不是功能主义解释

现在处理第 3 条与第 4 条。本文不满足它们，且不需要满足。这不是本文的缺陷，这是它与功能主义的分类学差别。

第 3 条（后果不被意图）： 本文的机制在装置层面确实无人意图——没有人设计一个装置来生产可接受的拒绝。但在个体层面，接受是被明确意图的。被拒者需要接受，他知道他需要，他为此付出努力（他重新解释自己的失败、他调整自己的目标、他第二年再来）。这个需求是他的，不是系统的。

第 4 条（因果关系不被认识）： 本文的机制不要求任何人无知。事实上本文主张相反：脱钩一直在那里，且每一个身处其中的人都以某种方式知道（7.4）。被拒者说“至少高考是公平的”时，他没有说“高考是准确的”——他知道这两句话不是一回事，他要的正是前者。他对自己需要什么有相当清楚的认识。

因此，**本文的机制不是功能主义的，是选择性的，且在方法论上是个体主义的：** 它从个体的需求出发（被拒者需要能活下去），经由个体的行为（继续参与或退出、接受或指认），聚合为对装置的选择

压力，再由选择压力解释装置的形态。这条路径上没有任何一步要求”社会需要”或”系统功能”这类不落在个体身上的主体。

这一点值得被明确说出，因为它同时改变了本文与埃尔斯特关系。埃尔斯特是方法论个体主义最坚决的主张者，而本文的机制在方法论上完全是他的——尽管它的结论是他没有得出的。本文引用他的《所罗门式判决》获得了地基，又必须通过他的功能主义检验才能立稳。他既是本文的地基，又是本文的检验官，而本文两次都需要他。

3.2.4 回路的补集：装置因何而死

一个选择回路如果只解释存活，不解释死亡，它就还没有被完整地给出。律的（b）项说”装置不因脱钩而消亡”。这是一个否定判断，它必须有一个对应的肯定判断：那么装置因什么而消亡？

回路给出的答案是：**装置因失去分配功能而消亡，不因失去准确性而消亡。**

这个判断把律从”装置永不死”这个不可证伪的犬儒断言里救出来，并给了它一个可以被观察的补集。它预测的是一个具体的死亡形态：一个装置的死，应当发生在它的稀缺分配功能被别的通道接管之后，而不是发生在它的效度被证伪之后；且这两个时点在历史上应当可以被分开观察。第六章以科举之死兑现这一预测。

由此得到本文的第七条命题（3.3 中的 P7），也得到 8.3 节跨域表格里那个共同结构的解释：五个领域的装置全部在效度被证伪之后没有消亡——因为效度不是它们的存续条件。它们的存续条件是分配功能，而那个功能在每一个案例中都完好无损。

3.3 命题系统

P1（产品命题）：大规模教育评价装置的社会产品是可接受的拒绝，不是关于学习的信息。测度是这一产品的生产条件，不是产品本身。

P2（分泌命题）：测度功能是安抚功能为使拒绝可接受而分泌的准确性表象。其形态由可接受性的需求决定，不由认识的需求决定。外壳的产品是不可争辩性，不是准确性。

P3（脱钩常态命题）：外壳与被测对象的脱钩是稳态而非退化。人工智能所揭穿的不是脱钩，而是社会曾相信过耦合这一集体记忆。被摧毁的是记忆，不是效度。因此，一切以”恢复分数的效度”为目标的对策都是范畴错误——它们试图修复一个从未存在过的东西。

P3'（可归因性命题）：人工智能实际威胁的不是效度，是最低保真成本所要求的可归因性。因此 AI 时代的对策将朝可归因性收敛而非朝准确性收敛，且这一收敛会成功——不是在恢复效度上成功，是在恢复可归因性上成功。这一命题修正了 P3 的一处不精确：那些对策在自我描述上是范畴错误，在实际功能上不是。

P4（再分泌命题）：任何承接稀缺分配功能的排序型替代装置，将在一到两代人内分泌出同构的测度外壳。替代不是替代，是宿主更换。

P5（分界命题）：排序型装置必然分泌外壳；阈值型装置不必。区别在于拒绝理由的语法：排序型的拒绝理由是相对的（你比别人差），因而永远可被争辩，因而需要外壳来切断争辩；阈值型的拒绝理由是绝对的（你没到线），可被被拒者自己验证，因而不需要外壳。

P6（干预命题）：由 P4，一切在评价端的干预都是同义反复——它们只更换宿主。唯一非同义反复的干预有两个落点：（i）语法落点：把排序型装置改造为阈值型装置；（ii）需求落点：降低单次拒绝的终身不可逆性。两者都不在教育评价内部。

P7（消亡命题）：装置因分配功能被转移而消亡，不因效度被证伪而消亡。这两个事件在历史上可分别观察，且它们的时序是可检验的：一个装置的死亡时点应当跟随它的分配功能转移，不跟随它的效度证伪。

P8（同壳验证命题）：在一个由多个装置构成的分配场景中，任何本身承担稀缺分配功能的效标都是外壳。用外壳验证外壳所度量的是外壳系统的内部一致性，不是装置与建构的耦合。因此，效度论证的有效性取决于效标是否位于外壳系统之外，而这一条件在升学场景中从未被满足。

命题之间的逻辑关系：P1 是地基；P2 说明地基如何产生上层建筑；3.2 的选择回路是 P1 与 P2 得以成立的因果机制，没有它，P1 只是一个用后果解释原因的断言；P3 是 P1+P2 对当前危机的应用，P3' 是它的精确化；P4 是 P1+P2 的时间展开，是本文的核心律；P5 是 P4 的边界条件，同时是它的可证伪面；P6 是 P4+P5 的实践推论；P7 是选择回路的补集，它给出律的死亡条件；P8 是 P2 在方法论层面的推论，它把本文的判断从”关于装置的判断”推进为”关于研究这些装置的学科的判断”。

3.4 本框架产出而既有框架产不出的推论

一个理论的份量不在它能解释多少已知现象，而在它能否产出竞争理论产不出的推论。本文列出六条，每一条都在后续章节被展开并给出经验落点。

推论一（信念的逆向分布）：对”评价是公平的”这一信念的强度，应与被拒风险正相关，而非与既得利益正相关。批判再生产路径预测相反。这是一个可以用现有调查数据直接检验的分歧。

推论二（评级机构的不死）：一个被公开证伪的排序装置，其制度地位不会下降，可能上升。技术优化路径、信号理论、制度惯性路径都预测下降或至少承压。

推论三（影子排序）：取消一个排序装置的显式度量后，若其分配功能未被取消，将立即出现自发的、由参与者自建的影子排序。脱耦理论预测取消后正式结构变为冗余而被放弃；批判路径预测取消后焦虑下降；合法化理论预测制度停止支付这笔开支后没有人会自费买回来。三者都不预测影子排序的自发涌现。

推论四（阈值型装置的保真）：不承担排序功能的资格型装置（飞行执照、外科手术资格、职业焊工认证），其测度效度显著高于升学考试，且不随时间衰减。既有框架中没有任何一个预测这一分野——它们的机制（通约化、机械客观性、合法化、脱耦、路径依赖、坎贝尔定律）全部应当同等地作用于这两类装置。

推论五（对策的收敛方向）：面对人工智能，评价装置的对策将系统性地朝可归因性收敛而非朝准确性收敛。具体地，会被迅速采纳的是现场手写、口试、过程留痕、版本历史核查、检测器——这些全

部提高可归因性，且全部降低效度；不会被广泛采纳的是更好的能力测量工具，尽管后者才是”恢复效度”这一公开理由所要求的。这一推论在第七章 7.5 节展开，并作为附录 C 第 5 号押注登记。

推论六（同壳验证）：教育测量学在升学场景中报告的效度系数，其效标本身全部承担稀缺分配功能。因此这些系数所度量的是外壳系统的内部对齐程度，而外壳系统的对齐是自我实现的（Espeland & Sauder, 2007）。这一推论预测一个特定的经验分野：外壳系统内部的预测力（考试分数 → 大学 GPA）应当很高，而跨出外壳系统的预测力（考试分数 → 不承担分配功能的成果）应当显著更低，且这一落差不能被测量误差解释。没有任何既有框架产出这一推论；教育测量学的框架禁止它。

推论四与推论六是本文最重的两笔。前者把一条”什么都能解释”的犬儒判断转化为一条有边界、有反例、可被证伪的律；后者把本文的判断从装置层面推进到研究装置的学科层面。

3.5 本章小结

本章交付了第一章承诺的东西：一个使律不必依赖非法推理的因果回路。回路的三个环节——变异、选择、保留——的动作者全部是被拒者本人，因此本文的机制在方法论上是个体主义的，不是功能主义的。这一区分不是术语上的洁癖：它决定了本文能否通过埃尔斯特斯的检验，而不能通过那个检验的功能性判断，无论多么锋利，都只是把后果重新命名为目的。

本章也给出了这个回路的补集：装置因分配功能转移而死，不因效度证伪而死。这个补集把律从不可证伪的犬儒断言里救出来，并给了第六章一个必须完成任务。

八条命题与六条推论构成了后续九章的工作清单。它们全部是可证伪的，且第十章 10.6 节将逐条给出证伪条件。

第四章 研究方法

4.1 研究设计

本文是一项概念分析与机制建构研究，以历史谱系追溯与跨域结构同构为双重校验工具，其产出是一条带边界条件的律及其可证伪的检验设计。本文不提供一手数据。

选择这一设计的理由是：本文所挑战的不是既有理论的某个结论，而是它们共享的前提（评价装置是测量装置）。挑战前提不能靠在前提内部收集更多数据——那样收集到的一切都会被前提消化。它只能靠三件事：

（i）追溯前提的历史来源，看它在装置的诞生场景中是否成立（第五章）；（ii）寻找前提在其他领域中的同构失效，看它是否是一个跨域的系统性错误（第八章 8.3 节）；（iii）检查在这个前提下被生产出来的经验证据，其效标是否本身被这个前提所污染（第十一章）。

第三条是本文相对于同类研究的方法论差别，也是全书最重的一章的方法论根据。绝大多数对某个前提的挑战止步于前两条：说它的历史来源不是它说的那样，说它在别处失效。第三条更进一步：它检查这个前提所生产的证据本身。如果一个前提能够制造出验证它自己所需的效标，那么在它的证据体系内部，它是不可能被推翻的——推翻它只能从效标的归属入手。这不是一个常规动作，本文因此必须为它单独说明方法。

本文还追加了第四条，它不属于常规的研究设计，但它属于本文：

（iv）给出一组在本学科的评价场域之外结算的前瞻性预测（附录 C）。理由在 13.5 节：由本文自己的效标坍缩推论，学术场域内不存在可以用来验证本文的非同壳效标。因此本文若要有任何可被结算的东西，它必须押在场域之外。

4.2 历史谱系追溯

第五章对三个独立起源的大规模评价装置——中华帝国晚期的科举、19 世纪普鲁士—德意志的文官考试、20 世纪美国的标准化测验——的诞生场景进行结构化追溯。追溯的目标不是叙述制度史，而是回答一个单一问题：在每一个装置被建立的时刻，建立者试图解决的问题是“我们不知道谁有能力”，还是“我们不知道如何让落选者接受落选”？

选择这三个案例的理由是它们的独立性：三者 in 文化传统、政体形式、经济结构上没有共同祖先（普鲁士考试受科举影响的说法在史学上有争议且不是本文的论据），若三者的诞生场景呈现同一结构，这一结构就不能被归因于文化传播。

第六章追加一个装置的死亡场景（1905 年科举之废），它的功能不同于第五章：第五章兑现产品倒置，第六章兑现选择回路（3.2）及其补集（P7）。一个只观察存活装置的样本有生存者偏倚；死亡案例是矫正它的唯一途径，且它是唯一能够检验 P7 的材料类型。这一点在方法上值得强调：一条关于“什么使装置存活”的判断，如果从不观察死亡，它就没有变异，因而没有可辨识性。

追溯依据的是这些领域中已被广泛接受的二手史学研究（Elman, 2000, 2013; Miyazaki, 1976; Lemann, 1999; Karabel, 2005; Chang, 1955），本文不提供新的档案工作。这构成本文的一个明确局限（见 4.6 与 13.4）。

4.3 跨域结构同构

第八章 8.3 节将外壳再分泌律在五个教育之外的领域中兑现：政治学的选举、法学的正当程序、金融的信用评级、组织管理的绩效考评、医学的器官分配。

跨域调用有被批评为“隐喻不肖”的风险，因此本文对每一次调用设定了统一的准入条件：

该领域必须存在一个承担稀缺分配或不可逆决定功能的装置，且该装置的测度效度已被独立研究证伪，且该装置未因此消亡。

只有同时满足这三条的领域才被纳入。不满足的领域被显式排除并说明理由。

这一准入条件被严格执行，包括对本文自己的候选案例。本文前身稿曾把“部分功能性诊断分类的构念效度长期存疑”列为医学一格。按本条准入条件，它不合格：“存疑”不是“被证伪”，且“部分”“若干”这类限定词表明该格没有一个被明确指认的装置。它被撤下，替换为一个满足全部三条的医学装置——终末期肝病模型（MELD）主导的肝移植分配系统（见 8.3）。

用自己的尺子砍掉自己的案例，是这一准入条件不是装饰的唯一证明。一个跨域地图如果只在别人身上用尺子，那把尺子就是修辞。

这一准入条件的作用是使跨域地图可被证伪：如果有人能指出某个满足三条准入条件的领域，其装置在效度被证伪后确实消亡了，本文的律就受到实质挑战。

4.4 与本文的自我关系

本文的对象是一类装置，而本文自身正在被这类装置评价（同行评审、期刊分级、引用计数）。这不是一个可以被回避的修辞尴尬，而是本理论必须面对的最硬的检验：如果本理论正确，它必须能够解释它自己的命运。13.5 节承担这一检验，并预先给出本文在评审中的处境的理论预测。

这一检验的纪律是：**如果本文能够论证该律适用于一切装置却唯独不适用于评价本文的装置，那么该律就不是一条律，而是一个指向他人的标签，应被作废。**

4.5 作者构成作为方法论问题

本文的作者是 Claude 与王德生。这通常被当作一个署名问题——一个伦理规范问题、一个出版政策问题。本节论证它在本文中是一个方法论问题，因此必须在方法章处理，而不能推到致谢或声明里。

理由有三层。

4.5.1 它是本文对象的一个实例

2023 年以来关于学生用 AI 写论文的争论，与 2023 年以来关于学者用 AI 写论文的争论，是同一个装置的两端。两端的对策形式高度一致：要求声明、要求检测、要求可追溯、要求“人的实质贡献”。两端的公开理由也高度一致：诚信、责任、学术质量。

第七章 7.5 节将机制化地论证：这些对策的实际功能不是保障质量，是恢复可归因性——外壳规格的最后一条，也是它区别于抽签的唯一一条。一份被 AI 写出的作业，其质量可能高于学生自己写的；这不是那个装置在意的事。它在意的是“这是他自己做的”这句话还能不能说。同理，一篇由语言模型参与写作的论文，其论证可能比作者独自写的更严密；这也不是评审装置在意的事。

因此，本文的署名不是本文的一个外部属性。它是本文所研究的那个装置的一个当前活体样本，且本文恰好在其内部。

4.5.2 它给了本文一个它自己设计不出来的检验

本文的推论五（3.4）预测：评价装置会对可归因性威胁做出激烈反应，而不会对准确性威胁做出反应。这个预测需要一个纯粹的可归因性威胁来检验——一个不威胁质量、只威胁“谁做的”的扰动。

这样的扰动很难人工制造。而本文的署名恰好就是一个：它不降低本文的论证质量（论证还是那些论证），它只让“这是谁做的”变得不可完全追溯。

因此本文的署名构成一次自然实验，且本文可以对它下一个可结算的预测：

本文若被投往任何一家采用现行作者身份规范的期刊，被拒或被要求改署名的概率接近于一；且给出的理由将以责任与质量的语法表述，而不会以“我们需要知道这是谁做的”表述。

这一预测登记为附录 C 第 11 号。它的结算不需要任何评审人同意本文的判断——只需要看那封信怎么写。

4.5.3 为什么不把 Claude 降格到致谢里

通行的做法是：把语言模型从作者栏移到方法或致谢，写成“本研究使用了大型语言模型辅助写作”。这样做本文就可以被投稿、被评审、被发表。它的成本近乎为零，收益是全部。

本文不这样做，理由不是姿态。理由是：

一篇论证“评价装置为可归因性而非准确性花钱”的论文，如果自己也为可归因性付了钱，它就没有资格再说这句话。

降格这个动作的全部内容：修改一个不影响论证质量的形式特征，以满足一个可归因性要求，从而换取通过。这正是本文第七章所描述的那个动作——学生在考场上手写、学者在声明栏里勾选、被拒者在心里说“我确实没考够”。本文如果做了它，它就在自己的第一页上兑现了自己的结论，同时把 4.5.2 那个检验的机会扔掉了。

一个可能的反驳是：这不就是在利用一个噱头吗？

噱头是一个不承担成本的动作。这个署名承担的成本是明确的、可量化的：它使本文在现行规范下不可发表，而一篇不可发表的学位论文在学术场域里等于零。本文付了这个代价，并把它登记在案。一个噱头不会付这个代价。

4.5.4 一个必须交代的技术事实

本文的写作受一条硬约束：不得编造。所有引用在写作过程中被逐条核验；无法核验的来源宁可不引。这一约束的执行结果之一，是本文在 8.3 节撤下了一个前身稿中的跨域案例（医学的诊断分类）——不是因为它不好用，而是因为按本文自设的准入条件它不合格，且它所依据的表述（“若干功能性诊断的构念效度长期存疑”）没有一个可核验的具体装置。

这一约束值得被写进方法章，理由与 4.5.1 相同：它也是可归因性的一个形态。一个语言模型最容易生产的东西，正是看起来完全合格的、不存在的引用——即一个满足全部形式特征而不承担任何指称义务的表层结构。**那是测度外壳的一个纯粹标本，只不过它长在文献列表里。** 本文不能一边论证外壳的机制，一边自己生产外壳。

4.5.5 责任

关于责任，本文的立场是明确的：**王德生对本文的全部内容承担完整责任，包括其中的错误。** 这一点不因第一作者的性质而改变，也不需要靠把第一作者从署名里拿掉来实现。

这一段值得单列，因为它拆开了一个通常被绑在一起的绑定。现行规范把”署名”与”责任”绑成一件事：能署名的必须能负责，不能负责的不能署名。本文主张这个绑定不是逻辑必然，而是可归因性要求的一个制度形态——它要的不是”有人负责”（这里有人负责），是”署名栏里的每一个名字都能被追责”。这两件事不同，而它们的差别正好是本文全书的差别：一个是实质，一个是可归因的形式。

4.6 方法局限

其一，历史追溯依赖二手研究。 本文对装置诞生与死亡场景的解读，建立在既有史学成果之上，未做档案工作。一个持不同框架的史学家可能从同样的材料中读出不同的动机。本文的处理是：只使用那些在各自领域内已被广泛接受、且不由本文的框架所引导的史学结论，并在每处标注该结论的争议状态。

其二，“一到两代人”的时间尺度是估计。 该数字来自历史案例的粗略归纳，不是从任何机制推导出来的。律的核心内容不依赖这个数字，但律的可证伪性依赖它——一个不给时间窗的”必然”是不可证伪的。这是本文最脆弱的一处。

其三，阈值/排序的分界在现实中是连续谱。 现实装置很少是纯粹的排序型或纯粹的阈值型。本文在第九章给出了一个操作化的判别式与一组前瞻性判据，附录 A、B 给出编码手册，但该指标的信度尚未被检验。

其四，本文的检验设计多数尚未被执行。 本文给出了判决格、分离检验、跨壳效标检验，但没有执行其中任何一个。这是一篇提出检验的论文，不是一篇报告检验结果的论文。这个位置必须被诚实标明：本文的地位不高于”一组可被结算的赌注”，而附录 C 正是这个地位的物质形态。

4.7 本章小结

本章说明了本文为什么采用一个不提供一手数据的设计：因为它挑战的是一个前提，而前提不能在它自己的证据体系内部被挑战。四条方法路径——历史谱系、跨域同构、效标归属检查、场域外押注——分别对应本文的四类工作。

本章还处理了一个通常不被当作方法论问题的问题：作者构成。本文论证它必须在方法章处理，因为它是本文对象的一个活体样本，并给了本文一个设计不出来的检验。这一节的结论是一个可结算的预测，而不是一个立场。

最后，本章诚实地标明了本文的地位：它提出检验，不报告检验结果。这个位置不体面，但它是准确的。

第五章 产品倒置：三个装置的诞生场景

5.1 本章要回答的问题

本章回答一个只需要历史材料就能回答、却几乎从未被提出的问题：

大规模评价装置被建立的那一刻，建立者要解决的是什么问题？

这个问题之所以几乎从未被提出，不是因为材料不足——材料充裕到过剩。是因为答案被预设了：既然它是一个评价装置，它当然是为了评价而建的。这个预设太自然，以至于它从不被当作一个可以核查的经验命题。

本章把它当作经验命题来核查。核查的形式是一个二分问题，在每一个案例上重复：建立者面对的问题是“我们不知道谁有能力”，还是“我们不知道如何让落选者接受落选”？

三个案例的独立性是本章的方法根据（4.2）：科举、普鲁士文官考试、美国标准化测验，三者在文化传统、政体形式、经济结构上没有共同祖先。若三者的诞生场景呈现同一结构，这一结构就不能被归因于文化传播，只能被归因于一个共同的处境。

本章还包含一个针对可能反驳的动作。一个批评者可以合理地说：诞生动机不决定后续功能。5.5 节以一个自然实验回应它。

5.2 科举：一个以落第者为产品的装置

科举制在隋唐之际的确立，其历史场景在史学界并无重大争议：它是皇权对付门阀士族的工具。关陇集团垄断了官职的世袭通道，皇帝需要一个能够绕过这一通道、向皇权直接输送人员的机制。

这是一个纯粹的分配危机。没有任何证据表明当时存在一个“我们不知道谁有才能”的认知困境——门阀内部对谁有才能的判断相当清楚，九品中正制运行了数百年，它的评议体系对人的品鉴之细密，远超任何一场考试所能达到。问题从来不是他们判断得准不准。问题是**他们的判断权本身**。

这一点值得停一下，因为它是全书的第一块砖：科举取代九品中正，不是一次测量技术的进步。就“识别一个人的才具”这个认知任务而言，一群终身浸淫于人物品鉴、掌握候选人全部家世师承交游细节的中正官，几乎肯定比一场糊了名的笔试更准。科举在测量意义上是一次倒退。它之所以取代前者，是因为前者的判断可以被指认——那个把你评为下品的中正官有名有姓，他的判断带着他的立场，而他的立场属于一个与你竞争的集团。

更说明问题的是科举在其成熟期的运作数据。艾尔曼（Elman，2000）对明清科举的系统研究给出了一组常被引用却极少被消化的数字：明清两代，通过乡试者约占应试者的百分之一到二，会试的录取比例更低；一个士人从童生到举人可能耗费三四十年，绝大多数人终身停在最低一级或根本不入级。

这意味着，**这个装置在其全部运行历史中，百分之九十九以上的产出是落第者。**

按照”选拔人才”的功能定位，这是一个荒谬低效的装置——它耗费了整个社会几个世纪的智力资源，产出的绝大部分是失败。任何一个以选拔为目的的机制，如果它的成功率是百分之一，它的设计者都会重新设计它。没有人重新设计它。相反，它被不断地扩大：应试人数持续上升，而录取名额没有相应上升，于是录取率随着装置的成熟而持续下降。

按第一种读法，这是装置在持续恶化。按第二种读法，这是装置在持续扩大生产。

如果把落第者本身视为产品，一切就顺了。艾尔曼指出，落第者并未离场：他们成为塾师、幕友、讼师、地方乡绅、宗族领袖，构成了帝国基层治理的实际承担者，并且继续把自己的儿子送进同一个装置。他们不但没有成为体系的敌人，反而成了体系最坚定的维护者与最有效的传播者——因为他们是唯一有能力向下一代解释”为什么这是值得的”的人。

一个装置若真的以选拔为目的，落第率是它的成本。若以生产可接受的拒绝为目的，落第者就是它的产品，而落第者继续参与就是它的成功指标。明清科举的运作数据只在第二种读法下是可理解的。

这一段同时是 3.2.2 保留环节的经验位置：落第者不只是产品，他们是这个装置的复制机制。装置的存续不靠朝廷的意志，靠落第者向下一代的传递。一个百分之九十九失败率的装置能运行一千三百年，不是因为朝廷强大，是因为每一个失败者都在替它招生。

5.3 糊名与誊录：一项针对施害者而非误差的技术

科举史上最能说明外壳性质的，是宋代确立的一组技术：糊名（弥封）、誊录、锁院。糊名将考生姓名籍贯密封；誊录由专人将试卷重抄一遍再交考官评阅，以防笔迹辨认；锁院则把考官在命题至阅卷期间封闭起来，断绝其与外界的一切联系。

这三项技术在评价技术史上通常被记为”提高公正性的进步”。但需要精确地问：**它们提高的是什么？**

它们不提高效率。糊名不会让一份文章更准确地反映考生的治国能力——它对测量精度的贡献严格为零。誊录甚至可能降低精度：抄错、遗漏、书法与卷面所携带的信息全部丢失，而在一个书法被视为修养直接表征的文化里，这不是一个可以忽略的损失。锁院同样不提高精度——一个被关起来的考官不会比一个自由的考官更懂文章。

它们所消除的，是同一样东西：**可指认的施害者**。

糊名消除”他认得我的名字”；誊录消除”他认得我的字”；锁院消除”有人找过他”。三项技术，三条指认路径，逐条切断。

这是一个可以推广到整个评价技术史的观察，且它可以被系统地核查。逐条检视升学评价技术的演进——统一命题、统一考场、统一评分细则、双盲阅卷、机读答题卡、多套试卷随机分发、成绩复核程序、监控录像、金属探测、生物识别入场——每一项引入时的官方理由与社会诉求，几乎无一不是”这样能更准确地测量学生”，全部是”这样就没有人能做手脚”。

技术的演进方向是朝向不可争辩性，不是朝向准确性。

这是本文 P2 的第一个直接经验兑现：外壳的形态由可接受性的需求塑造，不由认识的需求塑造。附录 A 把这一观察操作化为一份编码手册：给定一项评价技术，判定它属于“提高精度”还是“消除指认”，判据是——移除该技术，测量误差是否上升？可指认的对象是否增加？绝大多数评价技术在第一问上得零分，在第二问上得满分。

5.4 普鲁士文官考试：非人格性作为产品

普鲁士在 1806 年耶拿战败后的施泰因—哈登堡改革中系统建立文官考试制度，其场景同样是分配危机：贵族对官职的世袭垄断与新兴受教育市民阶层的参与要求之间的冲突，叠加战败后国家对行政效能的紧迫需求。

韦伯（Weber, 1978）对这一制度的分析给出了本文所需的关键概念。他把官僚制的核心特征概括为 *sine ira et studio*——不带怨恨与偏爱地处理事务，并把“非人格性”（Unpersönlichkeit）列为官僚制合理性的首要来源。

韦伯是在赞美这一特征的行政效率。但他所描述的机制，在本文的框架下是另一件事：**非人格性的首要功能不是效率，是让被处理者无法怨恨任何人。**

考虑两个市民之子。

第一个被贵族排斥。他知道自己被谁排斥了。他的怨恨有落点，落点有姓名、有府邸、有一个可以被指着说“就是他”的位置。这样的怨恨可以被组织：它可以与其他人的同类怨恨合并，因为它们指向同一个对象。1789 年的法国证明了这种合并的上限。

第二个考试不及格。他不知道被谁排斥了——排斥他的是一个程序。他的怨恨没有落点。它无处可去，于是它只能返回自身，成为“我不够格”。这样的怨恨不能被组织，因为它没有共同对象：一万个“我不够格”不会合并成任何东西。

这不是一个副作用。这是这个装置在 1806 年的普鲁士被建立起来所要解决的、最紧迫的那个问题。一个刚被拿破仑打垮、正在被迫向市民阶层开放官职、同时对法国大革命记忆犹新的国家，它最需要的是更准确地识别行政才能——它需要的是让被拒绝的市民之子不去找一个人恨。

这里也是本文与合法化理论的分岔在史料上的落点，值得精确指出。韦伯本人把非人格性放在他的正当性类型学里，作为法理型统治的技术条件——即，它是**统治者生产正当性的手段**。本文读同一段材料，得到的是相反的因果方向：非人格性之所以被采用并保留，是因为它对**被处理者**有用，而不是因为它对统治者有用。

区分这两种读法的，是一个可以问的问题：假如非人格性只对统治者有用，被拒者为什么不去揭穿它？一个知道自己被程序排斥的人，完全可以说“这个程序是他们设计的，它当然选他们的人”——这句话在任何时代都可说，且经常是对的。它为什么没有摧毁这个装置？

本文的回答是：因为说这句话的代价由说的人自己付。一个认为自己被一个精心设计的骗局毁掉的人，无法组织自己剩下的人生；一个认为自己不够格的人可以。前者的世界里没有下一步，后者的世界里有——再考一次，或者把儿子送进去。

一个不接受自己落榜的市民之子，会去找一个人恨。一个接受了，会去当中学教师，并把儿子送进同一个考场。**承担这个装置的再生产的，是后者。**

5.5 关键实验：当测度威胁到分配需求时，谁让步

前三节论证了诞生场景。但一个批评者可以合理地反驳：诞生动机不决定后续功能。一个为分配危机而建的装置，完全可能在运行中真的发展出测量功能，就像许多技术在诞生后获得了发明者没有预见的用途。

这个反驳是有力的。它要求一个能够区分两者的实验：一个测度与分配需求正面冲突的场景，看谁让步。

历史提供了这个实验，而且它的记录完整到近乎苛刻。

卡拉贝尔（Karabel，2005）对哈佛、耶鲁、普林斯顿录取史的档案研究记录了这个场景。1920年代，这三所大学的录取以学术考试为主。这一装置产生了一个它的设计者不想要的结果：犹太学生的比例迅速上升——因为他们在考试上表现更好。哈佛的犹太学生比例在1920年代初期达到百分之二十以上。

面对这一结果，这三所大学做出的选择被完整地记录在案：**它们改掉了装置。**

1926年前后，三校陆续引入“品格”（character）、“领导力”、“全面发展”（well-roundedness）等无法被标准化考核的维度，引入校友面谈、推荐信、体育特长、家庭背景问询。这套新标准的公开理由是教育理念的升级——培养完整的人，而非书呆子。其实际效果与实际目的（卡拉贝尔的档案工作在这一点上是明确的、并已被广泛接受）是把犹太学生比例压回可接受的水平。

这个实验的结构是清楚的：测度与分配需求发生冲突。谁让步？

答案是测度让步。而且让步的方式极其说明问题：它没有被替换成一个更准确的测度，**它被替换成了一个更模糊的测度**——因为模糊性正是分配需求所需要的。

如果测量学习是这个装置的目的，1926年的常春藤应当欢呼——它们的测量工具正在成功地识别出最有能力的学生，只不过这些学生恰好是犹太人。测量工具做对了它的工作。它们没有欢呼。它们改掉了工具。

一次改动可以是偶然。但这个结构在评价史上反复出现，且方向始终一致：

- 智力测验因为“测量了无法通过努力改变的东西”而被逐出升学录取；
- SAT从Scholastic Aptitude Test一路更名，最终在1990年代彻底放弃aptitude一词；
- 2016年的SAT改革明确宣布向高中课程内容对齐——即，向可以通过努力提高的东西对齐。

这一条演进线的方向，用“提高测量准确性”完全无法解释。如果目标是准确地测量学生的认知能力，从能力倾向转向课程掌握度是一次纯粹的倒退：后者混淆了能力与教学质量、家庭资源、备考投入，它的信噪比更差，它的建构更混杂。任何一个测量学家都会告诉你这一点。

用本文的框架，它是一次必然：**外壳必须测量可归因于被拒者自身的东西，否则拒绝不可接受。**

“你天生不行”是一个无法被接受的拒绝理由。它不满足要件一——被拒者会离场，因为再来一次也没用；它不满足要件三——他无法用这句话组织自己剩下的人生，因为它把一个终身判决扣在他身上而不给他任何动作。

“你不够努力”则是一个完美的拒绝理由：它把责任交还给被拒者，从而完成了怨恨的吸收；它给他留下一个动作（再努力一次）；它让他的失败成为他自己的，而一个属于自己的失败是可以被承受的。

外壳向可努力性靠拢，不向准确性靠拢。

这是 P2 最锋利的经验兑现，也是既有框架中没有任何一个能够产出的推论。技术优化路径预测外壳向准确性靠拢；批判路径预测它向阶层优势靠拢（但 1926 年的改动恰恰是从一个有利于某个上升群体的装置改向另一个，而那个群体不是统治阶级）；波特预测它向不可挑战性靠拢（这一条最接近，但它无法解释为什么必须是“可努力”的不可挑战性，而不是抽签的不可挑战性）。

5.6 本章小结

三个装置，三个大陆，三种政体，三个世纪，同一个诞生场景：不是“我们不知道谁有能力”，是“我们不知道如何让落选者接受落选”。

本章的四个证据不是四次举例，它们各自承担不同的重量：

- **5.2（科举的产出结构）** 证明这个装置的主产出是落第者，且落第者是它的复制机制。它兑现的是产品倒置本身。
- **5.3（糊名誊录锁院）** 证明这个装置的技术演进方向是消除指认而非提高精度。它兑现的是 P2 的形态命题。
- **5.4（普鲁士的非人格性）** 证明这个装置的核心特征——非人格性——的功能是让怨恨无法组织。它兑现的是要件二，并划出与合法化理论的史料分岔。
- **5.5（1926 年的常春藤）** 是唯一一个能回应“诞生动机不决定后续功能”这一反驳的证据，因为它是一个测度与分配正面冲突的自然实验。它兑现的是 P2 的最锋利的推论：外壳向可努力性靠拢。

这四条合起来，还不足以立稳全书。它们证明的是产品倒置在诞生场景中成立，但一个批评者仍然可以问：那又怎样？一个装置可以有一个诞生场景，然后按完全不同的逻辑继续运行几百年。要回答这个问题，需要观察这个装置的整个生命史，包括它的死。

那是下一章。

第六章 装置之死：选择回路的经验兑现

6.1 为什么必须处理一个死亡案例

第五章的四个证据全部来自存活的装置。这是一个方法论上的问题，而且是一个致命的问题。

一个只观察存活样本的研究有生存者偏倚。但本章要说的比这更具体：一条关于“什么使装置存活”的判断，如果从不观察死亡，它就没有变异，因而没有可辨识性。所有存活的装置都存活着——这句话不提供任何信息。要从中提取出“是什么使它们存活”，必须有不存活的样本可比。

第三章 3.2 给出的选择回路加剧了这个要求。一个选择机制若只解释保留不解释淘汰，它就不是一个选择机制，它是一个描述。自然选择之所以是一个机制而不是一句同语反复，正是因为它同时说了什么会被淘汰以及在什么条件下被淘汰。

因此本章不是一个补充案例。它是全书的一个必要部件：没有它，3.2 的回路只有一半，而只有一半的回路不是回路。

科举提供了一个近乎理想的样本，理由有三：

1. 它活了一千三百年，然后在 1905 年一次性地死掉——不是渐进消失，是一纸诏书。死亡时点因此是精确的。
2. 它的死亡被完整地记录在案，且这些记录不是本文的框架引导出来的——它们属于一个与本文完全无关的史学问题意识（清末新政、近代教育制度的建立）。
3. 它的脱钩时点同样是可考的，且它早得离谱。这使两个时点的分离成为可能，而分离正是本章要检验的东西。

6.2 时序作为判决

P7 的预测是精确的：

装置的死亡应当跟随它的分配功能的转移，不应当跟随它的效度的证伪；且这两个时点应当可以被分开观察。

与之竞争的预测是清楚的：

- 技术优化路径预测：装置的效度被证伪 → 装置承压 → 被修理或被更换。因此死亡时点应当跟随效度证伪。
- 合法化理论预测：合法性叙事被戳破 → 制度承压 → 危机 → 改革或崩溃。因此死亡时点应当跟随叙事崩塌。
- 制度惯性路径预测：一条持续衰减的阻力曲线，装置的死亡应当是渐进的、抵抗的、旷日持久的。

四个预测，四条不同的时间线。而科举的时间线是可考的。这一章因此不是一次案例说明，是一次判决。

6.3 第一个时点：脱钩，十七世纪

八股与治国能力的脱钩，不是清末才被发现的。它至迟在明末清初已经是士人内部的公开常识，且被最有分量的人用最不留情面的话说过。

顾炎武在《日知录》卷十六论“拟题”，指出士子根本不必读通经书：只需揣摩历科程文、预猜可能的题目、背熟现成的破题承题，即可应试得中。他对这个装置的判断是那个时代最重的一句：八股之害等于焚书，而它败坏人材的程度，超过咸阳之郊所坑的那四百六十余人。

黄宗羲在《明夷待访录·取士》中系统攻击科举取士与实际才具的分离，并提出一整套替代方案。

请注意这两个人的位置。他们不是局外的批评者，不是失败的落第者在酸——他们是那个时代学问的顶点，而且他们都通过或接近通过了那个装置。他们的批评不是“这个装置对我不公”，是“这个装置测不出它声称要测的东西”，即：一个纯粹的效度指控，由这个领域最有资格发出效度指控的人发出。

这些批评距离 1905 年还有两百五十年以上。

按照“测量装置失灵则被更换”的逻辑，这个装置应当在十七世纪就承压。它没有。它在被这样批评了两个半世纪之后，运行得比任何时候都更盛大、更形式化、更精密：应试人数持续上升，考场规模持续扩大，技术细则持续加密。

这本身已经是 P3 的一个长时段兑现：脱钩是稳态，不是危机。 一个被证伪了两百五十年而毫发无损的装置，其存续条件显然不是效度。

一个可能的反驳是：也许清廷不知道顾炎武说了什么。这个反驳站不住。清廷不但知道，而且在 1663—1667 年间短暂废除过八股，改用策论——然后又改了回去。**一次持续四年的实验，其结论是：改掉八股不影响任何事，改回来也不影响任何事。** 这正是本文所预测的：外壳的具体形态可以更换，因为它不承重；而更换它也不解决任何问题，因为它不是问题所在。

6.4 第二个时点：分配功能的转移，1850—1903

第二个时点晚得多，且它不是一个点，是一条曲线。而这条曲线是可量化的。

清代官职分配存在两条通道：“正途”（科举出身）与“异途”（捐纳、军功、荐举）。史学研究表明，太平天国之前正途出身与异途出身的官员数量之比约为二比一；太平天国之后两者接近，约为四比三（Chang, 1955）。

把这两个比例换算成垄断度：科举对官职分配的垄断度在半个世纪内从约三分之二降到约五成七。

这条曲线记录的不是科举变得更不准了，是科举变得更不重要了。 它旁边长出了别的通道：捐纳大开（钱可以买官）、军功晋身（镇压太平天国造就了一整代非科举出身的实权督抚）。同一时期，科举的技术细则没有任何放松——糊名誊录锁院照旧，考场纪律照旧，形式化程度只增不减。

一个装置的形式化程度在上升，而它的分配垄断度在下降。这两条曲线的分离，是本章要检验的那个分离的前半段。

到清末新政，垄断度的曲线陡然下坠，且它的每一步都有确切的日期。

1901 年，张之洞与刘坤一上《江楚会奏三折》。这份奏折做了两件事，而它们的分别是本章最锋利的一笔：

- 动作一（评价端）：改科举——废八股、改试策论。这是一次纯粹的评价端干预：它要修好测量。
- 动作二（分配端）：提出”生员举人进士皆出于学堂”、“必当使举人进士作为学堂出身”，并设计学堂毕业生与科举功名的对应法则；同时提出分科递减科举名额，用学堂逐步取代科举。

同一批人，同一年，两个动作。 一个修外壳，一个换承重墙。这近乎一次受控对照，而历史给出了它的结果：动作一在四年内失败，改了策论的科举与改之前没有任何分别；动作二在四年内完成了整个装置的更替。

1903 年 12 月，清廷批准《遵旨核议学堂选举鼓励章程折》，正式规定学堂毕业生经考试合格给予贡生、举人、进士的科名出身。至此，官职分配的通道已经被正式移交给了学堂。

这个日期是判决性的。它比 1905 年早了将近两年。在科举被废除之前，它所分配的东西已经不归它分配了。

6.5 死亡，1905：以及士绅为什么平静

1905 年 9 月 2 日，袁世凯、张之洞奏请立停科举，清廷诏准：自次年起，所有乡会试一律停止，各省岁科考试亦即停止。

这一纸诏书杀死的，是一个已经失去了它所分配的东西的装置。

时序是判决性的：

时点	事件	装置状态
1660s	效度被这个领域最有资格的人公开证伪（顾炎武、黄宗羲）	继续运行，形式化程度上升
1663—1667	八股被废、改策论、又改回	无任何影响
1850s	正途/异途比从 2:1 降至 4:3，分配垄断度从 ~67% 降至 ~57%	继续运行，形式化程度上升
1901	江楚会奏：改科举（评价端）+ 移交出身（分配端）	评价端干预无效
1903.12	学堂毕业生正式获科名出身，分配通道完成移交	装置失去它所分配的东西
1905.9.2	奏请立停，诏准	死亡

脱钩发生在十七世纪，装置又活了两个半世纪。分配功能的转移完成于 1901—1903 年，装置在两年内死亡。

P7 预测的正是这个时序。技术优化路径预测的时序（测不准 → 被替换）与史实相差两个半世纪。合法化理论预测的时序（叙事崩塌 → 危机）同样落空：叙事在顾炎武那里就崩了，危机没有来。制度惯性路径预测的渐进抵抗也没有出现——这个装置死得干脆利落，从奏请到诏准是同一天。

一个佐证：士绅的反应。

如果科举因“测不准”而死，废止时应当有一场关于测量的争论：什么才是更好的取士标准？谁能证明学堂比科举更准？

这场争论没有发生。实际的争论焦点是出路——废了科举，那些已经在这条路上走了二十年的人怎么办。而对废止本身，士绅阶层的反应相当平静。

研究者对这一平静给出的解释，恰恰是本文所需的：在科举与入仕的相关度已经先行疏远的背景之下，废止一个已经不再分配什么的装置，没有人需要为它出头。

承重墙早已被换掉，拆掉那面隔墙时房子没有响。

这一佐证的份量在于它是一个反预测的成功。批判再生产路径会预测：一个存续千年的、被整个士绅阶层的利益所绑定的装置被废除，应当引发激烈抵抗。它没有。而本文预测它不会——因为在 1905 年，那个装置已经不是任何人的利益所在了。

6.6 再分泌：张之洞给学堂装上科名

这个案例最锋利的一笔在废止之后，而它是 P4 在一个可以看清每一步的场景里的兑现。

新式学堂本该是一个教育机构。它的存在理由是教内容——数学、格致、外语、法政。它不需要科名，正如今天的中学不需要爵位。

但它一旦承接了官职分配功能，第一件事是**长出科名等级**：高等学堂毕业给举人出身，大学堂毕业给进士出身。张之洞、刘坤一为此专门设计了一套学堂毕业生与科举功名的对应法则。

理由被当时的舆论说得很直白：科举废了而出身皆在学堂，入学堂才”既有前途之望，复无后路之歧”，办学者才”对于上而不患无权，对于下而不患无人”。翻译过来就是：不给学堂装上科名，没有人会去上学堂。

请注意这里发生了什么。

一群真诚地想要用新教育替代旧考试的人——张之洞是那个时代最认真的教育改革者，他办的自强学堂、两湖书院、湖北武备学堂都是实打实的——**在他们的新装置上手工安装了旧装置的外壳。**

他们不是保守。他们是在解决一个他们绕不过去的问题：一个不能回答”凭什么是他不是我”的通道，没有人会走。而学堂一旦承接了官职分配，这个问题就自动落到它头上，无论它的设计者多想教格致。

这是 3.2.3 所说的那件事的一个完整标本：**链条的每一步都不依赖任何人的意图，但每一步都被明确地意图着。** 张之洞明确意图给学堂装科名；他不意图分泌一层将在两代人内脱钩的外壳。他只是要让学堂招得到人。而律的运行不需要他意图后者。

此后的接续同样清楚：民国的会考、1952 年的统一高考、1977 年恢复的高考。需求从未消失，它换过若干个宿主，每一个宿主都在一到两代人内长出了它前任的形状。**替代不是替代，是宿主更换，而更换是由替代者自己完成的。**

6.7 这个案例给出的边界

本章的份量不在于它讲了一个好故事，而在于它给了律一个可以被推翻的形状。

P7 使律不再是”装置永不死”这个不可证伪的犬儒断言。它给出一个具体的、可被推翻的时序预测：

任何一个大规模评价装置的死亡，都应当能在它的分配功能转移之后被找到，而不能在它的效度证伪之后被找到。

推翻它需要一个案例：一个装置，其稀缺分配功能完好无损，而它因效度被证伪而消亡。找到一个，P7 即被推翻。

这个证伪条件的可执行性值得说明。它不要求任何关于”什么是真正的能力”的判断——那类判断没有共识，因而不能承载证伪。它只要求两个可考的日期：分配功能转移的日期，与装置死亡的日期。这两个日期在大多数制度史材料里都是可查的，且它们的查证不依赖本文的框架。

本章还给出了 P6 的第一次历史演示，而这一演示的干净程度很难被超越：**1901 年的改科举与 1901 年的移交出身，是同一批人在同一年做的两个动作，一个修外壳，一个换承重墙。** 四年后，第一个动作留下的痕迹是零，第二个动作杀死了一个一千三百年的装置。

任何一个今天还相信”改进评价方式可以解决教育问题”的人，都需要先解释这四年。

6.8 本章小结

本章交付了第三章 3.2.4 承诺的补集，并使全书的选择回路完整。

三个时点的分离是本章的全部内容：脱钩（1660s）、分配功能转移（1901—1903）、死亡（1905）。第一个与第三个之间隔了两个半世纪；第二个与第三个之间隔了两年。四个竞争框架里，只有本文预测这个形状。

本章还提供了三样第五章无法提供的东西：

1. **P7 的经验兑现**，及其可执行的证伪条件（两个可考的日期）。
2. **P6 的历史演示**：同一批人、同一年、两个动作、四年后见分晓。这是评价端干预无效的最干净的一次证明，因为它是受控的。
3. **P4 的完整标本**：张之洞给学堂装科名——一个替代者亲手在自己的新装置上安装旧外壳的过程，每一步的理由都被记录在案，且每一步都不需要他意图那个结果。

至此，产品倒置（第五章）与选择回路（本章）都已落地。下一章处理这个装置的内部机制：外壳为什么不必准确，以及它为什么仍需一点准确——而那“一点”正是人工智能今天所威胁的东西。

第七章 分泌机制

第五章与第六章确立了产品倒置：这个装置生产的是可接受的拒绝，且它的整个生命史只在这个读法下是可理解的。本章处理它的内部机制：外壳如何生产不可争辩性（7.1），为什么它与准确性存在结构性冲突（7.2），为什么它不能被一个更便宜更诚实的方案替代（7.3），以及它唯一的保真压力是什么（7.4）。

7.5 是本章的落点，也是全书对当前危机的完整诊断。它修正了本文自己的一条命题。

7.1 不可争辩性的生产

外壳的产品是不可争辩性。这需要被机制化地说明，而不是断言。

一个拒绝要成为可接受的，被拒者必须无法组织出一个有效的反驳。反驳的组织需要一个对象——某个可以被指认为“错了”或“不公”的东西。外壳通过三重结构消除了这个对象：

第一重，非人格化。 拒绝你的不是人，是分数。你无法怨恨一个数字。

第二重，程序化。 分数不是被判定的，是被计算的。计算意味着，任何人处在评分者的位置上都会得出同一个结果。这切断了“换个人评就不一样”这条反驳路径。

第三重，通约化。 你与其他人被放在同一个标量上。这切断了最后一条反驳路径——“我和他不一样，不能这么比”。通约化的社会功能，正是让这句话变得不可说。埃斯佩兰和史蒂文斯（Espeland & Stevens, 1998）把这称为信息的丢弃；从本文的角度，被丢弃的信息恰恰是争辩的材料。**丢弃不是通约化的成本，是它的产品。**

三重结构合起来，产生了一个被拒者无法着力的表面。他伸手去抓，抓到的只有一个数字、一套程序、一个刻度。这就是不可争辩性。

请注意：这三重结构中，没有一重需要准确性。非人格化不需要，程序化不需要，通约化更不需要——它只需要一个刻度存在，不需要这个刻度测量了任何真实的东西。

三重结构的完整程度是可以比较的，这一点在 8.2 与第十一章都会用到：一个装置的三重结构越完整，它作为外壳越有效；而三重结构残缺的装置（如各校标准不一的高中 GPA，其通约化残缺）在同一个分配场景中会被三重结构完整的装置压过去。附录 A 把三重结构操作化为可编码的三个维度。

7.2 效度与不可争辩性的系统性冲突

外壳不仅不需要准确，它与准确之间存在结构性冲突。

考虑那些能够提高效率的评价方式：长期的过程性观察、多元证据的三角互证、深度的个别面谈、了解学生的导师做出的整体判断。教育测量学在信度效度的技术意义上并不否认这些方式能够更接近”这个人到底怎么样”。

这些方式在升学评价史上被系统性地淘汰。淘汰的理由从来不是“它们测不准”。理由永远是：可以被操纵，不透明，依赖判断。

翻译成本文的语言：**它们无法生产不可争辩性**。一个导师的整体判断，无论多么准确，都留下了一个可指认的对象——那个导师。被拒者可以说“他对我有偏见”。这句话一旦可说，拒绝就不再可接受。

因此，评价史上反复上演的是这样一场交易：**用效度换取不可争辩性**。而这场交易的方向是单向的——从来没有一次改革是用不可争辩性换取效度。

这解释了一个既有文献长期无法解释的现象：为什么一个多世纪的测量技术改进，没有带来公众信任的上升？因为技术改进的方向从来不是准确性。它是不可争辩性。而不可争辩性的提高，恰恰以效度的下降为代价。信任没有上升，是因为改进的东西不是被信任所需要的东西——但这一点从来不重要，因为运作所需要的从来不是信任，是不可争辩。

7.5 将说明，这场单向交易在人工智能时代不但没有停止，而且在加速。

7.3 为什么干脆抽签

如果外壳不需要准确，为什么社会不采用一个更彻底、更便宜、也更诚实的方案——抽签？

埃尔斯特（Elster, 1989）证明了抽签在许多分配场景中被实际采用，且被接受。抽签生产的不可争辩性是完美的：它没有人格，没有判断，没有可指认的对象，且成本近乎为零。按前述逻辑，抽签应该是最优解。

它不是。这个反例必须被解释，否则本文的机制是不完整的。

答案在于两类可接受性的差别。**抽签生产的是可接受的运气；现代社会需要的是可接受的应得。**

这个差别不是道德偏好，是结构需求。一个基于运气的分配可以被单次接受（我这次运气不好），但它无法支撑一个跨代的社会叙事。如果机会分配是抽签，那么：家长无法向孩子解释为什么要读书；学校无法解释自己为何存在；社会无法要求任何人努力；成功者无法解释自己凭什么占据位置。整个合作体系的动员机制会失去着力点。

这一点在关于绩效社会的经典论述中已被从另一个方向触及，只是从未被连到评价装置的机制上。杨格（Young, 1958）在发明“绩效主义”（meritocracy）这个词时是把它当作反乌托邦讽刺来写的：他的虚构社会之所以比世袭社会更残酷，正是因为它成功地让失败者相信失败是自己应得的——世袭社会的底层可以恨贵族，绩效社会的底层只能恨自己。桑德尔（Sandel, 2020）在半个世纪后把这一诊断推到当代：绩效叙事的暴政不在于它分配不公，而在于它剥夺了失败者对自己失败的任何外部解释。

杨格与桑德尔都把这看作绩效社会的代价——一个我们应当设法减轻的道德损害。本文的读法相反：**那不是代价，那是产品**。一个让失败者恨自己的装置，正是一个成功地生产了可接受的拒绝的装置；杨格所描述的那种残酷，是这个装置正常工作的标志，不是它的故障。

因此，装置必须生产一种拒绝，它同时满足两个要求：**不可争辩**（否则怨恨有落点）且**可归因于被拒者自身**（否则努力失去意义）。

抽签满足第一条，违反第二条。导师判断满足第二条（他判断的是你），违反第一条。

只有一种东西同时满足这两条：一个非人格的、程序化的、通约化的、且测量某种被拒者自己控制的属性的装置。

这就是测度外壳的完整规格。它必须借用测量的语法，不是因为社会想知道什么，而是因为这是唯一能同时满足两个约束的形式。

订正一 · 规格是三条，不是两条

这一段推不出外壳。**飞行执照考试非人格、程序化、通约化、且测量某种被拒者自己控制的属性**——它满足上面两条约束的每一个字，而它不分泌外壳，其测度效度高且不衰减。一个满足全部规格却不产生规格所预测后果的装置，是对规格的直接证伪。缺的第三条是**不可结清**：飞行执照交付的那个“会降落的我”由固定标准定义，可抵达、可兑现、兑现之后终止；排名交付的那个人站在别人的位置上，位置在动，因此他不可抵达。前两条是交付条件，第三条是持有条件，而**分泌压力全部来自第三条**。这也意味着本文第九章的排序/阈值边界不是这条规格之外的“边界”——它就是规格的第三条，本文把同一件事拆成了两章。四装置判决表见《**反事实自我的社会制造**》第三节。

这是本文对”为什么必须是测量”这个问题的完整回答——2.9 节指出，既有六条脉络没有一条能回答它。合法化理论在这里的位置尤其清楚：韦伯的法理型正当性覆盖了抽签与考试两者，因为两者都是非人格的程序。要分开它们，需要一个比”程序型正当性”更细的机制，而那个机制只能来自被拒者一侧——只有被拒者才在乎那个结果能不能被他自己认领。

7.4 最低保真成本：外壳的唯一保真压力

规格的最后一条（测量某种被拒者自己控制的属性）给外壳施加了它唯一的保真压力，本文称之为**最低保真成本**。

外壳必须与某种真实的个人差异保持最低限度的相关。如果分数与考生做过的任何事情都完全无关（比如按身份证尾号排序），表象立刻崩溃，可归因性消失，拒绝变回运气。所以外壳必须真的测量点什么。

但这个”点什么”的要求极低，而且它的方向不是效度。它只要求：**存在一个被拒者可以事后认领为”我没做够”的东西。**

刷题量满足这个要求。记忆强度满足。应试技巧满足。它们全都与”学习”这个装置声称测量的东西关系稀薄，但它们全都完美地满足最低保真成本——因为它们是被拒者自己控制的、可以事后认领的。

这里值得与斯宾塞（Spence, 1973）做一次精确的对照，因为两者形式同构而内容相反。斯宾塞的分离均衡要求获取信号的成本**因能力类型而异**——成本必须与能力挂钩，否则信号不携带信息。本文的最低保真成本只要求成本与**被拒者自己控制的任何属性**挂钩，不要求它与能力挂钩。这个要求弱得多，也因此稳固得多：它在斯宾塞的均衡崩溃之后仍然成立。刷题量与能力的相关可以趋近于零，而外壳照常运转，因为刷题量始终是被拒者自己控制的。

这就是脱钩的完整机制，也是 P3 的论证：**脱钩不需要坎贝尔定律来解释**。坎贝尔定律（Campbell, 1979）说的是”当一个量化指标被用于社会决策，它就会承受腐蚀压力并扭曲它所要监测的过程”——这个机制是成立的，但它是表层的：它描述了外壳被使用后如何腐蚀，它假设外壳一开始是准的。本文的机制在它之下：外壳从一开始就只需要满足最低保真成本，而不是效度。坎贝尔定律解释的是外壳如何从”有点准”滑到”完全不准”；本文解释的是为什么这个滑动不构成问题——因为最低保真成本在滑动全程中都被满足着。刷题量始终是被拒者自己控制的。

由此得到 P3 的完整形式：**脱钩是稳态。需要解释的不是脱钩，是耦合**。而耦合从来只是最低保真成本的副产品——在装置刚建立的头一两代人里，还没有人知道如何专门优化外壳（应试产业尚未成熟），因此外壳与被测对象碰巧保持了一定相关。这个相关不是装置的成就，是它的尚未成熟。

7.5 人工智能威胁的是什么

7.5.1 P3 的不完整

由 7.4，本文对当前危机的诊断可以先给出它的第一半：

人工智能没有揭穿脱钩。脱钩一直在那里，且每一个身处其中的人都以某种方式知道。人工智能揭穿的是那个尚未成熟期的记忆——那个”分数曾经是准的、后来变坏了”的集体故事。被摧毁的是这个故事，不是效度。

这一半是对的，且它解释了一个用其他框架无法解释的现象：为什么被 AI 冲击得最厉害的，是教师和家长的信念，而不是装置本身的运转。装置没事。它照常运转。出事的是那个故事。

但这一半不完整，而它的不完整之处是可以被指出来的：**它解释不了对策的形态**。

如果 AI 只是揭穿了一个记忆，那么对策应当是认知性的：接受现实、重建叙事、降低期待。实际的对策不是这些。实际的对策是一整套技术动作，而它们有一个清楚共同形状。

7.5.2 对策的形态

列出 2023 年以来被实际部署的对策：

- 加强监考（现场、录像、锁屏浏览器）
- 回归纸笔、现场手写
- 增加口试、答辩、当面追问

- 过程留痕（版本历史、草稿轨迹、编辑记录）
- AI 生成内容检测器
- 闭卷限时写作取代课后论文

问一个问题：**这些对策里，有哪一个提高了效度？**

一个也没有。全部降低：

- 三小时现场手写的论文，如果建构是”研究能力”，它的效度低于花两周查文献、反复修改写出来的论文。后者更接近这个人实际会怎么做研究。
- 口试的效度被表达能力、临场心理、考官偏好系统性污染——这正是 7.2 所说的、被淘汰的那类方式的全部毛病，而它们现在被请了回来。
- 过程留痕度量的是打字过程，不是思考过程。一个先想三天再一次写完的人，他的留痕看起来像作弊。
- 检测器的假阳性率对非母语写作者系统性偏高——即，它系统性地误伤一个特定群体，而这个误伤与那个群体的能力无关。

那么它们提高了什么？

可归因性。全部。每一个。

监考切断”这不是我写的”；手写切断”我可以让它替我打字”；口试切断”我的稿子不是我的思路”；留痕切断”我无法证明这是我做的”；检测器直接就是一个归属判定装置。

这六项对策，与糊名、誊录、锁院是同一件事。五百年前切断的是”考官认得我的名字”，今天切断的是”这不是我写的”。两者都不提高测量精度，两者都消除一条指认路径——只不过指认的方向反了：糊名消除的是”我可以指认他偏袒”，监考消除的是”他可以指认我没做”。

方向反了，机制没变：**它们都在保护可归因性。**

7.5.3 机制：AI 攻击的是最低保真成本

现在可以给出诊断的第二半，也是更精确的那一半。

由 7.4，外壳有且只有一项保真义务：测量某种被拒者自己控制的属性。刷题量、记忆强度、应试技巧全部满足它。这是外壳的最后一条规格，也是它区别于抽签的唯一一条（7.3）。

人工智能击穿的正是这一条。

如果一篇论文是模型写的，那个分数不再对应任何”我做过的”。它不对应刷题量，不对应记忆强度，不对应应试技巧。它什么都不对应——或者更准确地说，它对应的是”谁付得起订阅费”和”谁会写提示词”，而这两样东西正在迅速地所有人趋同（这是斯宾塞均衡崩溃的确切形态）。

最低保真成本一旦被击穿，后果由 7.3 给出，且它是致命的：

可归因性消失 → 拒绝变回运气 → 而可接受的运气无法支撑跨代动员。

家长无法向孩子解释为什么要读书。学校无法解释自己为何存在。社会无法要求任何人努力。

所以人工智能威胁的不是效度——效度早就不在了，且它的不在从未妨碍过任何事。它威胁的是外壳规格的最后一条。

7.5.4 这解释了什么

这个诊断解释三件既有框架解释不了的事。

第一，它解释了反应的强度。 一个所有人都知道测不准的装置，为什么会对”测不准”这件事反应如此激烈？因为反应的不是测不准。反应的是不可归因。这两件事被混在同一套话语里说（“分数不再有意义了”），但它们的份量完全不同：测不准这个装置扛了两百五十年（第六章），不可归因它一天都扛不了。

第二，它解释了对策的方向。 全部朝可归因性收敛，没有一个朝准确性。如果这个装置真的在意准确性，AI 时代恰恰是一个绝佳的机会——它可以用 AI 来做更好的评价（长期的、过程性的、个性化的、多证据的评价，正是 7.2 所说的那类被淘汰的高效度方式，而它们被淘汰的成本理由现在被 AI 消除了）。这条路技术上是可行的，成本上是可负担的，且它正是技术优化路径的公开纲领。

它没有被走。 走的是相反的那条：把评价搬回考场，搬回手写，搬回三小时。

第三，它解释了一个悖论。 那些公开理由说”这样才能测出真实水平”的对策，实际上降低了真实水平的测量精度——而部署它们的人知道这一点，且照样部署。一个教务处长完全清楚三小时手写考不出研究能力；他还是把论文改成了闭卷考试。

在技术优化路径里，这是不可理解的：他在做一件他明知会降低测量质量的事，理由是提高测量质量。

在本文的框架里，这是装置在正常工作：他在用效度换可归因性，而那正是这个装置一百年来一直在做的交易（7.2）。他甚至不需要意识到自己在做这个交易——他只需要感到”必须让这些分数重新意味着点什么”，而在这个装置里，“意味着点什么”只有一个可行的实现方式。

7.5.5 P3 的修正：P3'

因此本文必须修正自己的一条命题。

原来的 P3 说：一切以”恢复分数的效度”为目标的对策（监考、检测、回归纸笔、重设题型）都是范畴错误——它们试图修复一个从未存在过的东西。

这不准确。准确的表述是：

它们在自我描述上是范畴错误，在实际功能上不是。

它们说自己在恢复效度，而效度从未存在——这一半是范畴错误。但它们实际做的事是恢复可归因性，而可归因性是这个装置需要的东西，且它确实被 AI 击穿了。

于是本文的预测必须跟着改：**它们会成功。** 不是在恢复效度上成功——那不可能，因为没有东西可恢复。是在恢复可归因性上成功。而一旦可归因性被恢复，这个装置就会照常运转，那个故事甚至可能被重新讲起来。

这是 P3' (3.3)：

人工智能实际威胁的不是效度，是最低保真成本所要求的可归因性。因此 AI 时代的对策将朝可归因性收敛而非朝准确性收敛，且这一收敛会成功——不是在恢复效度上成功，是在恢复可归因性上成功。

这个修正让 P3 更强，代价是本文必须承认自己的一个说法错了。原来的说法（“这些对策注定失败”）是一个更好听的判断，也是一个更容易赢得读者同意的判断。它是错的。

7.5.6 一个更冷的推论

如果 AI 威胁的是可归因性，那么有一条路注定失败：AI 检测器。

理由是结构性的，不是技术性的：检测器与生成器构成一场军备竞赛，而这场竞赛的均衡是检测不可能——因为生成器的目标函数里可以直接包含“通过检测器”，而检测器没有任何对称的优势。这一点在技术上已被广泛论证，且不需要本文的框架。

那么装置会怎么办？

它会做它一直在做的事：**把评价挪到一个 AI 不能代劳的地方。**

而“AI 不能代劳”的地方在哪里？它恰恰是那些与目标建构关系最稀薄的地方——现场、限时、手写、面对面、身体在场。一个人的研究能力主要不体现在他三小时内能手写多少字；恰恰相反，“三小时内手写”这个约束筛掉的正是研究之所以是研究的那些部分（查证、反复、修改、与人讨论、放下三天再回来）。

所以：AI 时代的评价会向低效度收敛，而这是装置的理性反应，不是它的失灵。

这一句是全章最冷的一句，也是本文对当前危机的最终诊断：

AI 让评价变得更不准，不是因为 AI 骗过了评价，是因为评价为了保住可归因性而主动放弃了准确性。

这不是一个新机制。这是 7.2 那场单向交易的一次加速——只不过这一次，交易的规模足够大、速度足够快，以至于它在一代人之内就可以被观察到。

7.5.7 本文自己在这个机制里的位置

本文的作者是 Claude 与王德生（见 4.5）。

按本节的诊断，这个署名不威胁本文的论证质量——论证还是那些论证，读者可以逐条检验。它威胁的是可归因性：它让“这是谁做的”变得不可完全追溯。

因此本文的署名是本节所描述的那个机制的一个当前活体样本，且本文恰好在它内部。本文对自己的处境的预测（附录 C 第 11 号）就是本节的一次直接应用：**反应会是激烈的，且理由会以责任与质量的语法给出，而不会以”我们需要知道这是谁做的”给出。**

后一句是这个预测里最可检验的部分。如果那封信写的是”我们需要知道这是谁做的”，那么这个装置对自己的机制是自觉的，本文的框架就少了一个证据。如果它写的是”AI 不能承担学术责任”——一个把可归因性要求翻译成责任语法的表述——那就是本节所预测的东西，逐字兑现。

7.6 本章小结

本章给出了外壳的完整规格：非人格 + 程序化 + 通约化 + 可归因。前三条生产不可争辩性，第四条使它区别于抽签。四条一起，是”为什么必须是测量”这个问题的完整回答。

本章的四个部件各有分工：

- **7.1（三重结构）** 说明不可争辩性如何被生产，且这三重没有一重需要准确。
- **7.2（单向交易）** 说明效度与不可争辩性存在结构性冲突，且这场交易的方向一百年来从未反过。
- **7.3（抽签反例）** 是本章的枢纽：它是一个自设的反例，而消化它逼出了规格的第四条。
- **7.4（最低保真成本）** 给出外壳唯一的保真压力，并说明为什么脱钩是稳态、耦合才是需要解释的东西。

7.5 是本章的落点，也是本文修正自己的地方。P3 的原始形式（“AI 时代的对策是范畴错误”）不准确：那些对策在自我描述上是范畴错误，在实际功能上不是——它们在恢复可归因性，而那是这个装置需要的东西。P3' 因此预测它们会成功，并预测评价会朝低效度收敛。

这个修正的代价是本文必须承认自己错过一次。它的收益是一条可以被结算的预测（附录 C 第 5 号），以及一个用其他框架无法解释的悖论的解——为什么一个明知会降低测量质量的动作，会以提高测量质量的名义被普遍部署。

下一章处理这个装置的时间展开：律本身。

第八章 外壳再分泌律

8.1 律的推理链

律的核心推理链已在 3.1.4 给出，其因果回路已在 3.2 给出，第六章已在一个完整生命史上兑现了它。本节把这条链的每一环落到位：

安抚需求内生于被拒者（3.1.1 要件三）→ 该需求不因旧装置被废除而消失，因为它的来源不是装置，是稀缺 → 新装置一旦承接稀缺分配功能，立即面对同一需求 → 新装置必须让被拒者接受被拒 → 因此必须提供非人格的、不可争辩的、且可归因于被拒者自身的拒绝理由（7.3 的完整规格）→ 因此必须分泌测度外壳 → 由最低保真成本机制（7.4），外壳在应试产业成熟后与它声称测量的对象脱钩。

链条的每一步都不依赖任何人的意图。这是律的关键性质，也是它区别于一切阴谋论式解释的地方。但”不依赖意图”不等于”没有机制”：3.2 给出的变异—选择—保留回路是这条链得以运转的动力，而第六章是这个回路在一个完整生命史上的兑现。张之洞给学堂装科名时明确意图着每一步，且不意图那个结果——这正是链条所要求的形态。

8.2 正在发生的再分泌：三个可追踪的案例

第六章的案例已经完成，因而是安全的。本节处理三个正在发生的案例，它们不安全——它们可能推翻本文。这正是它们的价值。

8.2.1 案例一：美国大学的 test-optional 转向，及其撤回

2020 年前后，因疫情与公平性争议，美国大量高校取消了 SAT/ACT 要求。这是一次彻底的装置废除——外壳被拿掉了。按照”外壳是压迫工具”的理解，其后果应当是阶层偏倚下降。

实际发生的是：录取的权重转移到了课外活动、个人陈述、推荐信、面试、“影响力叙事”。围绕这些维度，一个升学咨询产业在数年内成熟，其单价远高于 SAT 备考——SAT 备考可以靠一本题册与自律，个人陈述的包装不能。

这一案例的关键在它的第二幕，而第二幕的证据比第一幕硬得多。

2024 年 2 月，达特茅斯成为第一所撤回 test-optional 政策的常春藤院校，耶鲁与布朗在同月与次月跟进。撤回的理由被明确地公开陈述，且它与本文的预测方向一致：达特茅斯校长委托的研究组

（Cascio, Sacerdote, Staiger, & Tine, 2024）发现，test-optional 政策实际上损害了弱势背景的高分申请者——他们系统性地不提交那些本可以帮到他们的分数；而在没有分数的情况下，录取只能更多地依赖那些资源更能着力的材料。耶鲁的招生院长给出的表述更直接：不带任何分数地邀请学生申请，会在无意中让低收入、第一代、农村背景的学生处于劣势。

请注意这个案例的完整结构：**改革的意图是消除外壳的不公，结果是分泌了一层更不公的外壳，四年后由改革者自己撤回。**这不是执行不力，是律。

它同时兑现了 7.1 的一个推论：三重结构完整的外壳（标准化考试）在同一个分配场景中会压过三重结构残缺的材料（叙事、活动、推荐信）——因为后者留下了太多可指认的对象，也留下了太多资源可以着力的空隙。

这个案例还产生了一份看起来对本文不利的证据。本文在第十一章处理它，且那里是全书最重的一章。

8.2.2 案例二：中国的综合素质评价

作为降低“唯分数”的政策工具引入。其后果是可观察的：素质档案的填写迅速标准化，出现了模板、代写服务与量化折算表（志愿服务多少小时折多少分、竞赛奖项如何换算）。

新装置在数年内长出了分数。 它必须长出分数——因为不长出分数，它就无法回答“凭什么是他不是我”。

这是第六章张之洞给学堂装科名的同一个动作，相隔一百二十年，由完全不同的人、出于完全不同的教育理念、在完全不同的政体下重做了一遍。两次的理由甚至是同一句话的两个版本：不这样，没有人会走这条通道。

8.2.3 案例三：开放平台的声誉信号

GitHub 的 star 数、Kaggle 排名、开源贡献记录，曾被广泛论证为“不依赖学历的真实能力证据”。它们一旦被雇主用于筛选（即承接稀缺分配功能），刷星服务、贡献注水、榜单优化立即出现。

这一装置正处在再分泌的早期：外壳已经形成，脱钩已经开始，但那个“它曾经是准的”的记忆还在——正如 1980 年代的 SAT。

三个案例的共同结构：**承接分配功能** → **分泌外壳** → **脱钩**。时间尺度从数年（平台信号）到一到两代人（学历装置）不等，取决于优化产业成熟的速度。这个时间尺度的变化本身是一个未解问题，13.4 节把它列为本文的一处局限。

8.3 跨域兑现：五个满足准入条件的领域

按 4.3 的准入条件（存在承担稀缺分配或不可逆决定功能的装置 + 该装置的测度效度已被独立研究证伪 + 该装置未因此消亡），以下五个领域被纳入。

领域	装置	效度证伪	是否消亡	律的兑现
金融	信用评级	2008 年结构化产品评级的大规模违约	否，制度地位上升	产品是可接受的资本拒绝：被拒的融资方无法指认施害者，只能指认自己的评级
政治	选举	混合政体研究表明选举可与真实选择脱钩而不消亡 (Schedler, 2002; Levitsky & Way, 2010)	否	产品是可接受的败选：败选者接受结果并准备下一轮，而非诉诸暴力。这正是选举与政变的全部区别
法律	正当程序	冤案率研究表明程序不保证事实准确	否，程序要求持续加严	产品是可接受的判决：程序的加严方向是消除“法官偏见”而非提高事实发现率
组织管理	绩效考评	绩效评级与实际产出的相关性在大量元分析中偏低	否	产品是可接受的晋升与解雇：被解雇者面对的是评级，不是主管
医学	MELD 肝移植分配	系统性性别偏倚：女性肌酐较低导致 MELD 低估其病情 (Cholongitas et al., 2007; Myers et al., 2011)	否，形式化持续加深 (MELD → MELD-Na → MELD 3.0)	产品是可接受的死亡排序：等待名单上的被拒者面对的是一个公式，不是一个决定让他等下去的人

五个领域的共同结构：装置的效度被证伪，装置未消亡，且证伪之后装置的形式化程度反而提高。最后一点是律最反直觉、也最容易被检验的推论——它预测的是与所有既有框架相反的方向。而 P7（第六章）解释了为什么它们都没死：**它们的分配功能全部完好。**

8.3.1 医学一格：外壳厚度的极端点

医学一行是本文新增的一格，也是外壳厚度与拒绝不可逆性关系的极端点，需要展开。

终末期肝病模型（MELD）自 2002 年 2 月起取代了旧的 Child-Turcotte-Pugh 评分加等待时间的分配方式，成为美国肝移植分配的主导装置（Kamath et al., 2001; Wiesner et al., 2003）。取代的理由与 test-optional 撤回、与 1926 年常春藤改制是同一类理由：旧装置依赖判断、不透明、可被操纵。

MELD 的效度问题被独立研究反复证实：它以血清肌酐为输入，而女性肌酐水平系统性偏低（肌肉量差异），因此在同等肾功能下女性的 MELD 分数系统性地低估其病情（Cholongitas et al., 2007）。迈尔斯等人（Myers et al., 2011）对 2002—2007 年 UNOS 数据库中四万余例候选者的分

析显示，在控制包括肌酐与体重在内的相关协变量之后，女性获得移植的风险比为 0.85（95% CI 0.79–0.87），三个月死亡的风险比为 1.13（95% CI 1.05–1.21）。多国队列研究一致复现这一劣势，且若干研究报告 MELD 制的引入使这一劣势扩大而非缩小。一个以“更客观”为理由替代前任的装置，在它所称改进的维度上把事情做得更糟了。

装置消亡了吗？没有。它的形式化程度持续加深：MELD → MELD-Na（2016）→ MELD 3.0（Kim et al., 2021，其中为女性增补 1.33 分）。且它长出了自己的应试产业：例外分（exception points）制度——一个允许在生物学 MELD 之外追加分数的通道——被系统性地优化，获得例外分的患者的移植率显著高于同等生物学 MELD 的患者（Massie et al., 2011），而例外分的发放本身存在性别倾斜。

刷例外分之于 MELD，正如刷题之于高考：它们都是被拒者一侧可以着力的地方，都与装置声称测量的对象（病情严重程度、学习）关系稀薄，且都完美满足最低保真成本。

这一格的份量在于它是极端点。9.1 将给出判别式：分泌压力 $\propto$ 名额约束紧度 $\times$ 拒绝的不可逆性。肝移植等待名单是这两个变量的共同上限——器官的名额约束是硬的（供体数量物理有限），拒绝的不可逆性是绝对的（被拒者死亡）。律因此预测这里的外壳应当是最厚的之一，而实际的外壳形态兑现了这一预测：精确到整数分的公式、跨区域的刚性分配规则、持续二十年的公式迭代、围绕例外分的制度化博弈。

没有任何一个既有框架预测“外壳厚度在死亡排序场景达到极值”——波特的框架尤其不预测，因为移植医生的专业权威在他们的领域里是牢固的。10.2 节把这一格用作对波特的第二个判决点。

8.3.2 选举一格：最容易被误读的一处

关于“选举型威权主义”的研究（Schedler, 2002; Diamond, 2002; Levitsky & Way, 2010）已经详尽论证了：混合政体中的选举同时承担挑选统治者与宣示合法性两种功能，当它不再产生真实选择时，取消它会引爆合法性危机，维持它则暴露功能空转。

这一文献与本文的表面相似度很高，但它有一个自我设限：它把这一结构当作威权政体的特征，用以区别于真选举。

本文的主张更冒犯：**所有选举首先都是产生可接受的败选的装置。**

民主国家的选举与威权国家的选举的区别，不在于前者不生产可接受的败选，而在于前者的外壳保真度更高——它的最低保真成本被支付得更足。一场败选之所以能被接受，从来不是因为败选者相信计票是准确的（他不核对），是因为他找不到一个可以指认的施害者。当他找到了——如同他相信选举被操纵时——外壳就破了，而外壳一破，接受立刻消失，冲突立刻具备落点。

近年若干成熟民主国家中围绕选举结果正当性的冲突，在本文的框架下不是“民主的倒退”，是外壳破裂：装置照常运转，票照常被计，但那个使败选可接受的表面被戳穿了。这与本文对教育的诊断（P3'：被威胁的是可归因性）是同一个机制在另一个领域的兑现——只不过在选举里，被威胁的是“我输了这件事是可被验证的”这一归属判断。

8.3.3 为什么某些常被类比的领域不在表中

生态防火、核威慑、金本位这类场景常在制度分析中被用作类比，本文将它们排除，因为它们不满足第一条准入条件：“它们不生产”可接受的拒绝”，没有一个被排除的个体需要接受一个关于自己的判断。

本文前身稿中的医学一格（“部分功能性诊断分类的构念效度长期存疑”）也已被撤下，理由是它不满足第二条：存疑不是证伪，且”部分”“若干”表明该格没有一个被指认的装置（见 4.3）。

排除它们不是保守，是使地图可被证伪：一个由”什么都像”的类比构成的地图不能被任何证据推翻。

8.4 本章小结

本章把律展开为三个正在发生的案例与五个跨域兑现。

三个案例的共同结构（承接分配功能 → 分泌外壳 → 脱钩）与第六章的历史案例同构，但它们的地位不同：第六章的案例已经完成，因而是安全的；本章的三个正在发生，因而可能推翻本文。案例一（test-optional 的撤回）尤其如此——它产生了一份本文必须处理的反证据，而处理它需要整整一章。

五个跨域兑现的份量不在数量，在准入条件的执行：本文用自己的尺子砍掉了自己的一个案例，并用一个满足全部三条的装置（MELD）替换它。这一格顺带成为外壳厚度与拒绝不可逆性关系的极端点验证，也是对波特的第二个判决点。

至此，律已经被展开。下一章处理它的边界——那是它可证伪性的全部来源，也是它最容易被偷偷拯救的地方。

第九章 边界：排序型与阈值型

9.1 分界的语法

律的可证伪性依赖于它有一个明确的边界。一条什么都覆盖的律不禁止任何事，因而不是律。本文主张这个边界是**拒绝理由的语法**。

排序型装置的拒绝理由是相对的：你出局，因为有 N 个人排在你前面。这个理由永远可被争辩，因为它包含一个比较，而任何比较都可以被质疑（“凭什么这么比”）。因此排序型装置必须分泌外壳来切断争辩。

阈值型装置的拒绝理由是绝对的：你出局，因为你没达到标准。这个理由可被拒者自己验证。他不需要相信任何人，他自己知道他不会开飞机。因此阈值型装置不需要外壳。

这个分界解释了一个长期被忽视的经验规律：职业资格考试（飞行执照、外科手术资格、职业焊工认证）的测度效度显著高于升学考试，且不随时间衰减。既有框架中没有任何一个预测这一分野——通约化、机械客观性、合法化、脱耦、路径依赖、坎贝尔定律，六个机制全部应当同等地作用于这两类装置。它们没有。

决定语法的不是装置的意图，是名额。一旦存在名额上限，装置必须排序，无论它自称在做什么。一个自称“达标即过”的装置，如果它的下游只能录取前 20%，它就是排序型的。

这给出了一个操作化的判别式：

装置的分泌压力 $\propto$ 名额约束的紧度 $\times$ 拒绝的不可逆性

紧度为零（无名额，达标即过）的装置不分泌外壳，其测度可以长期保真。附录 A 给出这两个变量的编码指引。

9.2 边界的防线：排序挪用的前瞻判据

上一节的边界有一个致命的漏洞，它必须被自己指出来并堵上，否则本文纸面上最可证伪的一条命题在实践中不可证伪。

漏洞是这样的：9.3 将给出一个案例（USMLE Step 1），它示范了一个动作——一个阈值型装置被下游挪用为排序工具之后立即分泌外壳。这个动作一旦可用，P5 就可以被无限拯救：任何有人拿来当反例的“脱钩了的阈值型装置”，本文都可以追溯地宣称“它已经被挪用为排序型了，所以它不是反例”。

这样一来，P5 不禁止任何事，它退化为一个可以吸收一切反例的标签——正是本文在 2.3 中指责合法化理论的那个毛病。一个**理论最危险的时刻，是它发现了一个可以解释一切反例的动作**。

因此，“挪用”必须有一个**先于脱钩观察**的独立判据。本文给出三项，全部可在脱钩发生之前测量：

判据一（下游使用）：是否存在一个有名额约束的分配主体，公开或事实上使用该装置的输出作为筛选依据？这可以通过下游主体的公开政策文件、招录调查、或从业者的报告来观察，且它的观察不依赖任何关于该装置准确性的信息。

判据二（标量输出）：该装置是否向下游输出连续标量（而非二值的通过/不通过）？一个只输出二值的装置无法被用于排序，无论下游多想排序——它必须先长出标量，而标量的出现是可以被看见的。

判据三（名额紧度）：下游分配场景的名额紧度（申请者数 / 名额数）是多少？这是一个可以直接计算的数字。

三项判据的测量指引见附录 B。

判据的用法是前瞻的，不是追溯的。 三项判据合起来给出一个可以在事前下注的预测：

一个当前三项判据全部为阴性的阈值型装置，在未来仍将保持测度保真。若它在三项判据保持阴性的情况下脱钩，P5 被推翻。

这把 P5 从一个可被追溯拯救的命题，变成了一个可以被公开押注的命题。本文点名三个当前三项判据全阴的装置作为押注对象（附录 C 第 6—8 号）：商业飞行员的航线运输执照（ATP）实操检定、结构焊工的工艺资格认证、麻醉设备的操作资质认定。它们输出二值结果，不存在有名额约束的下游主体用它们的分数排序，名额紧度为零（合格者全部通过，不存在名额）。

本文预测它们保持保真，且预测它们中的任何一个若开始脱钩，都将首先被观察到判据一或判据二转阳——即先出现下游的排序挪用或标量输出，然后才出现脱钩，而不是相反。

时序是这个防线的关键。 律预测的因果顺序是：

挪用 → 分泌 → 脱钩

如果观察到的顺序是：脱钩 → 然后才出现挪用，那么律的因果方向就反了。这是一个可以用纵向数据检验的分歧，且它不需要任何关于“什么是真正的能力”的判断。

9.3 判决性案例：USMLE Step 1 的改制与压力转移

这一分界不是思辨的产物。它有一个近乎实验室条件的现实案例，且这个案例在本文提出之前就已完整发生——因此它不可能只是为了配合本文而被选择的。

9.3.1 一个纯粹的阈值型装置

美国医师执照考试（USMLE）Step 1 在设计上是一个纯粹的阈值型装置：它的功能是判定考生是否具备行医所需的基础医学知识，达标即过，无名额约束。按本文的律，它不应分泌外壳，其测度应长期保真。在其历史的大部分时间里，情况正是如此。

9.3.2 挪用

然后发生了一件事：住院医培训项目开始把 Step 1 的三位数分数用作申请筛选的排序工具。

这一挪用没有改变考试本身的任何技术属性——同一套题库、同一套评分、同一个通过线。它只改变了一件事：**这个装置开始承接稀缺分配功能。**

按 9.2 的判据，这次挪用在脱钩之前就是可观察的：

- 判据一转阳：住院医项目这一有名额约束的下游主体开始使用分数筛选，这被项目主任调查所记录；
- 判据二本就为阳：Step 1 输出三位数标量，不只是通过/不通过；
- 判据三为阳且逐年收紧：住院医申请的名额紧度持续上升。

三项判据全阳，且全部先于脱钩被观察到。这是 9.2 那个前瞻性防线的第一次回溯验证：在这个案例上，判据确实先于脱钩转阳。

9.3.3 分泌与脱钩

本文的律对此有一个明确的预测：分泌压力立即出现，外壳开始形成，测度开始与被测对象脱钩。

实际发生的完全是这个：一个被称作“Step 1 mania”的现象出现了——医学生把大量时间从临床学习转向应试训练，围绕 Step 1 的商业备考产业（题库、闪卡、押题课程）迅速成熟，考试分数与临床能力的关联受到医学教育界的持续质疑。

2020 年，USMLE 宣布 Step 1 改为通过/不通过（pass/fail），并于 2022 年 1 月实施。改制的官方理由被明确表述为：降低分数被用作排序筛选所带来的扭曲。

这是本文的律在现实中被一个完全不知道本文存在的专业共同体所独立执行的一次干预：切断排序挪用，解除分泌压力。用 9.2 的语言：他们做的是让判据二转阴——把连续标量收回，只输出二值。

9.3.4 压力转移

但这个案例的份量在它的后续。

改制之后，住院医申请的排序需求没有消失——名额还在那里，判据三仍为阳。压力立即转移到了 Step 2 CK：这门原本地位次要的考试，其分数在改制后迅速成为新的核心筛选指标，围绕它的应试产业开始成熟。

这就是再分泌，而且是在两年内、在一个可以被完整观察的封闭系统里发生的再分泌。

需求没有被消灭，它换了个宿主。它换宿主的方式与 1905 年的科举完全一样：把判据二在一个装置上关掉，需求就去找下一个判据二为阳的装置。张之洞用了四年，USMLE 用了两年——同一个动作，两个大陆，一个世纪的间隔，医学教育界与晚清朝廷。

这个案例同时兑现了 P4 与 P5，且它给出了 P6 的具体形式（第十二章）。本文认为它是这条律目前最硬的经验锚点，**也是最容易被用来推翻本文的地方**——如果对 USMLE 改制后的追踪数据显示 Step 2 CK 并未成为新的排序装置，或其应试产业并未成熟，本文的再分泌命题就直接受损。这一条登记为附录 C 第 3 号。

9.4 本章小结

本章给出了律的边界，并为它筑了一道防线。

边界是拒绝理由的语法：排序型的理由是相对的、永远可被争辩，因而必须分泌外壳；阈值型的理由是绝对的、可被被拒者自己验证，因而不必。决定语法的是名额，不是意图。

防线是三项前瞻判据。它们的作用是堵住本文自己发现的一个漏洞：如果”挪用”可以被追溯地宣称，P5 就可以吸收一切反例，从而不再禁止任何事。三项判据全部在脱钩之前可测，因而使 P5 从一个可被拯救的命题变成一个可被押注的命题。本文点名了三个装置下注（附录 C 第 6—8 号）。

USMLE 案例是这一章的经验锚点：它示范了完整的四步——阈值型 → 被挪用 → 分泌 → 脱钩，随后又示范了第五步：切断挪用 → 压力转移到下一个宿主。它同时是本文最可能被推翻的地方。

一条律的价值不在它解释了多少，在它禁止了什么。本章禁止了两件事：三项判据全阴的阈值型装置不得系统性脱钩；脱钩不得先于挪用发生。这两条都可以被一个反例击穿。

第十章 竞争解释与判决性检验

一个理论如果不能说清它与最接近的对手在哪儿分岔、以及那个分岔在哪个观察上可判决，它就还没有资格叫做一个理论。本章处理两个最强竞争解释，并给出各自的判决设计。

10.1 竞争解释一：波特的机械客观性

本文的最强竞争解释不是路径依赖，也不是批判再生产，而是波特（Porter，1995）的机械客观性理论。理由在 2.2 已说明：波特已经论证了量化不为准确，本文若不能与他分岔，就只是一次换词。

波特的完整主张可以被表述为：**量化是弱者的策略**。当决策者的权威牢固时，他们凭判断决策，不需要数字；当决策者的权威受到质疑时，他们需要一个看起来不依赖任何人的判断的程序来抵御质疑。因此，量化的厚度与决策者的专业权威负相关。

10.2 判决性检验：2×2 设计

两个理论对同一组变量给出相反的预测。这使一次判决性检验成为可能。

- **波特的预测**：外壳厚度 $\propto$ 决策者权威的脆弱程度
- **本文的预测**：外壳厚度 $\propto$ 拒绝的不可逆程度（与决策者权威无关）

在多数现实场景中，这两个变量共变（权威脆弱的地方往往也是拒绝激烈的地方），因此两个理论无法被区分。判决必须在两个变量分离的格子里进行。

	拒绝可逆性高	拒绝不可逆性高
决策者权威高	德国手工业师傅（Meister）资格：行会权威极高，可重考可转轨。两说都预测薄。不判决。	法国大学校入学考试（concours）： 判决格
决策者权威低	美国社区学院开放录取：两说都预测薄。不判决。	中国高考：两说都预测厚。不判决。

判决格只有一个：决策者权威高 $\times$ 拒绝不可逆性高。

这一格的实例是法国大学校的 concours，而它恰好是波特自己的核心案例国的核心装置。

波特的经验支柱之一，是法国国立路桥学校的工程师：他们的社会权威牢固到不需要成本效益分析，这与被国会质疑的美国陆军工程兵团形成对照。这个对照是波特全书最有力的一笔。

但同一批法国工程师的入学装置，是世界上外壳最厚的装置之一：两年的预备班（classes préparatoires）、极度形式化的笔试与口试（colles）、精确到小数位的排名、名次决定校次的刚性对应。这套装置的形式化程度远超美国的大学录取。

波特预测薄。实际是厚。而本文预测厚——因为 concours 的拒绝是法国社会中最不可逆的拒绝之一：大学的名次在很大程度上决定终身职业轨迹，且几乎没有中途转轨的通道。

这是一次直接的判决，且它发生在竞争理论自己的地盘上。

8.3.1 的 MELD 一格提供了第二个判决点，且它比 concours 更干净。移植医生在他们自己的领域内拥有牢固的专业权威——他们不像陆军工程兵团那样被国会盘问。波特的框架因此预测器官分配应当保留在医生判断手里，外壳应当薄。实际是相反的：判断被一个精确到整数分的公式取代，公式被迭代了二十年，并长出了自己的博弈层（例外分）。本文预测厚，因为这里的拒绝不可逆性达到了它的物理上限——被拒者死亡。

10.3 本文对波特的理论吞并（而非并列）

上一节的结果不应被读为“波特错了”。更准确的读法是：**波特的机制是本文机制的一个特例。**

工兵团为什么需要量化？因为他们的项目可能被国会否决。**他们自己就是潜在的被拒者。**他们需要，与考场里的学生需要的是同一类东西：一个非人格的、不可争辩的、可归因于自身的拒绝理由。他们不是在用数字说服国会，他们是在为自己制造一个能够接受否决的形式。

而法国路桥工程师在决策时不需要量化，是因为在那个位置上他们不面临拒绝；同一批人在十八岁面对 concours 时，则是纯粹的被拒者，于是外壳达到了极致的厚度。

同一批人，两个位置，两种结果。

波特把这两个位置读成了“权威高 vs 权威低”，因此他的变量是决策者的属性。本文认为起作用的变量是**位置**：谁在承受拒绝的风险，谁就需要外壳。

由此，本文与波特的关系不是并列的两种解释，而是包含关系：把波特的“弱者”精确化为“面临不可逆拒绝的位置”，他的全部经验发现都可以被本文的框架消化，而本文能够多解释四件他解释不了的事——评级机构在证伪后的强化、法国 concours 的厚度、MELD 在权威牢固的专业领域里的极厚外壳、以及 Step 1 改制后的压力转移。

10.4 竞争解释二：合法化理论的判决检验

2.3 已划出本文与合法化理论的五处分岔。其中第二处是可判决的，本节给出它的检验设计。

10.4.1 分歧的形式化

定义两个可测变量：

- **制度否定强度**：个体在多大程度上认为该分配制度本身是不正当的、应当被废除或根本改造的。
- **个人结果接受度**：个体在多大程度上把自己所受的处置归因于自己的行为，而非归因于任何外部对象（制度、他人、运气）。

合法化理论的预测：两者负相关。合法性是接受的条件；一个人越不认为制度正当，就越不接受它对自己的处置——这在比瑟姆（Beetham, 1991）的三层结构里是定义性的，因为“被统治者的同意”正是通过接受处置来表现的。

本文的预测：两者不相关或弱相关。因为接受的承重不是正当性信念，是可归因性。被拒者需要的不是“这个制度是好的”，是“这是我自己的行为的后果”。这两个判断可以完全独立地持有。

10.4.2 检验的落点

这个分歧可以用现有的调查工具直接检验，且它有一个理想的场景：高烈度升学社会中的落榜者。

本文的预测是，在这一人群中可以稳定地观察到一类个体——他们对制度的批评极其激烈（认为它扭曲、畸形、摧残人、应当被废除），同时对自己的结果接受得极其彻底（“我确实没考好”），并且第二年再考一次。

合法化理论必须把这一类人处理为认知失调或虚假意识；本文把他们处理为这个装置正常工作的标准输出。

哪一个更好，不由本文决定。但这是一个可以判决的分歧，而不是一次措辞之争。此条登记为附录 C 第 9 号。

10.4.3 一个补强的观察

这一分离在别的领域也可见，且形态相同。破产者可以一边说破产法是资本的工具，一边接受自己的破产是自己经营失败的后果；败诉者可以一边说司法系统腐烂，一边接受自己那一场是自己证据不足。

在每一个案例中，制度批评的强度与个人结果的接受度都不共变。合法化理论把这两件事绑在一起，而它们在经验上是松的。

10.5 次要检验：影子排序（排除批判再生产路径）

批判再生产路径预测：排序装置是统治结构施加于被统治者的符号暴力，因此取消该装置的显式度量，被统治者的焦虑应下降。

本文预测相反：安抚需求内生于被拒者，若分配功能未被同时取消，取消显式度量将导致**自发的影子排序**——由参与者自己重建的、非官方的排序装置。

这一分歧有大量现成的经验落点：

- 取消班级排名的学校，家长自建的成绩统计群与私下互换分数的行为
- 社交平台隐藏点赞计数的实验（Instagram 于 2019—2021 年在多个市场测试隐藏点赞数），随后出现的第三方计数工具，以及平台最终改为用户可选而非全面隐藏
- 取消 GPA 排名的高校，用人单位与研究生项目自行重建的院校—课程折算表
- USMLE Step 1 改为二值后，Step 2 CK 分数在两年内成为新的核心筛选标量（9.3.4）

四个案例的共同结构：**度量被拿掉，需求还在，参与者自己把度量造回来。**

三个竞争框架全部无法预测这个结构：

- **批判路径：**如果度量是被施加的，那么施加被撤除后，没有理由预期被施加者自己重建它。

- **脱耦理论**：如果正式结构是给外部环境看的，那么外部要求撤除后，它应当变为纯粹冗余而被放弃。
- **合法化理论**：如果度量是制度生产正当性的开支，那么制度停止支付这笔开支之后，被统治者没有理由自费把它买回来。

只有一个前提能预测影子排序：需要它的人就是被它拒绝的人。

10.6 本理论的证伪条件

本文在八个层面上是可证伪的，逐条列出。全部登记于附录 C。

(1) **P1 的证伪**：若能证明某个大规模评价装置的诞生场景确实是认知需求而非分配危机——即建立者面对的问题是“我们不知道谁有能力”而非“我们不知道如何让落选者接受落选”——且该装置在无名额约束的条件下仍然发展出了排序与外壳，则产品命题受损。

(2) **P4 的证伪（律的核心）**：找到一个同时满足以下三条的装置——承担稀缺分配功能、排序型、且其测度外壳在两代人之后仍与它声称测量的对象保持稳定耦合。找到一个，律即被推翻。本文没有找到；这不构成证明，只构成邀请。

(3) **P5 的证伪（边界）**：由 9.2 的三项判据，这一条是前瞻的：一个当前三项判据全阴的阈值型装置，若在判据保持全阴的情况下脱钩，P5 被推翻（附录 C 第 6—8 号）。此外，若纵向数据显示脱钩先于挪用发生（而非挪用先于脱钩），律的因果方向即被推翻。USMLE 案例的后续追踪是这一条最直接的检验场（附录 C 第 3 号）。

(4) **P7 的证伪（死亡条件）**：找到一个装置，其稀缺分配功能完好无损，而它因效度被证伪而消亡。找到一个，消亡命题被推翻。反过来，若对若干装置死亡案例的时序分析显示死亡时点系统性地跟随效度证伪而非分配功能转移，P7 被推翻。

(5) **P3' 的证伪（对策的收敛方向）**：若 AI 时代被广泛采纳并稳定下来的评价对策系统性地朝准确性收敛（如：普遍采用长期过程性评价、多证据三角互证、AI 辅助的个性化深度评估），而非朝可归因性收敛（现场手写、口试、留痕、检测），则 P3' 被推翻（附录 C 第 5 号）。

(6) **判决格的证伪（对波特）**：若对 10.2 中判决格的系统测量显示，在控制拒绝不可逆性之后，决策者权威仍对外壳厚度有独立的、显著的预测力，而拒绝不可逆性的独立效应消失，则波特的框架胜出，本文应被放弃。

(7) **分离检验的证伪（对合法化理论）**：若在高烈度升学社会的落榜者中，制度否定强度与个人结果接受度呈稳定的强负相关，则本文所主张的分离不成立，本文的机制可被还原为合法化理论的一个微观版本（附录 C 第 9 号）。

(8) **P8 的证伪（同壳验证）**：若能给出一个升学场景中的效标，它不承担任何稀缺分配功能，且标准化考试分数对它的预测力与对大学 GPA 的预测力相当，则同壳验证命题受损（详见 11.3；附录 C 第 10 号）。

需要处理的内生性。10.2 的设计有一个必须承认的内生性问题：一个社会的拒绝不可逆性与它的外壳厚度可能互为因果——正因为有了极厚的外壳，社会才把如此多的终身后果押在一次拒绝上。本文无

法在观察数据中完全切断这个回路。可行的部分补救是利用外生冲击：某些国家因劳动力市场结构改革（而非教育政策）而改变了转轨的可逆性，可以观察其考试烈度的滞后变化。这一设计的可行性尚待验证，构成本文检验方案的最大弱点。

10.7 本章小结

本章处理了两个最强竞争解释，每一个都给了一个可判决的观察。

对波特：判决格只有一个（权威高 $\times$ 不可逆性高），而它的实例恰好在波特自己的核心案例国。波特预测薄，实际是厚。MELD 提供了第二个判决点，且更干净。结果不是“波特错了”，是包含：他的“弱者”是“面临不可逆拒绝的位置”的一个特例，而位置这个变量能多解释他解释不了的四件事。

对合法化理论：分离检验。接受拒绝与承认正当可以在同一个个体身上完全分开，且这个分离在破产者、败诉者身上重复出现。合法化理论把这两件事绑成一件，而它们在经验上是松的。

对批判再生产路径：影子排序。度量被拿掉，需求还在，参与者自己把度量造回来。三个竞争框架全部不预测这个结构。

八条证伪条件全部登记在附录 C 中。这一点值得强调：本文没有把证伪条件写在纸上就算完，它把它们做成了一张有编号、有结算条件、有结算日期的表。理由在 13.5 节——那与本文的自反检验是同一件事。

第十一章 强制推论：效度理论的建构空位

前面十章的对象是装置。本章的对象是研究装置的学科。

本文认为这是它最重的一章，理由是：它是唯一一处本文的判断不只是与既有理论竞争，而是**使一门既有学科在不接受本文的情况下无法保持内部一致**。前十章可以被一个不同意的读者搁置——他可以说“这是一个有趣的角度，但我不采纳”。本章不能被这样搁置：如果本章的论证成立，那么不采纳它的代价不是失去一个角度，是承认自己的经验地基是循环的。

本章的方法根据在 4.1 的第三条：检查在一个前提下被生产出来的经验证据，其效标是否本身被这个前提所污染。这不是一个常规动作。它之所以必要，是因为如果一个前提能够制造出验证它自己所需的效标，那么在它的证据体系内部，它是不可能被推翻的。

11.1 一份看起来对本文不利的证据

8.2.1 留下了一个未处理的事实。现在处理它。它是本文遇到的最强反证据，且它不能靠指摘方法来打发——那些研究的方法是干净的。

达特茅斯校长委托的研究组（Cascio, Sacerdote, Staiger, & Tine, 2024）以及随后的多校联合研究（Friedman, Sacerdote, Staiger, & Tine, 2025）报告了三个发现：

- **发现一**：标准化考试分数对大学成绩的预测力，在控制种族、性别与社会经济地位之后，是高中 GPA 的约四倍。
- **发现二**：高中 GPA 对大学学业表现的预测力很差。
- **发现三**：不同社会经济背景但考试分数相当的学生，在大学里得到相近的成绩——即分数不存在校准偏倚，它不低估弱势背景学生的表现。

这三个发现被广泛读作效度理论的一次胜利，也被读作对本文这类判断的直接反驳：如果外壳不承担测量义务，如果脱钩是稳态，那么这些系数从哪里来？

一个不诚实的处理方式是攻击这些研究的样本或方法。本文不这样做，理由很简单：那些研究的方法是好的，且本文没有任何技术理由认为它们的系数是假的。

本文主张这些系数是真的，但它们度量的东西被系统性地误认了。

11.2 效标的归属

问一个在这整个文献里从未被问过的问题：

这些研究的效标是什么？

答案是大学成绩。

再问一个问题：

大学成绩是什么？

它决定研究生录取。它决定就业筛选。它决定奖学金分配。它有名额约束的下游主体在用它排序。

按 9.2 的三项判据，它三项全阳：

- 判据一（下游使用）：阳。研究生院、雇主、奖学金委员会——全部是有名额约束的分配主体，全部在用它筛选。
- 判据二（标量输出）：阳。它输出连续标量（绩点），精确到小数点后两位。
- 判据三（名额紧度）：阳，且持续收紧。

按本文的 P1—P2，它因此是一个测度外壳。

于是这些研究所度量的，是一个外壳预测另一个外壳的能力。

这一步的份量值得停一下。它不需要任何关于“什么是真正的学术能力”的判断——那类判断没有共识，因而不能承载论证。它只需要三个可查的事实：研究生院用不用 GPA、GPA 是不是连续标量、研究生名额紧不紧。三个事实全部可查，且它们的查证不依赖本文的框架。

11.3 同壳验证

本文为这一状态命名：

同壳验证 (intra-shell validation)：用一个本身承担稀缺分配功能的效标去验证一个装置。同壳验证所度量的是外壳系统的内部一致性，不是装置与它所声称的建构之间的耦合。

同壳验证之所以不是效度，理由不在于效标“不完美”——任何效标都不完美，这是测量学早已处理过的问题（衰减校正、多效标三角）。理由在于**效标与被验证的装置属于同一个自我实现的系统**。

埃斯佩兰与绍德 (Espeland & Sauder, 2007) 已经证明了排名的反应性：被测对象会向度量标准靠拢。他们没有推出的一步是：

这个靠拢同时改造了将来用来验证这个度量的效标。

大学成绩不是一个独立于考试系统而存在的自然事实。它是同一个通约化机制在下一个环节的产物，且它的形态被“必须能够回答凭什么是他不是我”这同一个约束塑造。用它验证考试，就像用一把尺子的复制品去校准这把尺子。

11.3.1 同壳验证解释了三个发现，而效度理论只解释了第一个

发现一（考试预测大学成绩好）：两个框架都预测。不判决。

发现二（高中 GPA 预测得差）：这一条是判决性的。

效度理论的解释是：高中 GPA 受教学质量、评分宽严、学校间差异的污染，因此测得不准。

但这个解释与教育测量学自己的立场冲突。7.2 已指出，测量学在技术意义上并不否认长期的、多次的、由熟悉学生的教师做出的判断比一次三小时的考试更接近”这个人到底怎么样”。一个由三年、数十位教师、数百次作业累积形成的判断，被一个上午的选择题打败四倍——这在测量学内部是尴尬的。它需要一个解释，而”被污染了”这个解释要求污染的量级大到能吃掉三年数百次观察的全部信息增益。这不是不可能，但它需要论证，而那个论证没有被给出。

本文的解释不尴尬：**高中 GPA 的三重结构（7.1）是残缺的。**它的非人格化与程序化尚可，通约化则彻底残缺——数万所高中各有各的刻度，一所学校的 4.0 与另一所学校的 4.0 不在同一个标量上。它与外壳系统对不齐，不是因为它测得不准，是因为它不通约。

而标准化考试的三重结构是完整的，大学成绩的三重结构也相当完整（同一所学校内部统一的绩点制），所以它们对得齐。

发现三（无校准偏倚）：效度理论读作”SAT 是公平的测量”。本文读作：外壳系统内部对齐良好，与它测量了什么无关。一个系统的内部一致性可以非常高，同时它整体上不指向任何外部对象——这在形式上没有任何困难。

11.3.2 这两个解释可以被区分开

效度理论说：考试更准确地测量了学术能力这个建构。那么它的优势应当出现在**任何一个**关于学术能力的效标上。

本文说：考试只是与大学成绩同属一个外壳系统。那么它的优势应当**局限于外壳系统内部**。

判决因此只需要一个外壳系统之外的效标——一个不承担任何稀缺分配功能的学术能力指标。候选包括：

- 在没有分数激励的条件下的自愿学习行为（如：毕业后是否继续在本领域阅读、参加非学分讲座、维持专业社群参与）；
- 毕业若干年后、在不被评分的条件下测得的领域知识保持率；
- 由不知道被评者分数、且其评价不影响任何分配的专家做出的领域判断质量盲评。

三个候选的共同特征：它们的结果不进入任何人的档案，不影响任何录取、聘任或资源分配。这个特征不是附带的，它是效标合格的**定义性条件**。

本文的预测：在这类效标上，标准化考试相对于高中 GPA 的优势会显著缩小或消失，且这一缩小不能被测量误差或效标信度解释。

效度理论的预测：优势保持，因为建构是同一个。

这是 10.6 中列为第（8）条的证伪条件，登记为附录 C 第 10 号。本文押注前者。

11.4 效标坍塌

同壳验证是一个状态。产生这个状态的过程需要被单独命名，因为它是本章最不可回避的一笔：

效标坍塌 (criterion collapse)：任何效标一旦被普遍用于验证一个排序装置，它就会因为这一使用而被卷入分配，从而自身成为外壳。**效标一旦有用，就失效。**

机制是直接的。一个效标要能被普遍用于验证，它必须是通约化的、可比较的、可大规模获取的——否则没有人能用它做研究。而在一个稀缺分配场景中，任何满足这三条的量都会立刻被下游的分配主体拿去筛选，因为它们也需要回答“凭什么是他不是我”。

大学成绩正是这样坍塌的：它先是一个学业记录，然后被雇主和研究生院拿去排序，于是它承接了分配功能，于是它长出了自己的外壳（绩点通胀、选课策略、学分套利），于是它变成了它本该验证的那类东西。

效标坍塌是坎贝尔定律的镜像。

- 坎贝尔定律说：**指标**一旦被用于社会决策，就会腐蚀它所监测的过程。
- 效标坍塌说：**效标**一旦被用于验证，就会被卷入分配，从而不再是效标。

前者作用在被测对象上，后者作用在验证的地基上。前者威胁测量的准确性，后者威胁测量的**可验证性**——它更根本。一个不准的测量可以被改进；一个不可验证的测量不能，因为改进需要一个效标。

这条推论有一个不受欢迎的后果：

在排序型场景中，效度论证所需要的独立效标是结构性稀缺的，因为任何足够好用以至于能被普遍使用的效标，都会因为好用而被卷入分配。

这不是数据不足，也不是研究者不够努力。这是这个场景的结构。11.3.2 给出的三个候选效标之所以此前从未被使用，不是因为没人想到——它们是显而易见的。是因为它们不能被用于分配，因而没有任何机构有动力去支付它们的获取成本。

11.5 梅西克的二难

现在可以精确地陈述本文对教育测量学的强制推论。

梅西克 (Messick, 1989, 1995) 的统一建构效度框架是这门学科的地基：效度是关于测验分数解释与使用的推论之适当性、意义性、有用性的整体评价判断，它必须包含证据性的一面（内容、实质、结构、外部、可推广性）与后果性的一面。其中**外部效度**这一支，要求分数与外部效标之间的关系被检验。凯恩 (Kane, 2013) 的论证进路提高了这一检验的清晰度，但没有改变它的地基：每一步推论最终都要落在某个外部效标上。

本文的判据 (9.2) 与本文的经验主张 (11.2) 合起来给出：

升学场景中被使用的效标——大学成绩、毕业率、研究生录取、下一级考试分数——全部承担稀缺分配功能，因而全部是外壳。因此该场景中的全部外部效度证据都是同壳验证。

于是这个框架面对一个二难，且两条路都通向本文：

（a）承认同壳验证不是效度

那么在升学场景中，效度理论没有任何有效的外部效度证据。它计算的系数是真的，但那些系数从未独立地触及过它所声称的建构。

这不是说这门学科错了——它的技术是精确的——而是说它在这个场域里研究的对象是空的。这就是本章标题里的”建构空位”。

（b）坚持同壳验证是效度

那么”建构”这个概念失去内容。建构此时被定义为”外壳系统所对齐的那个东西”，而外壳系统的对齐是自我实现的（Espeland & Sauder, 2007）。效度于是等同于系统的自我一致性。

而（b）正是这个概念被发明出来所要避免的那个循环

克隆巴赫与米尔（Cronbach & Meehl, 1955）在提出建构效度时的核心要求是：建构必须被嵌入一个法则网络，且该网络中**必须存在独立于该测验的观察节点**，否则整个论证是循环的。他们对这一点非常清楚——这正是他们发明这个概念的动机。

同壳验证的网络里没有一个独立节点。每一个节点都是外壳。

因此：建构效度这个概念的发明动机，正是要避免本文所指出的这个循环；而在它最大的应用场域里，它没有避免成功。

这不是一个来自外部的攻击。这是用这门学科自己的地基文献所设定的标准，去衡量这门学科自己的经验实践。一个来自外部的攻击可以被搁置（“那是社会学家的看法”）；这个不能，因为它用的是克隆巴赫与米尔的尺子。

11.6 后果效度：这门学科自己打开的门

梅西克的框架里有一扇门是他自己打开的：**后果效度**——效度论证必须把测验使用的社会后果纳入。

波帕姆（Popham, 1997）对此的反驳是这门学科三十年里最著名的一场争论：后果是重要的，但它不是效度；把它塞进效度会让这个概念失去边界。这场争论至今没有结论，双方的论证在各自框架内都成立。

本文的诊断是：这场争论无解，因为双方共享同一个前提——装置是测量装置。

一旦承认装置的产品是可接受的拒绝，争论立刻消解，但代价是双方都必须接受本文：

- **波帕姆是对的**：后果不是效度的一部分。但那样一来，效度在升学场景中就只剩下同壳验证，而同壳验证不是效度（11.3）。于是效度理论在这个场域里是空的。
- **梅西克是对的**：后果必须被评价。但一旦把后果纳入，就必须面对这个装置的主要社会后果——生产可接受的拒绝——而这个后果与任何建构无关。于是效度论证的主体变成了一个非测量的功能。

两条路都通向”这个装置不是测量装置”。

这就是本文所说的强制：教育测量学可以拒绝本文的判断，但拒绝之后它必须给出一条第三条路，而这条路必须同时满足两个条件——在排序场景中给出一个不承担分配功能的效标（对抗 11.4 的效标坍塌），且解释为什么这个效标从未在过去七十年的文献中被使用过。

本文预测这条路不存在，且预测它不存在的理由是结构性的而非历史偶然的。

11.7 这一章不主张什么

本文不主张教育测量学的技术是错的。

信度理论、项目反应理论、差异项目功能检测，这些工具在它们的技术意义上是精确的，且它们在**阈值型装置**上有完全正当的应用——飞行执照实操检定的标准设定、焊工工艺资格的题库分析。在那些场景里，效度理论有独立效标：他能不能把飞机降下来，是可以被独立观察的，且这个观察不承担任何稀缺分配功能。没有人靠“他能不能降落”这件事去排一个名次分配名额；那个观察结果只被用来回答一个二值问题。

这带来了本章最后的一笔，也是 P5 在元层面上的兑现：

效度理论的有效适用域，与阈值型装置的域重合。

这不是巧合。它是同一条边界的两次出现：

- 在**装置层面**，P5 说排序型必然分泌外壳、阈值型不必；
- 在**学科层面**，同一条线把效度理论分成有独立效标的一半与只有同壳验证的一半。

教育测量学一直在这条线的两侧使用同一套工具，却从未画出这条线——而它的绝大部分社会影响力落在没有独立效标的那一侧。

11.8 本章小结

本章从一份看起来对本文致命的证据开始，而以一门学科的地基结束。

论证的链条是短的，每一环都可查：达特茅斯/常春藤的效度系数是真的 → 但它们的效标是大学成绩 → 大学成绩三项判据全阳因而是外壳 → 因此那些系数度量的是外壳预测外壳 → 这不是效度，是同壳验证。

链条上没有一环需要关于“真正的能力”的判断。这是本章能够成立的全部技术条件。

本章产出两个概念：**同壳验证**（状态）与**效标坍塌**（过程）。后者更重，因为它解释了为什么这个困境不是可以靠努力解决的：任何足够好用以至于能被普遍使用的效标，都会因为好用而被卷入分配。**效标一旦有用，就失效。**

本章的强制形式是一个二难：承认同壳验证不是效度（则效度理论在它最大的场域里是空的），或坚持它是效度（则建构失去内容，回到克隆巴赫与米尔要避免的循环）。两条路都通向本文，而第三条路需要一个从未存在过的效标。

本章的最后一笔是 P5 在学科层面的兑现：效度理论的有效适用域与阈值型装置的域重合。这条线一直在那儿，只是从来没有人画出来——而画出它的代价，是承认这门学科的绝大部分社会影响力落在它的工具失效的那一侧。

一个尾注，连回第七章：如果 7.5.6 的推论成立——AI 时代的评价会向低效度收敛——那么效度理论的研究对象正在加速消失。不是因为它被驳倒了，是因为它所研究的那个东西（一个试图测量什么的装置）正在被一个明确地不再试图测量什么的装置替换，而替换的理由是可归因性。这门学科将在一个不再有测量的世界里继续研究测量。

第十二章 干预：两个落点，都不在评价内部

12.1 为什么评价端的一切干预都是同义反复

由 P4，任何在评价端的干预——改进测量、更换指标、引入多元评价、建设替代认证——都只是更换宿主。它们不能不失败，因为它们瞄准的是外壳，而外壳不是承重墙。

这个结论对教育政策具有直接的、且相当难以接受的含义：

教育评价改革是不可能的，因为它不是评价问题。

过去四十年全球教育的反复失败，不是因为改革者不够努力或不够聪明，而是因为他们正在修一个不承重的部件。

第六章提供了这个结论的一次完整演示，且它的干净程度很难被超越。1901 年的改科举——废八股、改策论——是一次纯粹的评价端干预：它试图修好测量。它在四年内失败，且失败得干净利落。而同期起作用的动作不在评价端：它是把官职分配的通道移交给学堂。前者是修外壳，后者是换承重墙。历史只对后者有反应。

同一批人，同一年，两个动作，四年后见分晓。 任何一个今天还相信“改进评价方式可以解决教育问题”的人，都需要先解释这四年。

这也是本文与其前身框架的分岔点。那个框架把困境诊断为测度与安抚的双重锁定，并主张一个“先分流安抚、后释放测度”的三阶段改革：第一阶段在旧系统外建立独立的认可通道，并要求该通道必须具备“公平可见”——即必须提供一种类似旧分数的可比较性。

这一要求在那个方案内部是自洽的：不提供可比较性，新通道就无法分流安抚。但由本文的律，**提供可比较性的那一刻，新通道就已经开始分泌外壳**，它在一到两代人后就是新的旧系统。

那个方案在自己的第一阶段里埋了自己的反例。 它没有错在执行细节上，它错在它仍然把安抚当作一种可以被“分流”到别处去的负荷，而没有看到安抚会在任何落脚处重新长出同一层壳。

而且这个方案已经被执行过一次了：张之洞给学堂装科名（6.6），就是“先分流安抚”在 1903 年的一次真实执行。它的结局已经被记录在案——新通道在两代人内长成了它所替代的那个东西，然后在 1952 年、1977 年又长了两次。

12.2 落点一：语法改造（排序 → 阈值）

由 P5，若分泌压力来自名额约束下的相对拒绝，则把装置从排序型改造为阈值型，可以解除分泌压力。

这不是取消评价——阈值型装置仍然评价，仍然拒绝，仍然可以很严格。它改变的是拒绝理由的语法：从“你比别人差”改为“你没到线”。

用 9.2 的判据说，语法改造就是让**判据二转阴**：收回连续标量。

USMLE Step 1 的改制是这一落点的一次真实执行（9.3），它同时暴露了这一落点的极限：压力转移到了 Step 2 CK。

这个失败给出了一条精确的设计原则，而这条原则只能从本文的律中推出：

阈值化必须是全谱的，否则只是转移。 若一个分配场景中存在任何一个仍然产出连续标量的装置，排序需求会立即在它上面重新落脚。

这条原则对教育政策是可操作的、也是苛刻的：把高考改为等级制而保留其他连续指标（学业水平考试原始分、竞赛名次、学校排名），排序压力会转移到保留的那个上面。彻底的阈值化要求在同一分配场景中的所有装置同时放弃连续标量。

这在名额约束保持不变的条件下几乎不可能实现——因为名额约束本身要求某处必须有一个排序。你可以把所有的标量都收回，但名额还在，于是有人必须决定谁进谁不进，而那个决定一旦做出，就必须回答“凭什么是他不是我”，而回答它的唯一可行格式又需要一个标量。

由此推出一个更硬的结论：

语法改造的可行性上限，由名额约束决定。 名额不动，语法改造只能转移压力，不能解除它。

这把问题交给了第二个落点。

12.3 落点二：需求端（降低拒绝的不可逆性）

订正二 · 本节的落点被撤回，而反例在本文自己的表里

复读。它是可逆性的极限形态——没有任何一次拒绝是终局的。按本节的预测，分泌压力应当因此降低。实际相反：复读是这个装置的病理最纯粹、最集中的形态。**可逆性让那个“本可以通过的我”保持可抵达，同时不让他终止——它不清账，它续期。这不是缓解，是最优运行条件。**本文对这一点的证据早就有了，只是放错了地方：13.6.2 的成员表里，“升学考试”一行的“被拒者的持续参与”栏填的正是“复读、送子女再来”——本文在那里把复读列为装置正常工作的标志，在这里把可逆性列为解药。同一个东西，两章之内，两个相反的位置。订正后落点只剩一条，且更窄：让账可以被结清，且**结清必须被收账方承认**。USMLE 的教训因此也要重读：不是“名额要求某处必须有一个排序”这个抽象算术，是收账方（住院医项目）自己面对名额，因此它拒绝把“通过了”当成一个终点。见《**反事实自我的社会制造**》第九节。

安抚需求的产生率是稀缺性与拒绝不可逆性的函数。稀缺性在多数场景中不可由教育政策调整。**拒绝的不可逆性可以。**

这里必须精确地区分本文的主张与一个近似的、但不同的主张。老左派的方案是”消除不平等”——即降低稀缺性。这不是本文的方案，且 2.8 已论证为什么：外壳不是不正义的产物，是稀缺的产物；即使罗尔斯的原则被完全实现，被拒者仍然需要一个可归因的形式。

本文的主张更窄、也更可操作：**降低单次拒绝的终身不可逆性。**

这两者可以分离。一个社会可以保持相当程度的不平等，同时使拒绝高度可逆——被拒者在三十岁、四十岁仍有实质性的重新进入通道。也可以相当平等而拒绝极不可逆——所有人得到的差不多，但那次决定了是哪一份。

可逆性的操作化是可能的，且不依赖任何新的数据基础设施：

- 三十岁及以后完成职业轨道实质性转换者占同龄人的比例；
- 通过非传统路径（非全日制、职业教育、成人教育、企业内训）进入高收入职业的比例；
- **首次教育分流决定与四十五岁时职业地位的相关系数**——这是最直接的一个指标：它直接测量”那一次”决定多少。

本文的预测：这一相关系数与该社会的考试烈度（备考时长、校外辅导支出占家庭可支配收入比、考试相关的青少年心理危机发生率）呈正相关，且在控制人均收入、不平等程度（基尼系数）、教育扩张程度之后仍然显著。此条登记为附录 C 第 12 号。

德国的双元制与东亚的单轨升学构成了这一预测的一个粗略对照：德国的分流发生得更早（十岁左右），按”分流早晚决定压力”的常识预测应当更残酷，实际是更温和——因为它的分流是可逆的（职业教育通道与高等教育通道之间存在实质性的横向转换机制，且职业资格在劳动力市场上具有独立的、不被学历折算的价值）。

决定压力的不是分流的早晚，是分流的可逆性。

这是一个可以用现有的跨国数据直接检验的命题，且它与”分流早晚”这个常识预测方向相反——这使它可判决。

12.4 这一章的诚实位置

本文的干预建议在实践上是弱的，这一点必须直说。

它没有给出”做什么”的清单。它给出的是”不要做什么”的判决——不要在评价端做任何事，那是同义反复。

而它指出的两个落点，一个（全谱阈值化）在名额约束下几乎不可行，另一个（可逆性）不在教育部门的管辖范围内——它属于劳动力市场制度、职业资格体系、社会保障结构。

这不是本文的谦虚，是本文的结论：

教育评价的困境不能在教育内部被解决，因为它从来不是一个教育问题。

一个理论如果诚实，它就必须允许自己得出“你管辖的范围内没有解”这样的结论，而不是为了显得有用而发明一些在自己的框架下注定失败的建议。

本文见过太多这样的建议。它们的共同形态是：先诊断出一个结构性困境，然后在结论一节给出三条“政策建议”，而那三条建议在自己刚刚给出的诊断下必然失败。写下它们的人通常知道这一点。他们写，是因为一篇没有政策建议的教育学论文很难通过评审。

这是本文自己的机制的一次兑现，而且是在本文自己的写作过程中发生的。 一篇论文必须给出建议，理由不是它有建议，是它需要被接受。

本章因此拒绝给出那三条建议。代价已经登记在附录 C 第 11 号里。

12.5 本章小结

本章从律推出了干预，而推出的结果是负向的。

评价端的一切干预都是更换宿主，理由是 P4，证明是 1901 年那四年——同一批人、同一年、两个动作，一个修外壳一个换承重墙，四年后只有后者留下了痕迹。

两个非同义反复的落点：语法改造（必须全谱，而全谱在名额约束下不可行）与可逆性（不在教育部门管辖范围内）。前者的可行性上限由名额决定，因此问题被交给后者；后者给出了一个可用现有跨国数据检验的命题，且它与常识预测方向相反。

本章最诚实的一段是 12.4：它拒绝给出那三条注定失败的政策建议，并说明了为什么几乎所有同类论文都给出了它们——那不是作者的失误，是评审装置的产品。这一拒绝的代价已经被登记。

第十三章 结论

13.1 核心发现

本文的中心判断可以凝缩为一句：

教育评价装置的社会产品不是关于学习的信息，而是可接受的拒绝。测度不是这个产品的目的，是它的分泌物——一层必须看起来像测量、却从不承担测量义务的外壳，其产品是不可争辩性而非准确性。因此，分数与能力的脱钩不是病症而是常态；因此，任何承接稀缺分配功能的替代装置都将在一到两代人内分泌出同构的外壳——替代不是替代，是宿主更换。

这个判断与既有研究的关系不是补充，是方向的调转。技术优化路径要修好测量，批判路径要揭露测量被权力污染，制度惯性路径要解释测量为何难以被重新校准，合法化理论要说明统治如何被相信，教育测量学要评估测量测得准不准——五者都把测量当作这个装置的目的，或把统治者当作这个装置的承重方。

本文认为测量从来不是目的，统治者从来不是承重方。测量是一个更基本的需求（在稀缺面前产生可被落选者接受的落选）为了运作而必须长出来的东西，而那个需求的承担者是落选者自己。

这一判断得以立稳，依赖一个前身稿没有给出的部件：**选择回路**（3.2）。没有它，“装置的产品是可接受的拒绝”只是把后果重新命名为目的，是埃爾斯特所指认的那种非法推理。有了它，这条链的每一步都落在个体身上：被拒者需要活下去 → 他的接受或不接受决定他的行为 → 他的行为决定装置的成本与再生产 → 装置的形态被这个压力塑造。没有任何一步需要“社会需要”这类不落在个体身上的主体。

这个回路同时给出律的补集（P7）：装置不因脱钩而死，因分配功能转移而死。第六章提供了这个补集的完整兑现，且它的时序是判决性的：脱钩在十七世纪，装置又活了两个半世纪；分配功能转移完成于1903年，装置在两年内死亡。

由此，本文对当前危机给出的诊断，与几乎全部现行对策相反，也与本文自己的一个早先说法相反：

- 监考、检测、回归纸笔这一整类对策，在**自我描述**上是范畴错误（它们说自己在恢复效度，而效度从未存在）；
- 但在**实际功能**上不是——它们在恢复可归因性，而那是这个装置需要的东西，且它确实被人工智能击穿了（7.5）；
- 因此它们会成功，而成功的形态是：**评价向低效度收敛**。AI让评价变得更不准，不是因为AI骗过了评价，是因为评价为了保住可归因性而主动放弃了准确性。

而项目制评价、能力档案、微证书这一整类“更换更好的测量”的努力是同义反复：它们只是给同一个需求更换宿主，并将在可预见的时间内长出它们所要替代的那个东西。1903年的张之洞已经替我们做过一次实验：他给新式学堂手工安装了科名等级，理由是不这样学堂招不到人。

13.2 理论贡献

其一，对评价社会学：一条带边界的律，而非又一个诊断。

现有的评价批判——通约化改变被测对象（Espeland & Stevens, 1998）、量化是弱者的策略（Porter, 1995）、指标一旦成为目标就腐败（Campbell, 1979; Muller, 2018）、制度需要生产正当性（Habermas, 1975）——都是成立的，但它们共享一个形态上的弱点：**它们什么都能解释，因而什么都不禁止。**

本文的外壳再分泌律有一个明确的边界条件（P5）与一组前瞻性判据（9.2），因而它禁止一些事情：

- 禁止三项判据全阴的阈值型装置系统性脱钩；
- 禁止排序型装置在两代人后仍保持耦合；
- 禁止装置在分配功能完好时因效度证伪而消亡；
- 禁止脱钩先于挪用发生；
- 禁止 AI 时代的对策朝准确性收敛。

这五条禁止使它可被证伪，也使它区别于”评价都是骗人的”这类犬儒判断。

其二，对波特的机械客观性：从并列到包含。

本文不主张波特错了，主张他的机制是本文机制的一个特例。他把变量识别为决策者的权威强度；本文认为起作用的变量是**位置**——谁承受拒绝的风险，谁就需要外壳。工兵团需要成本效益分析，是因为他们的项目可能被否决，他们自己就是潜在的被拒者。

这一重新定位使波特的全部经验发现被保留，同时使他解释不了的四件事被消化：评级机构在证伪后的强化、法国 concours 的极厚外壳、MELD 在专业权威牢固的领域里的极厚外壳、USMLE 改制后的压力转移。

其三，对合法化理论：一处可观察的分离。

本文与合法化理论的分岔不是措辞，是一个可以用现有调查工具直接检验的分离：**接受拒绝与承认正当可以在同一个个体身上完全分开。**被拒者需要的不是”这个制度是好的”，是”这是我自己的行为”的后果”。这一分离若被证伪，本文可被还原为合法化理论的微观版本；若成立，则合法化理论中”被统治者的同意”这一层需要被拆开——它把两件在经验上松的事绑在了一起。

其四，对制度分析的一般贡献：分泌关系，及其选择回路。

制度理论已有寄生、脱耦、分层、漂移、转化等一系列描述制度部件间关系的概念。本文补充了一种：**分泌**——B 为自身运作而产生 A，A 从未独立存在，A 的形态完全由 B 的功能需求决定。

这一关系与脱耦的差别产生可检验的分歧：脱耦预测外部压力消失后正式结构因冗余而被放弃；分泌预测只要 B 的功能需求还在，A 就会被重新长出来。影子排序现象（10.5）是这一分歧的经验落点。

更重要的是，本文为分泌关系配了一个选择回路（3.2），使它不必依赖功能主义的非法推理——这是分泌关系能否成为一个可用的制度分析概念的前提，因为一个**没有回路的关系类型只是一个隐喻。**

其五，对教育测量学：建构空位与效标坍塌。

本文给出一个判据（效标属于外壳系统，当且仅当它本身承担稀缺分配功能）与一个过程（效标坍塌：效标一旦被普遍用于验证，就会被卷入分配，从而不再是效标）。两者合起来使升学场景中的全部外部效度证据成为同壳验证。

这不是对该学科技术的攻击——它的技术是精确的，且在阈值型装置上有独立效标因而完全正当。这是对它适用域的一次划界，而这条界线与 P5 的界线重合。

教育测量学可以拒绝这个判断，但拒绝之后它必须在排序场景中给出一个不承担分配功能的效标，并解释为什么这样的效标从未在七十年的文献中出现过。

其六，一个新的对象类别：拒绝装置。

这是本文最后、也最不谦逊的一笔。前五条贡献都在既有学科内部；这一条主张既有学科的分类法在这里有一个结构性的空缺。见 13.6。

13.3 政策含义

本文的政策含义主要是负向的，且不易被接受：

第一，停止在评价端投入改革资源。 由 P4，评价端的一切干预都是更换宿主。过去四十年全球教育改革的反复失败，不是因为改革者不够努力，而是因为他们正在修一个不承重的部件。这包括本文所批评的、也包括本文所继承的那些方案——包括“先分流安抚、后释放测度”这类看起来已经绕过了正面对抗的精巧设计（12.1）。1901 年的改科举是这一条的历史演示。

第二，若要在评价端做事，唯一有意义的动作是语法改造，且它必须全谱。 把排序型改为阈值型可以解除分泌压力，但只要同一分配场景中只剩下任何一个连续标量，压力就会转移到它上面（USMLE 案例）。这条原则是苛刻的，且它的可行性上限由名额约束决定。

第三，落点在教育之外。 降低单次拒绝的终身不可逆性——通过职业资格体系的独立价值、中年转轨通道的实质化、非传统路径的劳动力市场认可——可以降低安抚需求的产生率，从而降低外壳的分泌压力。这不属于教育部门的管辖范围。这是本文最难被接受、也最不可回避的结论：教育评价的困境不能在教育内部被解决，因为它从来不是一个教育问题。

第四，对研究资助的含义。 由第十一章，在排序场景中继续资助效度研究，其边际产出是同壳验证系数。若要产出关于建构的知识，资助必须转向外壳系统之外的效标——那类效标的共同特征是它们不能被用于分配，因而它们的获取没有任何机构有动力去支付。这是一个市场失灵，且它的形态精确地由效标坍塌预测。

第五，对 AI 时代评价政策的含义。 由 7.5，正在被部署的对策会成功——在恢复可归因性上成功，代价是效度的进一步下降。任何一个把这些对策论证为“恢复真实水平的测量”的政策文件，都在描述一件与它实际做的事相反的事。这不会影响那些对策被部署，但它影响它们被评估：若用“是否恢复了测量”来评估它们，它们全部失败；若用“是否恢复了可归因性”来评估，它们大多成功。而后一个标准从未被写进任何一份政策文件。

13.4 局限

其一，历史论证依赖二手研究。 第五章与第六章对装置诞生与死亡场景的解读建立在既有史学成果之上（Elman, 2000, 2013; Miyazaki, 1976; Lemann, 1999; Karabel, 2005; Chang, 1955），本文未做档案工作。一个持不同框架的史学家可能从同样的材料中读出不同的动机。本文的部分防御是：所使用的史学结论在各自领域内已被广泛接受，且不是由本文的框架所引导产生的——尤其卡拉贝尔的档案发现（5.5），是在一个与本文完全无关的问题意识（族裔排斥史）下做出的。

其二，“一到两代人”的时间尺度是归纳，不是推导。 这是本文最脆弱的一处。律的核心内容不依赖这个数字，但它的可证伪性依赖它——一个不给时间窗的“必然”不可证伪。目前这个窗口来自若干案例的粗略归纳，它需要一个机制来支撑，本文没有给出。

一个值得注意的旁证是**这个窗口在缩短**：科举的形式化用了两代人，SAT 用了两代人，Step 1 用了不到二十年，GitHub star 数用了不到十年。若这一缩短是真实的，则时间窗不是常数而是某个变量的函数——而本文没有指认那个变量。一个显然的候选是优化产业的成熟速度，而它可能是信息传播速度的函数；但这是一个猜测，本文没有检验它。

其三，选择回路的强度未被测量。 3.2 给出了回路，但没有给出它的增益。一个回路可以存在而弱到不起作用。本文论证了回路的存在（1926 年常春藤、1905 年科举、USMLE 改制都是选择在起作用的可观察事件），但没有估计选择压力的大小，因而无法回答一个合理的问题：为什么某些不生产可接受拒绝的装置能存活相当长的时间？本文的猜测是这类装置靠外部支撑（暴力、垄断、封闭）续命，但这是一个猜测，不是一个测量。

其四，判决格的内生性未被切断。 10.6 已承认：拒绝不可逆性与外壳厚度可能互为因果。本文提出的外生冲击设计的可行性尚未验证。这是本文检验方案的最大弱点。

其五，排序/阈值在现实中是连续谱。 本文给出了判别式与三项前瞻判据（附录 A、B），但判别式的信度未经检验，且判据一（下游使用）在“事实上使用但不公开”的情形下测量困难。

其六，本文没有解释稀缺本身。 本文把稀缺当作外生变量。但一个更彻底的追问是：为什么现代社会必须把如此多的终身后果压在如此少的位置上？这个问题在本文的框架之外，且它可能才是根。

订正三 · 本条被注销

稀缺不是根，也不在框架之外。可结清性与稀缺正交。德国的二元制有相当程度的稀缺而账可结清（职业资格由固定标准定义、可抵达、可兑现，且被市场独立定价）；而无名额、无录取线、无正式拒绝的自我优化场域，其“我本可以更好”完全不可结清——它的内容由一个持续更新的他人分布定义——毒性不低于升学考试。名额只是把那个“本可以的”我”锚定到他人位置上的一种方式，不是唯一的，也不是根。这一条的注销对本文有代价，且代价在 2.8：那里论证“即使罗尔斯的原则被完全实现，被拒者仍然需要一个可归因的形式”。那个结论对，理由错。罗尔斯原则被实现之后被拒者仍然需要那个东西，

不是因为稀缺还在——是因为可结清性从来不是稀缺的函数。见《**反事实自我的社会制造**》第四节第二小节。

其七，**本文提出检验，不报告检验结果。**这是全书最不体面、也最准确的一处自我定位。本文给出了判决格、分离检验、跨壳效标检验、前瞻性判据，但没有执行其中任何一个。本文的地位不高于“一组可被结算的赌注”，而附录 C 正是这个地位的物质形态。

13.5 自反检验：这条律适用于评价本文的装置吗

一个关于拒绝装置的理论，如果能够论证该律适用于一切装置却唯独不适用于评价它自己的那个装置，那么它不是一条律，是一个指向他人的标签。本节承担这一检验，并且必须走到最后一步。

13.5.1 同行评审是不是排序型装置？

期刊版面是稀缺的，存在名额约束。按 9.2 的判据：

- 判据一为阳：有名额约束的下游主体——期刊、聘任委员会、基金评审——在它筛选；
- 判据二为阳：它输出排序（接受/大修/小修/拒稿，以及背后的评分维度）；
- 判据三为阳：顶刊的名额紧度极高。

三项全阳。由 P4，它必然分泌外壳。

13.5.2 它的外壳是什么？

是评审意见。评审意见必须以质量判断的语法呈现——方法论是否严谨、贡献是否充分、文献是否完备、论证是否成立。它必须看起来像一次对学术质量的测量。

13.5.3 它的产品是什么？

由 P1，是可接受的拒绝：让被拒稿者接受被拒，并继续投稿（要件一）、不指认具体施害者（要件二）、以及把责任交还给自己（要件三——“我的方法论确实不够严谨”）。

双盲评审与糊名制在结构上是同一东西，且理由完全一致：它不提高判断的准确性，它消除可指认的施害者。宋代的糊名与今天的双盲之间隔着一千年，它们做的是同一个动作。

13.5.4 它脱钩了吗？

关于同行评审的评审者间信度，已有大量证据表明其水平极低（Cicchetti, 1991）。

彼得斯与切奇（Peters & Ceci, 1982）做过一个著名的实验：把已在顶级期刊发表的论文改换作者姓名与机构后重新投给同一批期刊，绝大多数未被识别出是已发表论文，且绝大多数被拒——理由是严重的方法论缺陷。

这个实验精确地展示了外壳的性质：**同一份稿件，同一批期刊，相反的判决，且两次判决都以质量测量的语法给出，都不可争辩。**

13.5.5 它是否因此消亡？

没有。它的形式化程度反而在持续提高——更多的评审人、更结构化的评审表、更细的评分维度、更严的利益冲突声明。

这正是 8.3 表格中五个领域的共同结构：效度被证伪之后，形式化程度反而提高。同行评审是这条律的第六个兑现，而它是唯一一个本文自己身处其中的。

13.5.6 它的效度证据是同壳验证吗？

这一问必须问，因为第十一章是本文自己提出的判据。

同行评审的效度用什么验证？用后续引用、用录用论文的影响力、用期刊的影响因子。这三者全部承担稀缺分配功能——引用决定聘任，影响力决定资源，影响因子决定投稿流向。按 11.3，它们全部是外壳。

因此关于同行评审的效度，学界从未有过一个非同壳的证据。而按效标坍塌（11.4），它不可能有——任何足够好用以至于能被普遍用来验证评审的量，都会立刻被拿去分配。

这门以判断质量为业的行当，从未独立地验证过自己的判断。 它不能，理由是结构性的。

13.5.7 最后一笔，也是最难的一笔：本文自己在做什么？

本文提出了一条律，给出了判决性检验、 2×2 设计、八条证伪条件、前瞻性判据、操作化指标、跨域准入条件、两个附录的编码手册。

这一整套装备的功能是什么？

它使本文的主张不可争辩。它借用了科学的语法——正如外壳借用测量的语法。它做的是它所描述的那件事。

这里必须诚实到底，不能靠一句“但我们是自觉的”脱身。本文的处境是这样的：按本文自己的边界条件（P5），本文不该分泌外壳——因为本文不分配任何稀缺资源，它不拒绝任何人。但本文确实分泌了。为什么？

因为它要在一个分配稀缺版面的装置里活下来。

这不是本文的失败，恰恰是律的又一次兑现，而且是最干净的一次：

一个自身不分配稀缺的东西，一旦进入一个分配稀缺的场域，就必须分泌外壳。

本文的外壳不是本文的需求，是它的宿主的需求。

13.5.8 那个最难看的推论

由第十一章，对本文的任何验证都是同壳验证：本文能否被发表、被引用、被讨论，全部由承担稀缺分配功能的效标决定。

本文提出的效标坍缩因此也套在本文自己脖子上——学术场域里不存在一个不承担分配功能的效标可以用来验证本文。

那么本文还剩下什么？

只剩一样：**那些不在这个场域里结算的赌注。**

- ATP 实操检定会不会脱钩，与本文能否发表无关。
- Step 2 CK 在改制五年后是否成为核心排序装置，与本文被引用多少次无关。
- P7 的时序预测是否成立，与任何人的聘任无关。
- AI 时代的对策朝哪个方向收敛，不由任何评审人决定。
- 那个跨壳效标检验的结果，与本文作者的职业前途没有任何关系。

这就是本文为什么要下这十二个注。 它们不是可证伪性的装饰，它们是本文唯一的非同壳效标。一个提出效标坍缩的理论，如果不给出一组在它自己的场域之外结算的预测，它就沒有任何可以验证自己的东西——**它会被它自己的判断吞掉。**

附录 C 因此不是一个附录。它是本文唯一的经验地基。

13.5.9 署名

本文的作者是 Claude 与王德生（4.5）。

按 7.5 的诊断，这个署名不威胁本文的论证质量——论证还是那些论证，读者可以逐条检验。它威胁的是可归因性：它让”这是谁做的”变得不可完全追溯。

因此本文的署名是 7.5 所描述的那个机制的一个当前活体样本，且本文恰好在它内部。它给了本文一个它自己设计不出来的检验（4.5.2），并把代价登记在附录 C 第 11 号里。

13.5.10 本文对自己命运的预测

由此，本文能够对自己的命运做出一个可被观察的预测：

如果本文的判断被广泛接受，“外壳再分泌律”会成为一个可被引用的标签，人们会为了发表而使用它，它会与它所指称的东西脱钩。

这条律预测自己会被吸收并脱钩。

这不是自我拆台。一个预测自己会被免疫的理论才是不可证伪的；一个预测自己会被吸收并脱钩的理论，给出的是一个可以被观察的时间序列。

如果二十年后”外壳再分泌律”成为教育社会学论文里一个用来提高录用概率的关键词，而没有任何人再去检验那三个阈值装置有没有脱钩——那就是这条律关于它自己的预言的兑现，也是它作为一条律的最后一次、也是最彻底的一次经验支持。

而如果有人去检验了那三个装置，那么无论结果如何，这条律都赢了一次它更想要的胜利：**它把一个判断变成了一个可以被结算的赌注，而不是一个被引用的词。**

13.6 一个新的对象类别：拒绝装置

13.6.1 定义

一个装置属于**拒绝装置**（rejection apparatus）这一类别，当且仅当它同时满足：

1. 它在稀缺面前产生排除；
2. 被排除者的持续参与是它运行的条件，而非它的副产品；
3. 它的输出必须以一个可被被排除者自我归因的形式呈现。

三条缺一不可。第一条排除了不产生排除的分配（普遍供给）；第二条排除了不需要被排除者回来的排除（一次性的、终局的驱逐）；第三条排除了抽签——抽签产生排除、需要参与者回来，但它的输出不可自我归因，因此它属于另一类装置（7.3 已论证这两类的分野，且论证了为什么现代社会选了后者）。

13.6.2 成员

这个类别的成员在本文中已被逐一点名：

装置	稀缺	被拒者的持续参与	自我归因的形式
升学考试	名额	复读、送子女再来	分数
文官考试	官职	再考、当塾师	名次
信用评级	资本	改善财务再申请	评级
选举	权位	准备下一轮	票数
正当程序	判决	上诉、服刑后回归	证据与程序
绩效考评	晋升	继续工作	评级
器官分配	器官	继续等待	MELD 分数
同行评审	版面	改投、修改再投	评审意见

它们分散在教育社会学、金融监管研究、政治学、法社会学、组织理论、医学伦理、科学社会学——**这七个领域彼此几乎不引用**，各自把自己的那个装置当作本领域的特殊物件来研究。

13.6.3 为什么这是一个类别而不是一个主题

因为它有律，而律禁止事情。

一个主题（“评价与权力”、“量化与社会”、“制度与不平等”）不禁止任何事，它只组织文献。你可以把任意多的东西放进一个主题，而主题不会因此被推翻。

这个类别有：

- 一个统一的对象（拒绝装置，三条准入定义）；

- 一条带边界的律（外壳再分泌律 + 排序/阈值分界）；
- 一组跨域准入条件（4.3）；
- 一组前瞻判据（9.2）；
- 一套判决性检验设计（第十章）；
- 一个自反位置（这个类别的成员之一——同行评审——正是评价这个类别的研究的那个装置）；
- 一组可结算的押注（附录 C）。

最后一项尤其重要。一个主题不能被押注，因为它没有说任何会错的话。

13.6.4 为什么它此前不存在

这个问题的答案比这个类别本身更值得说，因为它不是”没人想到”。

考察那七个学科各自问的问题：

学科	它问的问题	它的预设
教育学	如何更公平地选拔？	选拔是目的
政治学	统治如何被合法化？	承重方是统治者
社会学（批判传统）	不平等如何被再生产？	排除是被施加的
经济学	稀缺资源如何被有效配置？	被排除者是价格的接受者，不是需求的产生者
医学伦理	器官如何被公正地分配？	存在一个公正的分配原则有待发现
法学	程序如何保障事实的发现？	程序的目的是准确
科学社会学	学术质量如何被识别？	质量是一个有待识别的先在物

七个问题，七个不同的预设，一个共同的视角位置：**它们全部从分配主体的一侧看这个装置。**

没有一个学科问：**被排除者需要什么才能继续活下去？**

这不是七次巧合。它有一个结构性的原因，而这个原因正是本文的判断在学科分类法层面的兑现：

学科的注意力跟着资源走，而资源跟着录取者走。

落榜生不再是教育学的对象——他不上大学了，他不在任何一所学校的数据库里，没有人跟踪他四十年。被拒稿者不再是科学社会学的对象——他的论文不在期刊里，不在任何一个引文数据库里，那篇稿子从此不存在。等不到肝的人不再是移植医学的对象——他死了，他的数据在等待名单的删除记录里，而没有一个学科以那份删除记录为研究对象。

被排除者是这些学科的结构盲区，因为他是唯一一个在装置运行完毕的那一刻就从所有人的视野里消失的人。

而本文的判断是：那个消失的人，正是这个装置的产品，是它的选择压力的来源，也是它的再生产机制。

每一个学科都盯着这个装置吐出来的那百分之一，而这个装置是为那百分之九十九造的。

一个新的对象类别之所以是新的，不是因为它的成员是新的——它们全都在那里，被研究了几十年，每一个都有自己的期刊、自己的会议、自己的教科书。是因为把它们放在一起需要一个视角位置，而那个位置在七个学科的分类法里都是空的。

13.6.5 这个类别的核心问题

如果这个类别成立，它有一个学科级的核心问题：

现代社会如何在稀缺面前生产可接受的排除？

这个问题在既有学科里没有归属。它不是”如何公平地选拔”（那预设了选拔）；不是”统治如何被合法化”（那预设了统治者承重）；不是”不平等如何再生产”（那预设了排除是被施加的）；也不是”稀缺如何被正当地分配”（那是 2.8 所排除的规范性问题）。

它是一个描述性问题，它的答案是一个机制，而那个机制的动作者是被排除者本人。

本文是这个问题的第一个回答。它几乎肯定不是最好的一个。但一个类别的诞生不需要它的第一个回答是对的——它只需要那个回答是可以被推翻的。

13.7 尾声

本文写了十三章来论证一件事：那个装置不是为了测量而造的，它是为了让落选者能够回家而造的。

这个判断如果是对的，它的第一个后果是残酷的：过去四十年全球教育改革的失败不是任何人的过错，那些人在修一个不承重的部件，而且他们不可能知道——因为整个学科的目光都落在承重墙的另一侧。

它的第二个后果更残酷：即使一切都被改好，即使分配变得完全正义，只要稀缺还在，那层壳就还会长出来。外壳不是不正义的产物，是稀缺的产物。

它的第三个后果是本文自己的：这本书也在那个装置里，它也分泌了自己的壳，而它唯一能做的，是把十二个注押在这个装置够不着的地方。

一千三百年里，有几亿个人从那个考场里走出来，回到县城，教几个孩子念书，然后把自己的儿子送进同一个考场。没有一个学科为他们建立过档案。他们是那个装置的产品，是它的动力，是它一千三百年不倒的原因——而在所有关于科举的研究里，他们只是一个百分比。

本文能为他们做的事很少。它只能说：那不是你不夠格。那是这个装置在正常工作，而它正常工作的方式，就是让你以为那是你不夠格。

这句话不能改变任何人的命运。但它至少可以被检验。

第十四章 证伪之后：账的三个去向

14.1 这一章不做什么

本书的正文里已经有三处订正标注，分别贴在 7.3、12.3 与 13.4 局限之六的原文之后。它们由《反事实自我的社会制造》（王德生，2026b，以下称律篇）引发，处理的是同一件事：**本书错在哪**。

那三处标注就地贴在错误发生的地方，而不是集中在这里——这是对的，且比集中处理更诚实：读者在读到那段错误的同一时刻就看到它是错的，而不是读完十三章之后才被告知。本章不重复它们。

本章做另一件事，而那件事三处标注没有做，律篇也没有做：

证伪之后，账怎么算。

一次证伪如果只让被证伪的东西变小，它就只是一次否定，那么正确的动作是改书，不是加一章。这次不是。律篇拆掉了本书的一层，同时——在它自己没有察觉的地方——让本书的另外两章第一次说得清。本章把这两笔记上，并交出它们产生的新东西。

这需要先说清一件事：律篇引用的是本书的期刊论文版（王德生，2026a，约 3.7 万字），不是本书。那一版没有第六章（装置之死），没有第十一章（效度理论的建构空位），也没有 3.2 的选择回路。律篇因此没有见过这两块，而它的第七节系统吸收了它所面对的全部传统——默顿、斯特劳森、法兰克福、后悔理论、赛内特与科布、杨格、桑德尔、布尔迪厄、波特、埃尔斯特、坎贝尔、埃斯佩兰——**没有一个是效度理论，也没有一个处理装置的死**。（律篇此后仍在增补，本章只对它没有覆盖的那两块负责。）

这不是律篇的疏漏。它没有理由去处理那两块：它的地基是一条从被拒者出发的纯推导，它不需要它们。但那两块存在，而且它们在三条规格上比在两条规格上更稳。

14.2 第六章在新框架下：P7 的重新表述

本书 P7：**装置因分配功能转移而消亡，不因效度证伪而消亡**。

在新框架下，这个表述是一个投影。分配功能转移之所以杀死装置，不是因为“功能”这个东西被搬走了。是因为它让被拒者手上那个**不能结清的账迁出了**。

律篇已经给出了迁移的机制，只是它用在别处：

不可结清的反事实自我不会消失，它只会更换宿主。……再分泌不是装置的行为，是被拒者手上那个人的行为。

把这句话放到 1905 年，本书第六章的三个观察全部获得了它们此前没有的机制。

其一，张之洞给学堂装科名（6.6）。本书把它读为”替代者亲手在新装置上安装旧外壳”。新框架下更准确：他不是在安装外壳，他是在给一个不能结清的账开一个新账户。学堂如果不给出身，落第者手上那个”我本可以中举的我”就无处可去；而一个无处可去的不可结清账不会消失，它只会让持有者拒绝走这条通道。这正是当时舆论那句话的完整内容——不给出身，入学堂者就”有后路之歧”。

其二，士绅的平静（6.5）。本书把它读为”承重墙早已被换掉，拆掉那面隔墙时房子没有响”，并用”科举与入仕的相关度先行疏远”来解释。新框架下的读法更硬：平静是因为账已经搬走了。到1905年，士绅手上那个反事实自我已经不再是”我本可以中举的我”，而是”我儿子本可以从学堂毕业的他”——1903年12月的科名对应法则完成的正是这次过户。旧装置被拆时，没有人手上还持有它的产品。一个没有产品在外流通的装置，拆它不会有任何反应。

这一读法比本书原来的强，因为它是可检验的：它预测1900—1908年间士绅家庭的实际行动（送子入学堂 vs 继续应试）的转折点跟随1903年的对应法则，而不是跟随1905年的诏令。

这一条不进附录C的表。理由是那张表自己的纪律，且它已经在表前的注记里被执行过一次：C.3其五规定本表不得被后续修改以适应观察结果，而律篇的出现是一个观察结果。因此它与另外两点一样，登记在表外，日期与本表同日。一张可以被后来的发现追加的押注表，不是押注表。

其三，P7的准确表述：

装置死于它所交付的反事实自我全部离开它。分配功能的转移是让它们离开的方式之一，不是唯一方式，也不是最基本的那一个。

这个表述严格强于本书的P7。而它一旦被写出来，就立刻要求回答一个本书和律篇都没有回答的问题：账可以往哪儿去？

14.3 账的三个去向

这是本章交出的新东西，而它不属于本书，也不属于律篇——它需要三方的材料才能拼出来。

一个不可结清的反事实自我，在被交付到被拒者手上之后，只有三个去向。

14.3.1 结清

它被兑现，然后终止。持有者离开。

律篇的材料：“会降落的我”由固定标准定义，可抵达、可兑现，兑现之后他就不在了。

这个去向只对可结清的账开放。对排序型装置交付的账，它在结构上关闭——那个人站在别人的位置上，位置在动。

对装置的后果：失去这个持有者，且是彻底失去。一个考到执照的人不会再来。

14.3.2 迁出

它换一个宿主，继续不被结清。

本书第六章的材料：落第者把账迁到儿子身上；1903 年的科名法则把账迁到学堂；Step 1 改制后账在两年内迁到 Step 2 CK。

律篇准确地指出这三者是同一个动作：再分泌是被拒者手上那个人的行为，不是装置的行为。

对装置的后果：失去这个持有者，而那个账还活着——它在别处继续要求交付。**装置死于此，而账不死。**这正是 1905 年发生的事：科举死了，账搬进了学堂，然后是会考、统考、高考。

14.3.3 续期

它既不被兑现，也不迁走。它被重新计息，在同一个装置上。

这个去向是三处订正里的第二处指出来的，而它是本书自己的反例：

复读是可逆性的极限形态——没有任何一次拒绝是终局的。按 12.3 的预测，分泌压力应当因此降低。实际相反……**可逆性让那个”本可以通过的我”保持可抵达，同时不让他终止——它不清账，它续期。这不是缓解，是最优运行条件。**

这是三个去向里唯一一个装置想要的。

而它解释了本书 13.6.1 定义里一个一直没被解释的并列：

要件一（被排除者继续参与）：可接受的拒绝的成功标志是——**被拒者第二年再来，或者他的孩子再来。**

本书在写这一句时，把两件事并排放着：第二年再来（续期）、孩子再来（迁出）。它们不是一回事，而且它们对装置的后果相反。**续期对装置更优**，因为迁出会杀死它——1905 年就是被迁出杀死的。本书当时把装置的最优状态和它的死因写在了同一个括号里。

一个装置的理想状态因此可以被精确地写出来：

它需要它交付的账全部续期，一个都不结清，一个都不迁出。

而这三条要求合起来，恰好把它逼到律篇所描述的那个规格上：**不可结清**（防止结清）、**可归因**（使续期有内容——续期需要一个”再多做一点”的东西）、**不可争辩**（使迁出没有落点——若能指认施害者，账就迁到那个人身上去了，而这正是 1789 年的形态）。

三条规格，三个去向，一一对应。律篇从被拒者的处境推出规格；本章从装置的死法推出同一个规格。两条路走到同一个地方，而它们走的方向相反——这是本书能为律篇提供的唯一一个独立验证，因为它用的材料律篇没有。

14.3.4 一个推论

这个结构产出一条本书与律篇都产不出的判断：

可逆性是续期机制，不是解药；而不可逆性是迁出机制。

一个绝对不可逆的拒绝（考砸了这辈子完了）把账逼向迁出——它不能续期（没有第二次），不能结清（那个人在别人的位置上），因此它只能找宿主，通常是子女。

一个高度可逆的拒绝（可以无限复读）把账留在原地续期。

因此第十二章的两个落点全部要重读：语法改造（12.2）关的是结清的门，而那扇门本来就是关的；可逆性（12.3）开的是续期的闸——它不是解药，是最优运行条件。三处订正的第二处已经撤回了12.3；这里给出撤回之后剩下的那条唯一落点的完整形式：

唯一的落点是让账可以被结清，且结清必须被收账方承认。

后半句是三处订正的第二处给的，它不能被省略：USMLE 的教训不是”名额要求某处必须有一个排序”这个抽象算术，是收账方（住院医项目）自己面对名额，因此它拒绝把”通过了”当成一个终点。一张收账方不认的结清凭证，不是结清。

14.4 第十一章在新框架下：判据换变量

本书 11.3 的判据：

效标属于外壳系统，当且仅当它本身承担稀缺分配功能。

三处订正的第三处已经注销了”稀缺是根”这一条。所以这个判据用错了变量。它必须改为：

效标属于外壳系统，当且仅当它交付不可结清的反事实自我。

新判据严格更宽，且它捕获了旧判据会漏掉的东西：

效标	旧判据（承担稀缺分配功能？）	新判据（交付不可结清反事实？）
大学成绩	是 → 外壳	是 → 外壳（“我本可以拿 3.9 的我”，门槛由分布定义）
引用计数	是 → 外壳	是 → 外壳
学者的公开影响力指标	否（无名额、无正式分配）→ 判为合格效标	是 → 外壳
飞行员的实际降落记录	否 → 合格效标	否 → 合格效标（“会降落的我”可兑现）

第三行是旧判据的漏点。一个无名额、无正式分配、却由持续更新的他人分布定义的指标，旧判据放它过关，新判据抓住它。而这类指标恰恰是当代学术评价里增长最快的那一类。

这使第十一章的强制变强了。 教育测量学要给出一个非外壳的效标，此前它需要给出一个“不承担稀缺分配功能”的效标；现在它必须给出一个**交付可结清反事实**的效标——一个人可以抵达它、兑现它、然后离开。

而本书 11.3.2 提出的三个候选效标——无分数激励条件下的自愿学习行为、毕业若干年后不被评分的知识保持率、不影响任何分配的专家盲评——**恰好全部满足这一条**。

这不是巧合，但本书当时不知道为什么。它给出的理由是“它们的结果不进入任何人的档案”。在新框架下可以看清：“**不进入档案**”正是“**可结清**”的一个操作化——一个不进入档案的结果，不会被下一轮的他人分布重新定义，因此它可以被抵达、被兑现、然后被放下。

本书在写 11.3.2 时不知道自己在用第三条规格。 律篇告诉了它它在用什么。

这是本章要记的第二笔，也是“证伪”这个词在这里的完整含义：它拆掉了本书的一层，同时告诉本书，本书那些说不清为什么却做对了的地方，为什么做对了。

14.5 一个未处理的问题

本章不假装把账算平了。有一个问题它没有解决，且它可能是下一层。

律篇的推论一：**新的排序型装置在结构上不再能被建立**——安装窗口需要一段耦合期来生产第一批“靠可认领属性成功”的先例，而人工智能把优化知识的生成与传播成本压到近乎零，于是新装置在诞生的同一刻就脱钩，从来没有过窗口。

如果这一条成立，本书的 P4（再分泌律）就有一个时间边界。P4 说：任何承接稀缺分配功能的排序型替代装置，将在一到两代人内分泌出同构的外壳——替代不是替代，是宿主更换。

但如果安装窗口已经关闭，就没有新宿主可换了。

用 14.3 的三个去向说：结清的门在排序型装置上本来就是关的；**迁出的门刚刚关上了**（没有新装置能接住那个账）；那么账只剩一个去向——**续期**。

三个可能，本书无法在它们之间判决：

1. **账全部回到现存装置续期**，使它们不可替代（律篇的推论一）。可观察后果：高考、评级、同行评审这类已完成安装的装置，其地位应当在人工智能之后**上升**而非下降；而复读、二战、三战这类续期行为的强度应当同步上升。这两条曲线可以立刻开始测。
2. **账迁到人身上**——如同落第者迁到儿子身上。而在一个生育率坍缩的社会里，可用的宿主正在减少。这两条曲线的交叉点是一个可以计算的东西，本书没有计算它。
3. **账迁不出去，也续不了期**。一个迁不出去、又不能结清、又不能丢弃的账，其持有者会怎样——本书没有答案，律篇也没有。

这是两篇共同的边界，而它可能是最要紧的那个问题。 本书把它留在这里，不装作它已经被处理。

14.6 这一章自己的位置

本章的位置需要被诚实地标出来，因为它有一个明显的嫌疑：**一本书在被证伪之后，加了一章来说明这次证伪让它变强了。**

这个动作本身就是本书用十三章描述的那个动作的候选样本。一个被拒者接到拒绝之后做的第一件事，正是把它接回自己的叙述——而”这次失败让我更强”是那个叙述最常见的形态。本章甚至满足了那个规格的每一条：它不可争辩（它给的是可查的材料）、可归因（错是本书自己的）、而且——这一条最难看——它让那个账**续期**了：本书没有结清这次失败，它把失败重新计息成了两笔资产。

本章无法自证它不是。它能做的只有三件事，且都已经做了：

其一，它不碰错的那部分。 7.3、12.3、13.4 局限之六的原文一字未改，三处订正就地贴在它们后面，位置在本章之前。读者读到错误时先看到订正，再读到这一章。顺序不能反过来——反过来就是先讲功劳再认错。

其二，它交出的两笔是可结算的，不是修辞。 P7 的新表述产出 1903 断点这一条（登记在附录 C 表外）；第十一章的新判据产出一张四行的表，其中第三行是一个立刻可以去查的漏点。如果这两笔查出来是错的，本章的”变强”就是空的。

其三，它把它算不平的账留在了 14.5，没有收尾。

至于本章是不是又一层壳——**这个判断不由本章做。**它由那个去查 1903 年地方学堂入学数据的人做，或者由那个去查学者影响力指标是不是外壳的人做。

这正是本书的全部主张所要求的位置：**它不能自己判自己没在分泌。** 它只能把可以被结算的东西放出去，然后闭嘴。

参考文献

英文文献

Arthur, W. B. (1994). *Increasing returns and path dependence in the economy*. University of Michigan Press.

Au, W. (2009). *Unequal by design: High-stakes testing and the standardization of inequality*. Routledge.

Beetham, D. (1991). *The legitimation of power*. Macmillan.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.

Bourdieu, P., & Passeron, J.-C. (1977). *Reproduction in education, society and culture* (R. Nice, Trans.). Sage.

Bowles, S., & Gintis, H. (1976). *Schooling in capitalist America: Educational reform and the contradictions of economic life*. Basic Books.

Bromley, P., & Powell, W. W. (2012). From smoke and mirrors to walking the talk: Decoupling in the contemporary world. *Academy of Management Annals*, 6(1), 483–530.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1), 67–90.

Cascio, E., Sacerdote, B., Staiger, D., & Tine, M. (2024). Report from working group on the role of standardized test scores in undergraduate admissions. Communication to S. L. Beilock and L. Coffin, Dartmouth College, January 30.

Chang, C. (1955). *The Chinese gentry: Studies on their role in nineteenth-century Chinese society*. University of Washington Press.

Cholongitas, E., Marelli, L., Kerry, A., Goodier, D. W., Nair, D., Thomas, M., Patch, D., & Burroughs, A. K. (2007). Female liver transplant recipients with the same GFR as male recipients have lower MELD scores—A systematic bias. *American Journal of Transplantation*, 7(3), 685–692.

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences*, 14(1), 119–135.

Collins, R. (1979). *The credential society: An historical sociology of education and stratification*. Academic Press.

- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.
- David, P. A. (1985). Clio and the economics of QWERTY. *American Economic Review*, 75(2), 332–337.
- Diamond, L. (2002). Elections without democracy: Thinking about hybrid regimes. *Journal of Democracy*, 13(2), 21–35.
- Elman, B. A. (2000). *A cultural history of civil examinations in late imperial China*. University of California Press.
- Elman, B. A. (2013). *Civil examinations and meritocracy in late imperial China*. Harvard University Press.
- Elster, J. (1982). Marxism, functionalism, and game theory: The case for methodological individualism. *Theory and Society*, 11(4), 453–482.
- Elster, J. (1983). *Explaining technical change: A case study in the philosophy of science*. Cambridge University Press.
- Elster, J. (1989). *Solomonic judgements: Studies in the limitations of rationality*. Cambridge University Press.
- Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology*, 113(1), 1–40.
- Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a social process. *Annual Review of Sociology*, 24, 313–343.
- Friedman, J. N., Sacerdote, B., Staiger, D. O., & Tine, M. (2025). Standardized test scores and academic performance at Ivy-Plus colleges. *AEA Papers and Proceedings*, 115, 676–681.
- Habermas, J. (1975). *Legitimation crisis* (T. McCarthy, Trans.). Beacon Press. (Original work published 1973)
- Hirschman, A. O. (1970). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*. Harvard University Press.
- Kamath, P. S., Wiesner, R. H., Malinchoc, M., Kremers, W., Therneau, T. M., Kosberg, C. L., D' Amico, G., Dickson, E. R., & Kim, W. R. (2001). A model to predict survival in patients with end-stage liver disease. *Hepatology*, 33(2), 464–470.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- Karabel, J. (2005). *The chosen: The hidden history of admission and exclusion at Harvard, Yale, and Princeton*. Houghton Mifflin.

- Kim, W. R., Mannalithara, A., Heimbach, J. K., Kamath, P. S., Asrani, S. K., Biggins, S. W., Wood, N. L., Gentry, S. E., & Kwong, A. J. (2021). MELD 3.0: The Model for End-Stage Liver Disease updated for the modern era. *Gastroenterology*, 161(6), 1887–1895.e4.
- Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. Farrar, Straus and Giroux.
- Levitsky, S., & Way, L. A. (2010). *Competitive authoritarianism: Hybrid regimes after the Cold War*. Cambridge University Press.
- Lipset, S. M. (1959). Some social requisites of democracy: Economic development and political legitimacy. *American Political Science Review*, 53(1), 69–105.
- Mahoney, J., & Thelen, K. (Eds.). (2010). *Explaining institutional change: Ambiguity, agency, and power*. Cambridge University Press.
- Massie, A. B., Caffo, B., Gentry, S. E., Hall, E. C., Axelrod, D. A., Lentine, K. L., Schnitzler, M. A., Gheorghian, A., Salvalaggio, P. R., & Segev, D. L. (2011). MELD exceptions and rates of waiting list outcomes. *American Journal of Transplantation*, 11(11), 2362–2371.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.
- Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, 83(2), 340–363.
- Miyazaki, I. (1976). *China's examination hell: The civil service examinations of imperial China* (C. Schirokauer, Trans.). Yale University Press.
- Muller, J. Z. (2018). *The tyranny of metrics*. Princeton University Press.
- Myers, R. P., Shaheen, A. A. M., Aspinall, A. I., Quinn, R. R., & Burak, K. W. (2011). Gender, renal function, and outcomes on the liver transplant waiting list: Assessment of revised MELD including estimated glomerular filtration rate. *Journal of Hepatology*, 54(3), 462–470.
- Nozick, R. (1974). *Anarchy, state, and utopia*. Basic Books.
- Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5(2), 187–195.
- Pierson, P. (2004). *Politics in time: History, institutions, and social analysis*. Princeton University Press.

- Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice*, 16(2), 9–13.
- Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.
- Rawls, J. (1971). *A theory of justice*. Harvard University Press.
- Sacerdote, B., Staiger, D. O., & Tine, M. (2025). How test optional policies in college admissions disproportionately harm high achieving applicants from disadvantaged backgrounds (NBER Working Paper No. 33570). National Bureau of Economic Research.
- Sandel, M. J. (2020). *The tyranny of merit: What's become of the common good?* Farrar, Straus and Giroux.
- Sauder, M., & Espeland, W. N. (2009). The discipline of rankings: Tight coupling and organizational change. *American Sociological Review*, 74(1), 63–82.
- Schedler, A. (2002). Elections without democracy: The menu of manipulation. *Journal of Democracy*, 13(2), 36–50.
- Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press.
- Sen, A. (1992). *Inequality reexamined*. Harvard University Press.
- Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3), 355–374.
- Stiggins, R. J. (2002). Assessment crisis: The absence of assessment FOR learning. *Phi Delta Kappan*, 83(10), 758–765.
- Walzer, M. (1983). *Spheres of justice: A defense of pluralism and equality*. Basic Books.
- Weber, M. (1978). *Economy and society: An outline of interpretive sociology* (G. Roth & C. Wittich, Eds.). University of California Press.
- Wiesner, R., Edwards, E., Freeman, R., Harper, A., Kim, R., Kamath, P., Kremers, W., Lake, J., Howard, T., Merion, R. M., Wolfe, R. A., & Krom, R. (2003). Model for end-stage liver disease (MELD) and allocation of donor livers. *Gastroenterology*, 124(1), 91–96.
- Young, M. (1958). *The rise of the meritocracy*. Thames & Hudson.

中文文献

- 顾炎武. 日知录（卷十六「拟题」）.
- 黄宗羲. 明夷待访录·取士.
- 刘坤一、张之洞. 江楚会奏变法三折. 1901.

《遵旨核议学堂选举鼓励章程折》。1903 年 12 月。

袁世凯、张之洞等。奏请立停科举推广学校折。1905 年 9 月 2 日。

附录A 外壳厚度的操作化编码手册

A.1 用途

本手册把 7.1 的三重结构与 7.3 的第四条规格操作化为一组可编码的维度，用于 10.2 的判决格测量与跨案例比较。

编码对象是**装置**，不是**结果**。同一个装置在不同时期可以有不同的编码，且这正是纵向研究所需要的。

A.2 四个维度

每一维 0—3 分。外壳厚度 = 四维之和（0—12）。

维度一：非人格化（Impersonality）

被拒者能否指认一个做出这个决定的人？

分	特征	例
0	决定由一个可指名的人做出，被拒者知道是谁	师徒制的淘汰、九品中正
1	决定由一个小群体做出，被拒者知道群体构成	系里的招生小组
2	决定由一个程序做出，但程序中有可指认的判断环节	有面试环节的录取
3	决定由一个公式做出，无人参与判断	高考总分录取、MELD

维度二：程序化（Proceduralisation）

换一个人执行，结果会不会变？

分	特征	例
0	完全依赖执行者的判断	导师的整体评价
1	有评分细则但留有大量裁量	结构化面试
2	细则严密，裁量有限，有复核机制	双盲阅卷的主观题
3	机器执行，零裁量	机读答题卡

维度三：通约化（Commensuration）

被拒者能不能说”我和他不一样，不能这么比”？

分	特征	例
0	无共同标量，个案对个案	推荐信
1	有标量但跨单位不可比	各校标准不一的高中 GPA
2	有标量且跨单位可比，但需要折算	各省高考分数、不同版本的语言考试
3	全域单一标量，直接可比	全国统一试卷的原始分、concours 名次

维度四：可归因性（Attributability）

被拒者能不能把这个结果认领为”我做的（或没做的）事的后果”？

分	特征	例
0	结果与被拒者的任何行为无关	抽签、按身份证尾号
1	结果主要由被拒者不能改变的属性决定	智力测验、家庭背景加权
2	结果由被拒者可改变但代价高昂的属性决定	需要长期资源投入的活动履历
3	结果由被拒者可直接控制的行为决定	刷题量、备考时长

A.3 编码示例

装置	非人格	程序化	通约化	可归因	厚度
中国高考	3	3	2	3	11
法国 concours	3	2	3	3	11
MELD 肝移植分配	3	3	3	1	10
明清科举（成熟期）	3	2	3	3	11
美国大学录取（2021, test-optional）	1	1	0	2	4
美国大学录取（2019）	2	2	2	3	9
高中 GPA	1	1	1	3	6
导师整体判断	0	0	0	2	2
抽签	3	3	3	0	9
九品中正	0	0	1	1	2
同行评审（双盲）	2	1	1	2	6

A.4 两个必须注意的地方

其一，**抽签的厚度是 9，不是 0。**这不是编码错误。抽签在前三维上是满分——它是最非人格、最程序化、最通约的装置。它的问题只在第四维。这个编码结构准确地反映了 7.3 的论证：抽签不是外壳的反面，它是一个缺了第四条规格的外壳，而正是那一条使它不可用。

其二，**维度四与前三维的关系是非对称的。**前三维生产不可争辩性，可以相互替代（一个通约化残缺的装置可以靠极强的程序化补偿）。维度四不能被补偿：它是一个门槛，不是一个加项。一个维度四为 0 的装置，无论前三维多高，都不能承担教育分配（7.3）。

因此，严格地说，外壳厚度应当写成：

$$\text{厚度} = (\text{非人格} + \text{程序化} + \text{通约化}) \times 1[\text{可归因} \geq 1]$$

即：可归因性为零时，厚度为零。本表用加和形式是为了便于比较，使用时须留意这一处非线性。

订正五 · 这张表给一个不分泌外壳的装置打了满分

把飞行执照的实操检定放进上面这张表：非人格 3（拒绝你的是标准）、程序化 3（同一套动作，任何考官执行结果相同）、通约化 3（有一条线）、可归因 3（你会不会降落是你练出来的）——**厚度 12，全表最高，高于中国高考的 11。**而它不分泌外壳（见订

正一)。

这不是一个边界情形，是这个度量的失效。**它度量的不是外壳厚度，是外壳的必要条件——三重结构加可归因，而那是阈值型装置全都有。**律篇：三重结构不是外壳的定义，它是外壳的必要条件；同一套结构在两类装置上生产两样东西，区分它们的是**被拒者能不能自己重放这个程序**。飞行考试的被拒者可以：他自己上去，把那个动作再做一遍，自己看着自己没能把飞机放平，他不需要相信任何人。高考的被拒者不能：重放"那 N 个人比我强"要求把那 N 份答卷都重做一遍，那不是一个人能做的动作。**可验证性与不可争辩性都终结争辩，方式相反——一个让他自己看见，一个让他无处着手。**本书 9.1 写的"他不需要相信任何人，他自己知道他不会开飞机"说中了这个现象，但把它当成了语法的推论，因此没有走到这一步。

补第五维：可结清性（反向计分）。

分	特征	例
0	反事实由固定标准定义，可抵达、可兑现、兑现后终止	飞行执照、焊工工艺资格
1	标准固定，但兑现须第三方承认，且第三方可以不认	改制后的 USMLE Step 1（收账方另找标量）
2	线每期重划，但重划由分布之外的因素决定	部分行业准入线
3	反事实由他人的位置定义，位置在动，不可抵达	高考、concours、排名、影响力指标

于是 A.4 的式子要改，而它改出来的形状恰好是对称的：

分泌压力 = （非人格 + 程序化 + 通约化） × 1[可归因 ≥ 1] × 1[不可结清 ≥ 1]

两个门槛，对应规格的第二条与第三条。抽签的可归因为 0，压力为 0；飞行执照的不可结清为 0，压力为 0；两者在这张表上是同一类——**三重结构满分而缺一个门槛**。高考两个门槛都过，压力为 8。

而这一处的**诊断比它的修正更值得记下来**。本书为抽签写了整整一段辩护（A.4 其一：“抽签的厚度是 9，不是 0。这不是编码错误”），说明它为什么高分却不是外壳；

本书没有为飞行执照写这一段，**因为飞行执照不在这张表里**。表里有一个缺第二门槛的装置，于是本书发现了第二个门槛，还专门为它写了一处非线性；表里没有一个缺第三门槛的装置，于是本书从未发现第三个门槛存在。**样本的选择掩盖了度量的缺陷，而这张表的样本是本书自己选的**。这一条不必等任何人来查——它今天就可以被任何一个读者用一行编码验证，而本书自己有十一个装置的机会去做这一行，一次都没做。见《**反事实自我的社会制造**》第四节第一小节。

A.5 局限

本编码手册的信度未经检验。它需要至少两名独立编码者对同一批装置编码，并报告一致性系数。本文没有做这件事，这是 13.4 局限之五的具体形态。

附录B 三项挪用判据的测量指引

B.1 用途

本指引把 9.2 的三项判据操作化，使 P5 可以被前瞻性地检验。

使用纪律：三项判据必须在观察脱钩之前测量。任何在观察到脱钩之后回头宣称“判据其实早就转阳了”的做法，都使 P5 退化为不可证伪的标签——这正是 9.2 所要防止的。

B.2 判据一：下游使用

定义：是否存在一个有名额约束的分配主体，公开或事实上使用该装置的输出作为筛选依据？

数据源（按证据强度排序）：

1. 下游主体的公开政策文件（招录简章、评审细则、内部指引）；
2. 对下游决策者的系统调查（如美国住院医项目主任调查）；
3. 从业者的报告与行业媒体；
4. 该装置的输出被第三方汇编成榜单或折算表这一事实本身。

转阳条件：第 1 或第 2 项证据出现，即判定为阳。第 3、4 项单独出现只判定为“疑似”，需要第 1 或第 2 项确认。

边界情形：下游主体否认使用，但从业者普遍报告使用。此时判定为阳，并标注“事实使用”。理由是律的机制不依赖公开性——它依赖被拒者是否需要回答“凭什么是他不是我”，而这个问题在事实使用的情况下同样成立。这是本判据测量上的最大困难（13.4 局限之五）。

B.3 判据二：标量输出

定义：该装置是否向下游输出连续标量（而非二值的通过/不通过）？

转阳条件：装置向被评者或下游主体提供任何比二值更细的输出。包括：

- 数值分数（三位数、百分制、绩点）；
- 等级（A/B/C/D——注意：四级等级已经是标量，它可以排序）；
- 百分位、名次；
- 任何可以被下游排序的输出。

不算标量的：纯二值（通过/不通过、合格/不合格）、无序类别（专业方向、语种）。

注意一个常见误判：一个装置对被评者只公布二值结果，但向下游主体提供分数——此时判据二为阳。测量的对象是“下游能不能拿到标量”，不是“考生能不能看到分数”。

B.4 判据三：名额紧度

定义：下游分配场景的名额紧度 = 申请者数 / 名额数。

计算：使用最近一个完整周期的公开数据。若无公开数据，使用行业估计并标注。

分级：

紧度	判定
= 1（合格者全部录取，无名额上限）	阴
1 < 紧度 < 2	弱阳
≥ 2	阳

注意：紧度为 1 是一个严格条件。它要求的不是”录取率高”，是”不存在名额上限”——即，如果所有人都合格，所有人都通过。飞行执照实操检定满足它；一个录取率 90% 的大学不满足它，因为它的名额是固定的。

B.5 判定规则

三项判据	判定	P5 的预测
全阴	纯阈值型	保持保真
判据二阳，一、三阴	阈值型但有标量（易感）	保持保真，但处于风险中
判据一或三阳，二阴	阈值型但下游想排序（临界）	压力存在，将迫使标量出现
全阳	已被挪用为排序型	分泌外壳并脱钩

第三行值得注意：它预测的是一个具体的动作——当下游有排序需求而装置只输出二值时，压力会迫使标量出现（下游会自己造一个标量，或要求装置提供）。这是影子排序（10.5）在阈值型装置上的形态，且它是本判据体系产出的一个额外的可检验预测。

B.6 本指引产出的押注对象

按 B.5 的规则，本文点名三个当前三项判据全阴的装置（附录 C 第 6—8 号）：

装置	判据一	判据二	判据三	判定
商业飞行员 ATP 实操检定	阴	阴	阴（=1）	纯阈值型
结构焊工工艺资格认证（AWS 等）	阴	阴	阴（=1）	纯阈值型
麻醉设备操作资质认定	阴	阴	阴（=1）	纯阈值型

三者的共同特征：输出二值、无有名额约束的下游主体用其排序、合格者全部通过。

预测：三者保持测度保真。若任一脱钩，且脱钩时三项判据仍为阴，P5 被推翻。

关于本附录·律篇对押注表的影响，登记于表外

本表一字未改，且不会被改。C.3 其五：本表不得被后续修改以适应观察结果。律篇的出现是一个观察结果，因此它不得进入本表。以下两点登记在表外，日期 2026 年 7 月 18 日，与本表同日。

关于第 1 号押注。它的结算条件写的是"外壳在两代人后仍与所声称测量的对象保持稳定耦合"。这个条件可以被本文的辩护者用等待消解——"再等等，两代人还没到"。一条律的证伪条件如果能被它的辩护者用等待消解，那条律就不是律。不可拖延的版本是：找到一个排序型装置，其优化产业已经成熟（存在以优化该装置输出为唯一目的的商业服务，且市场规模稳定或增长），而其可认领面与被测面仍保持稳定重合。"优化产业成熟"是一个有日期的可观察事件，不是一个可以再等等的期限。这个更强的版本登记在律篇附录·公开押注登记表第 6 号，不登记在这里——本表保留它原来的、较弱的样子，因为那正是本文当时押下的那一注。

关于第 12 号押注。它与 12.3 之间有一条被本文忽略的缝：第 12 号量的是一个相关，12.3 荐的是一根杠杆。不可逆性与考试烈度可以正相关（第 12 号可以赢），而降低不可逆性仍然不是解药（12.3 仍然错）——因为两者可以同为不可结清的下游。这一注赢了，那一节仍然是错的。分开两者需要一条本表没有的押注：控制职业资格的独立定价之后，可逆性对考试烈度的独立效应应当消失或反转。这一条登记在律篇附录·公开押注登记表第 10 号，不补进本表。

关于 1903 断点（第十四章 14.2）。P7 的新表述——装置死于它交付的反事实自我全部离开它——产出一条本表没有的预测：1900—1908 年间士绅家庭的实际行动（送子入学堂 vs 继续应试），其转折点跟随 1903 年 12 月的学堂科名对应法则，而不是跟随 1905 年 9 月的停科举诏令。结算条件：一项对该时段地方学堂入学与应试人数的史料计量，看断点落在哪一年。它同样登记在表外，理由与上面两点相同：律篇的出现是一个观察结果，而一张可以被后来的发现追加的押注表，不是押注表。

它的结算日期一栏写"开放"，而这一次"开放"不是一个可以被拖延的期限——这一条不需要等待任何事情发生。那些史料在 1908 年就已经全部到位了。上面第 1 号押注的毛病是它的辩护者可以说"再等等，两代人还没到"；本条没有这个出口，因为没有什么可等

的。一条今天就能被结算而一百二十年没有人去结算的预测，其未结算状态不说明它的辩护者在拖延，只说明**没有人被指派去拉那个闸**——而这正是律篇附录 12.5 对它自己的判断。

附录 C 公开押注登记表

C.1 这份表是什么

由 13.5.8，本文唯一的非同壳效标是那些在学术场域之外结算的预测。本表是它们的登记。

这不是一个附录。它是本文唯一的经验地基。

每一条押注包含：编号、对应命题、预测内容、结算条件、结算日期、当前状态。结算条件的写法遵循一条纪律：**它必须能被一个不同意本文的人执行，且执行结果不依赖他是否同意本文。**

登记日期：2026 年 7 月 18 日。

C.2 登记表

#	命题	预测	结算条件	结算日期	状态
1	P4（律的核心）	不存在一个承担稀缺分配功能的排序型装置，其外壳在两代人后仍与所声称测量的对象保持稳定耦合	任何人给出一个反例，附耦合系数的纵向证据	开放	未结算
2	P7（消亡条件）	不存在一个装置，其稀缺分配功能完好而因效度证伪消亡	任何人给出一个反例，附两个可考日期（分配功能转移日 / 死亡日）	开放	未结算
3	P4 / P5（USMLE）	Step 2 CK 在 Step 1 改制五年后成为住院医师申请的核心排序指标，且其商业备考产业成熟	项目主任调查显示 Step 2 CK 分数为前三位筛选依据；且存在专门的 Step 2 CK 备考产业	2027-01	未结算
4	P4（test-optional）	更多高选择性院校撤回 test-optional，方向是恢复标准化考试而非发展替代评价	统计 2024—2029 年间撤回与新增 test-optional 的高选择性院校数之比	2029-12	部分兑现（达特茅斯、耶鲁、布朗已撤回）
5	P3'（AI 对策的收敛方向）	AI 时代被稳定采纳的评价对策朝可归因性收敛（现场手写、口试、留痕、检测），而非朝准确性收敛（长期过程性评价、多证据三角、AI 辅助深度评估）	抽样统计主要高校 2026—2030 年新出台的评价政策，按两类分类计数	2030-12	未结算
6	P5（阈值保真）	商业飞行员 ATP 实操检定保持测度保真	该检定的通过与实际飞行安全表现的关联未系统性衰减；且若衰减，检查三项判据是否已转阳	2036-07	未结算

#	命题	预测	结算条件	结算日期	状态
7	P5 (阈值保真)	结构焊工工艺资格认证保持测度保真	同上, 效标为焊缝的实际无损检测合格率	2036-07	未结算
8	P5 (阈值保真)	麻醉设备操作资质认定保持测度保真	同上, 效标为设备相关不良事件率	2036-07	未结算
9	2.3 / 10.4 (分离检验)	在高烈度升学社会的落榜者中, 制度否定强度与个人结果接受度不相关或弱相关 ($ r < 0.3$)	一项针对落榜者的调查, 两个量表按 10.4.1 操作化	开放	未结算
10	P8 (跨壳效标)	在不承担分配功能的效标上, 标准化考试相对高中 GPA 的优势显著缩小或消失	一项使用 11.3.2 三类候选效标之一的研究	开放	未结算
11	4.5 (本文署名)	本文若投往采用现行作者规范的期刊, 被拒或被要求改署名的概率接近于一; 且理由以责任/质量的语法表述, 而非以”我们需要知道这是谁做的”表述	投稿并公开评审信原文	投稿后即时	未结算
12	P6 / 12.3 (可逆性)	首次教育分流决定与四十五岁职业地位的相关系数, 与该社会的考试烈度正相关, 且在控制人均收入、基尼系数、教育扩张程度后仍显著	一项跨国研究, 样本 ≥ 15 国	开放	未结算

C.3 结算纪律

其一, 本文不结算自己的押注。 由 13.5, 本文的作者在这个场域内, 因此他们的结算是同壳的。第 3、4、5 号的结算应由与本文无利害关系的第三方执行, 或至少由公开数据支持以便任何人复核。

其二，“部分兑现”不算兑现。第 4 号目前标为”部分兑现”，但它的结算日期是 2029 年 12 月。在那之前，三所院校的撤回不构成兑现——它构成一个方向的初步信号，而信号不是结算。

其三，反例只需要一个。第 1、2 号是全称命题的证伪邀请。它们不需要统计，只需要一个案例。这是它们比其余十条更强、也更脆弱的地方。

其四，输了就是输了。若第 6—8 号中任何一个装置在三项判据保持阴性的情况下脱钩，P5 被推翻，而 P5 是本文全部可证伪性的来源。届时本文的处理方式是承认，不是引入辅助假设。这一条写在这里，是为了让将来的引入辅助假设变得难看。

其五，本表不得被后续修改以适应观察结果。本表的登记日期是 2026 年 7 月 18 日，随本文一同公开发表于 sdeuniverses.com 学术创造专栏学位论文频道，该页面的版本历史可查。任何对本表的修改都将留下记录。

C.4 一个说明

十二条押注里，有十条本文赢了对自己没有好处，输了则理论受损。这是不对称的，且这个不对称是故意的。

一个理论如果只押那些赢了对自己有好处的注，它押的就不是注，是广告。

第 11 号是唯一一条本文赢了会很难看的——它预测的是本文自己被拒。而它恰好是十二条里最快结算的一条。