成功性脆弱：当成功本身成为失能的发生学引擎

控制变量不是对标准的偏离，而是标准本身的位移——事后调查将找不到那位吹哨人

作者 鲍锦朝 · SDE 学员 · 约 2.6 万字 · 发表于2026年8月2日

摘 要

本文论证了安全科学长期忽视的一种独立的组织脆弱性形态——成功性脆弱。它并非由故障、违规或设计缺陷所驱动，而恰恰是由持续攀升的安全绩效与制度化代偿层的全面覆盖所共同生成的一个正反馈过程。论文的核心判断是：当一个专业系统越是成功地用标准化的操作界面替代从业者对真实情境的直接裁断，且这一“成功”在可审计的安全指标上反复得到确认时，该系统应对真正无先例极端事件的深层能力，便在一种“反馈稀缺”的静默引擎驱动下发生不可逆的衰退。本文分三步揭示这一引擎的运作机制：首先，长期零事故环境通过预测误差压制，系统性地抬升了组织对弱信号的感知阈值；其次，由制度背书的代偿层界面在全面覆盖操作生态位后趋于闭合，截断了从业者在不依赖外部认证的情况下建立独立判断通路的身体性条件；最后，由完美指标沉淀出的组织成功叙事，通过对残余异常的重新驯化，构成一股反向屏蔽力量，锁定了解释权边界。但本文的核心理论贡献在于论证：这三步并非三个独立的根源，而是同一个“成功–反馈稀缺”引擎在认知、操作与叙事三个界面上按明确的时间–序列结构逐层展开的递归显现。本文将这一新命名置于与正常事故理论、高可靠性组织理论、韧性工程、去技能化理论、认知卸载理论、偏差正常化以及“漂向失败”理论这七个最强近邻的严肃对质之中，逐次析出其不可还原的控制变量：标准本身的位移，而非对标准的偏离。论文以民航、核工业、航天与金融交易等领域的经验材料为机制提供了跨域痕迹证据，并通过一个纵深的行业案例展示了引擎在灾难发生前数年的渐进运作。文章自陈了当前实证链条的阴性案例盲区，并给出了一个可操作的前瞻性检验研究设计。最后，本文正面回应了最根本的一项反驳——即“这不过是偏差正常化的别称”——并给出了区分二者的判决性对比预测，并设立了严格的证伪条款以保持其可检验性。

关键词：成功性脆弱；发生学机制；正反馈引擎；代偿层；感知阈值；成功叙事；组织退化

引言

本文所追踪的问题，源自一个更原初的追问：工具延长了人的手，为什么同时也可能截断手的神经？

这是一个个体–工具–身体层面的发生学问题。它指向一种同源性——同一种延伸何以同时包含一种截除。然而，当我们试图将这一问题置入经验世界进行追踪时，立刻撞上了一个必须在分析层面予以回应的困境：在单纯的“人–工具”互动层级，我们固然能精妙地描述某一项具体技能在长期依赖外部介体后发生的废用性衰退——比如依赖导航软件导致空间记忆的神经基础萎缩——却无法切出本文将要追索的那种更具系统性、也更隐蔽的退化机制。因为后者并不发生在单一的工具通道内，而发生在整整一套由组织成功、制度激励与认知生态共同构成的正反馈回路之中。截断神经的，不是一柄更锋利的工具，而是一个由持续成功驱动的、将“不犯错”本身变成一种自毁加速器的组织演化过程。这一机制唯有在“专业系统”这一层级上，才能被发生学地完整追到。

基于这一裁定，本文将原初问题从“工具何以截断神经”改切为：在以制度化的代偿层系统性地替代从业者的身体性裁断与切身担责为特征的专业系统中，安全绩效的持续成功，如何同步注销了该系统处理真正无先例极端事件的发生基础？这一改切的理由在于：分析单位必须从“个体–工具”的单一交互，上移至“组织–制度–绩效”这一能够捕捉反馈稀缺、认知生态变迁与自我论证叙事之综合效应的层级。由此，本文追问的便不再是单一技能的废用性衰退，而是整个专业系统在通往完美安全指标的道路上，如何从内部生成出一种与故障无关、完全由成功滋养的脆弱性。我们将这一机制命名为“成功性脆弱”。

必须提前厘清的是，这一命名并非指一种更高级的安全状态，而专指一种特定的组织退化形态：它不是设计缺陷的暴露，而是成功自身所沉淀下来的一条演化路径。沿此路径，系统的可测量风险指标越是攀升，其应对完全落在既有标准之外、不具备被识别为“错误”之资格的根本性异常的能力，便越是无可挽回地走向枯竭。换言之，当每一个可被管理、可被审计、可被定价的风险都已被精确地降至极低时，剩下那些无法被纳入既有风险框架、因而无法触发警报的“未知之未知”，恰恰是未来真正的肇事者；而恰在此时，系统内部处理这类事件所必需的发生土壤——那些必须在真实的、不可撤回的担责情境中被一代代从业者身体性地生成为即时判断基底的东西——已然在“成功”的静默浇灌下，失去了其可被重新激活的物质基础。

这不是危言耸听。下文各节将依次做四件事：解构这一机制的三条运作原理并论证其同源性时间–序列结构；将这一新命名置于与七个最强近邻的严肃对质中，析出其不可还原的控制变量；通过一个纵深的行业案例与跨领域经验痕迹展示机制的实际运作；正面回应最有力的反驳；最后，设立严格的证伪条款、边界条件与一个可操作的前瞻性检验方案。

一、成功性脆弱的三条运作原理及其同源性引擎

成功性脆弱并非由某一单独原因触发，亦非三个独立根源的偶然叠加。它是同一个正反馈引擎——我们称之为“成功–反馈稀缺”——在专业系统的三个不同界面上按特定时间次序逐层展开、并最终形成相互咬合的循环的递归显现。这三个界面分别是：从业者个体与组织整体的感知层、由制度化工具构成的操作层，以及组织自我理解的叙事层。引擎的核心燃料只有一种：系统通过制度化防护持续地、成功地排除了它所定义的那类故障，从而使得一切本可以及早预警的真实危险信号，或者被彻底屏蔽在感知阈限之下，或者虽被接收却被代偿层界面所吸收，或者最终在组织叙事中被重新归类为“已受控的波动”。下文分述三者在各自界面上的运作机制，并在此基础上论证三者之间的时间–序列结构与同源性。

（一）感知阈值的系统抬升

任何安全系统在日常运行中都会生成海量信息：振动读数、维护日志、操作偏差、拦截记录。系统识别危险的能力，并不取决于这些信号是否客观存在，而取决于操作者与管理层对它们的感知基准线位于何处。这一基准线并非固定不变，而是在持久“零严重事故”的组织反馈环境中，经历着合乎神经加工规律的持续漂移。

预测编码理论为这一漂移提供了神经计算层面的精确描述：大脑并非被动接收信号，而是持续生成对下一时刻感官输入的预测，并将实际信号与预测之间的误差作为“需要处理的意外”向上传递。

当系统长期稳定运行时，高阶预测模型——“一切正常”——被反复验证，其先验精度不断提高，从而使得同等大小的偏差信号在误差层被压制，不再进入意识加工区间。这不是人的懈怠，而是任何具备层级预测加工机制的信息处理系统——无论生物神经网络还是组织化的注意力分配——在持续稳定的环境中必然发生的适应。脑成像研究为此提供了独立的证据：例如，伦敦出租车司机在长年累月的空间导航中，其海马体后部灰质体积显著大于控制组，且增厚幅度与从业年限正相关；反过来，长期依赖GPS导航的被试在空间记忆任务中不仅表现更差，而且其海马体与尾状核的激活模式已呈现与从未使用导航者不同的通路招募特征。这些发现表明，长期的环境反馈模式会真实地重塑执行某项功能所依赖的神经基础本身——不是暂时搁置，而是结构性改变。

在组织层面，这一机制表现为：曾经能够触发安全委员会专项会议的早期弱信号——反应堆冷却泵的一种非标准振动特征、某一航班起落架收放时超出设计时长的细微延时——逐渐被归入“可接受的正常运行波动”范畴。每一项“继续正常运行”的反馈，都在抬高感知阈值的基线，使得组织整体对于“什么事情值得警觉”的判断标准，在无痛、无过、完全合规的过程中缓慢上行。这不是管理层的无能，而恰恰是其依据过往经验做出的理性贝叶斯更新。当这一阈值最终越过某一临界点后，唯有那种已经造成可见后果的剧烈异常，方能重新激活组织的警觉机制——但成功性脆弱所面对的恰恰是另一类异常：它们在首次降临之际，便已是不可挽回的终局。

（二）代偿层界面的闭合

与感知阈值抬升同步推进的，是专业系统工作流程中一种更隐蔽的演化：由外部工具、标准化流程和制度化解释体系共同构成的“代偿层”界面，逐步覆盖了从业者原本用于直接感知材料、在第一线做出裁断、并亲身承受裁断失败所带来的全身性反馈的全部操作生态位。

“代偿层”并非某一具体工具，而是一整套提供“可被审计的正确路径”的完整界面系统。它在从业者与不确定情境之间预制了一张完备的命名网络、决策分支、操作指令与解释体系。在这套界面上，从业者不再需要启动自身复杂的经验模式去直接“读”面前那团含混的材料，而只需将材料精准归类到代偿层提供的类目格中，从格中调用预先设定的行动序列。一位外科医生面对停止出血后仍处于不可预测状态的手术创面，在过去代偿层尚未覆盖的年代，需要动员全部感官和前此所有类似病例的活体记忆去“感觉那个东西稳不稳”——这个过程同时是判断的生理基础和裁断的担责时刻。如今，代偿层为他提供了“创面稳定检查清单”与实时生命体征监测仪；他能够、也应当以“清单逐项打勾完毕”为根基来判定稳与不稳。判断被完成了——但完成判断所经由的认知通路，已经从一条依赖个体全面感觉和在承担真切后果中做出的裁断的通路，被置换为一条外接制度背书与标准分类的通路。

代偿层本身并不是问题。它当然是现代安全工程的伟大成就。问题发生在当这一层界面趋向于完全闭合之时。

所谓“闭合”，指的是：当从业者可以经由代偿层完成全部工作任务，并持续获得制度承认的正确性与组织嘉许的高绩效时，他们便不再有内驱力去启动、去维持那条与代偿层平行、但独立于制度背书之外的第二条判断通路。这条路不是被任何一条规定禁止的。它是被闲置的——被一整套将全部利害攸关的激励（绩效评估、职业晋升、法律保护）都绑缚在代偿层合规性之上的制度安排所系统性地挤出的。当从业者的时间预算被压缩至“每例X分钟，不容绕开清单”，当专业身份的建立从“我能裁断”转为“我严格依规”，当所有不可量化的身体性裁断都被归入“主观偏差”而无法进入审计档

案——闲置本身就不再只是一个选择，而成为一个由制度生态替从业者做出的安排。闲置到一定深度，那条通路所依赖的神经回路的可用性，就不再是一项可随时被个人意志唤起的备用技能，而是一个已被注销的发生事实。这正是认知卸载研究中长期观察到的从“暂时搁置”到“神经基础改变”的过渡——但在专业系统层面，这一过渡不是由个人选择驱动的，而是由组织化的成功绩效系统性地完成的。

（三）成功叙事对反馈通路的反向屏蔽

在一条稳定运行的系统中，任何好不容易穿透了感知钝化和代偿层界面的微弱信号，在抵达组织注意层之前，还必须面对最后一道防线：组织的成功叙事。

一个长期维持着优异安全绩效的组织，其内部必定已经生长出一套高度统一、逻辑自洽、并获得全体管理层情感认同的叙事结构：本组织的安全管理是卓有成效的，是经得起任何审计检验的，是值得持续优化、但绝不需要根本性重构的。这套叙事并非欺人之谈，而是从年复一年的零事故记录、不断压缩的故障率、以及不断精密的审计报告中有机地沉淀出来的。这套叙事的全部力量在于，它定义了何为“一个有意义的信号”。

当一个异常信号——例如反应堆压力容器的某类细微应力异常——穿透至此，它并不会被粗暴地否认。相反，它会被这套叙事吸收，并予之以一种无害的解释：这就是我们此前在应力测试报告中已经掌握的那种偏差，属于预置响应的管控范围，此前从未造成任何实害，因而没有理由偏离现有规程去启动针对未知事件的特别审慎。叙事的工作机制，不是封锁信息，而是消解信息中的指向性。它把所有到达此处的信号的可指涉范围，强制压缩到叙事自身所覆盖的熟悉解释区内。信号被降格为“已知波动”，组织的警觉便无须被唤醒。而当这一过程年复一年地成功运行时，每一次成功的驯服都反过来加固了叙事本身——“看，我们之前的判定是对的”——从而使得下一轮的信号更不可能逃出叙事的解释力边界。

这就是反向屏蔽。它不屏蔽信号，它屏蔽信号本该指向的未知。整个回路是自洽的。从任何一个环节单独拿出来，决策都是正确的、合规的、可辩护的。但恰恰是这些正确，将一个系统完整地封锁在它自己的成功之内。

（四）同源性论证：一个引擎的时间–序列结构

上述三条原理，不应被视作三个各自独立的根源，在此处偶然合流。它们共享同一个引擎：系统在持续成功中所经历的反馈稀缺。然而，如果仅止于宣称三者“共享同一引擎”，本文的理论贡献便仍停留在命名层——只是将三个相关现象捆绑在一起，赋予了一个共同标签。要使成功性脆弱成为一个真正的发生学概念，必须进一步刻画：这一引擎在三个界面上究竟按照何种时序展开？三者之间谁先动、谁被驱动、下一轮又由谁先动？

答案如下：引擎在三个界面上的作用，遵循一个可明确指认的因果时序，而非同步启动。

最先被引擎捕获的，是感知层。原因在于，感知阈值直接接受环境反馈的校准；它是正反馈回路中响应最快、也最不依赖于制度中介的界面。一个系统只要经历了持续的成功，其预测编码机制便会自动地、无需任何人做出决策地将同等信号的预警权重向下调整。感知阈值的抬升，是引擎启动的第

一站，也是最难被组织自我察觉的一站——因为它不表现为任何人的“错误”，而只表现为一种对正常波动的统计学习。

感知阈值的抬升，为代偿层的闭合提供了前提条件。这一因果传递的机制是：当弱信号不再能够进入组织的警觉通道时，从业者便失去了最重要的内源性驱动力去启动那条独立于代偿层之外的判断通路。因为那条通路的唯一激活条件，就是从业者“觉得不对劲”——而“觉得不对劲”恰恰依赖于感知阈值尚未被抬高到将当前异常滤除的程度。当感知层已将全部“可被归入正常”的信号滤除之后，从业者即使面对真正的异常，也不再有身体性的警觉去启动独立裁断。此时，代偿层不再是“辅助”认知的工具，而成为“全部”认知的工具——闭合由此完成。这是第一步（感知层）推动第二步（操作层）的具体因果链。

代偿层的闭合，又为成功叙事的反向屏蔽提供了它最需要的原料。成功叙事并非凭空生长；它需要源源不断的证据来维持自身。当代偿层全面覆盖操作并持续产生“完美”的可审计绩效时，这些绩效便成了叙事的铁证——任何外部的质疑都可以被一句“我们的安全记录是业界最佳的”所驳回。而一旦叙事层完成了对残余信号的全面驯化，它又会反过来加固感知阈值：因为叙事告诉组织“这些都是已知的、已受控的”，组织的预测模型便更有信心将未来类似的信号压制在意识阈之下。至此，三层之间的咬合完成了闭环：从感知到操作，从操作到叙事，再从叙事回到感知。引擎最初从感知层进入，但一旦三层形成闭环，三者便互相供养，不再有明显的“首动者”——而这恰恰意味着，试图通过“在某一层单独干预”来逆转整个过程，在理论上就是不可能的。

这就是成功性脆弱的同源性时间–序列结构：感知层先动——因为它最直接暴露于反馈稀缺；操作层继之——因为感知钝化截断了激活独立判准的内驱力；叙事层最后收紧——因为它需要前两层积累的“完美绩效”作为证据。三层形成闭环后，引擎便获得了自我维持的能力：成功喂养反馈稀缺，反馈稀缺抬高感知阈值，感知阈值促成代偿层闭合，代偿层闭合供给叙事以完美证据，叙事反向加固感知阈值。这便是本文的同源性论证。

二、近邻对质：析出“成功性脆弱”的控制变量

一个新的命名要站得住，就必须能够切出一条已有概念无法覆盖的辨别维度——不是在概念层面的精细区分，而是在经验上能产出相反预测的控制变量。本节将“成功性脆弱”分别与七个最强近邻进行对质，逐次析出其不可还原的位置，并给出一个系统的判决性对比表。

（一）与正常事故理论对质

查尔斯·佩罗的正常事故理论论证了在高复杂且紧密耦合的技术系统中，事故并非异常，而是系统结构本身的“正常”产物。本文对组织运行机制的深度追问，直接受惠于这一视角。然而，佩罗理论的基石，仍在于“事故必可追溯到某种失效”这一前设。他笔下的灾难，始终由多个微小故障的意外互动所触发——这些故障在事后均可被识别为某类“出了问题”的环节。

成功性脆弱所指向的却是另一完全不同的灾难形态：不存在可识别的故障点，所有操作界面上的决策都是依据当时的标准正确做出的，没有任何一个子程序可以被单独指认为“出错”。联合航空232号航班液压全失后，机长海恩斯与机组在没有任何规程可循的情况下，通过即兴协作实现了对飞机的

有限控制，使机上近六成人员生还。这一结果之所以可能，仅仅因为海恩斯的身体里有一套无法被任何规程描述、也无法被任何审计体系追溯的即时判断力。而这种能力的在场或缺失，在既有安全评价框架内根本就不是一个变量——因为它不属于“可被识别为故障”的那一类存在。正常事故理论需要“故障”来驱动其因果链，而成功性脆弱恰恰生存在那些已将所有可识别故障都排除干净的净空之中。

判决性对比预测：如果某次事故事后能够清晰地识别出若干个发生意外互动的失效点，则属于正常事故理论的解释范围，成功性脆弱不适用。如果事后最详尽的调查也只能将根本原因归于“长期零事故环境中身体性裁断力的代际消逝”，且找不出任何一个违反操作规程或偏离设计指标的具体环节，则成功性脆弱的解释力成立。

（二）与高可靠性组织理论对质

卡尔·维克和凯瑟琳·萨克利夫的高可靠性组织理论从实证研究中提炼出一套管理原则——专注失败、拒绝简化、对操作的敏感性、对专业知识的尊重——并论证了即使在极高风险的条件下，组织仍可依此将事故概率压缩至极低。高可靠性组织理论的真正贡献，在于揭示了组织并非只能承受复杂性，而是可以主动建构一套认知文化去管理复杂性。

成功性脆弱与高可靠性组织理论的根本分歧，不在原则层面，而在观测层面。高可靠性组织理论的研究植根于一类天然的幸存者样本：所有被识别为“高可靠性”的组织，都是那些直到被研究的时间点为止尚未发生灾难的组织。它从这些组织的成功做法中提炼出“安全”的定义，却无法观测到同一套成功绩效在持续运行时，也在从内部注销着某些它自身的观测仪器无法读取的机能。一个在全部高可靠性组织标准上都表现优异的组织，只要它尚未遭遇一次真正的无先例极端事件，它的所有正念实践在既有框架内就将继续呈现为“运行良好”。这是因为高可靠性组织的全部探测仪器，都部署在已被它自己定义为“失效”的那片区域里。它无法探测到：在同一套“专注失败”的实践中，从业者正在丧失识别那些尚未被定义为“失败”的危险的身体基础。

这里需要回应一个重要的问题：维克后期的工作——特别是关于“宇宙论插曲”和“工具失落”的论述——是否已经捕捉到了我所描述的现象？答案是：部分触及，但远未覆盖其核心。维克指出，当组织成员面临一个无法被既有认知框架收编的巨大意外时，他们会经历一种类似于“世界塌陷”的意义丧失感，并可能丢弃原本能够帮助他们应对的工具。这一观察敏锐地捕捉到了临界点上的认知崩溃，但它描述的仍是危机时刻的应激状态，而非日常成功中逐年累积的退化机制。成功性脆弱要解释的，不是为什么人在危机中会失能，而是为什么同样的组织在日复一日的成功中，正静默地丧失使他们能够在危机初始就做出正确裁断的发生条件。这不是危机中的认知塌陷，而是成功中的发生基础沙化。二者不在同一层面。

判决性对比预测：如果一个组织在所有高可靠性组织标准上均持续改善，但其从业者在首次面对真正无先例的极端异常时的初始独立裁断质量，显著低于在真实不可撤回压力中训练出同等绩效指标的对照群体，则高可靠性组织理论无法覆盖，成功性脆弱成立。若两者的初始裁断质量无显著差异，则高可靠性组织理论覆盖充分。

（三）与韧性工程对质

霍纳格尔等人所开创的韧性工程，代表了安全科学的最前沿批评性重构。其核心判断是：安全并不来自“不出错”，而来自系统在面对意料之外的震动时吸收、适应并恢复正常功能的能力。这一立场所看见的，正是成功性脆弱所关切的那个核心——在系统危险指标最为完美之时，它在面对从未被编程处置的极端危险时的恢复能力，极可能恰恰处于最低点。

然而，韧性工程将此能力表述为一种“应当被保留”的韧性，自2014年起发展的Safety-II框架则进一步将焦点转向“日常工作中真正发生的事”，强调理解系统在绝大多数时间正常运行的基础。这一进路确实比早期的韧性工程概念更为实然化，但与成功性脆弱之间仍存在一道关键的分界线。韧性工程的提问是：在正常运行中，哪些能力使系统保持了韧性？而成功性脆弱的提问是：在持续成功中，这些能力的发生条件是如何被成功本身所注销的？前者是正向追问“什么维持了适应力”，后者是反向追问“成功如何静默地取消了适应力的生成基础”。二者并不矛盾，但在因果机制上方向相反：韧性工程给出的是适应性行为的描述性清单，成功性脆弱给出的是这些行为在特定条件下无法持续的动力学原因。本文的工作不是站在韧性工程的对立面，而是站在其后方：不是“安全-I错在哪”，而是“当安全-II的理论被充分实践之后，还有什么是它看不见的”。

判决性对比预测：如果能够在不恢复“不可撤回的担责情境”这一条件的前提下，仅通过持续追加适应性训练和局部调整权限等韧性建设干预，便将感知阈值已明显抬升、代偿层已高度闭合的从业者群体的初始独立裁断能力恢复至与对照群体同等水平，则韧性工程足以覆盖成功性脆弱的核心诊断。如果必须重新引入不可撤回的担责经验这一条件本身，才能触发该能力的复苏，则成功性脆弱的不可还原性成立。

（四）与去技能化理论对质

哈里·布雷弗曼及其后继者的去技能化理论，揭示了现代劳动组织如何通过将概念权从执行者手中剥离并集中至管理层，从而系统性地贬损劳动者的专业裁断空间。这一传统对成功性脆弱所描述的一部分现象——从业者的独立裁断被制度性移除——提供了重要的思想资源。

然而，“被剥夺”与“从未长出”是两种完全不同的状态。去技能化讲述的是一个工人原本拥有完整的裁断能力，而后被工作设计分解并剥夺的故事。成功性脆弱所指涉的当代专业系统中相当大比例的新一代从业者，并不是被剥夺了某种他们此前具备的能力，而是该能力赖以生成的发生生态——必须在承担不可撤回后果的紧迫情境中，亲自做出第一次裁断——从未在他们身上被建造过。表现出的“不会”在表面上相同，但生成的谱系截然不同。对前者而言，重建是“再次调用”；对后者而言，重建是“初次生成”——且这一生成所必需的条件，在当前制度安排下已被系统性地隔绝于从业者的日常工作之外。因而，恢复的路径不可能是“教育培训”，而只能是在特定节点上重新引入真实担责这一发生条件。

判决性对比预测：如果在被去技能化的传统制造业中，被分解了技能的工人重返原始全流程工艺环境后，其独立裁断能力可通过再培训在短期内恢复，则去技能化理论仅适用于“能力被拿走”的形态。如果在高度专业化的某些领域中，即使将新一代从业者重新置于真实的、带有不可撤回后果的环境内，其初期仍表现出根本性而非熟练度层面的裁断缺失，则成功性脆弱的解释力成立。

（五）与认知卸载理论对质

认知卸载精确地描述了人将认知任务委托给外部系统以降低内部负荷的过程。导航软件卸载了空间记忆，决策支持系统卸载了临床推理，评分细则卸载了评价裁断。从卸载角度来叙述成功性脆弱，表面上不存在任何逻辑障碍。

但当代认知卸载研究的前沿结论远比“可卸载即可重建”复杂。里斯科和吉尔伯特的系列实证工作已经精细地揭示，长期卸载并非仅仅将认知任务“放在一边”，而是导致了被卸载机能的神经基础发生实质性退化——其效应不只是“暂时没用”，而是“长期未激活后，重新激活所需的基础条件已经改变”。在空间记忆的经典研究中，麦奎尔等人的伦敦出租车司机脑成像研究从正面证实了：当一项空间认知功能被长期、高强度地使用时，相关脑区的灰质体积会真实地增厚；而反向的研究则表明，长期依赖GPS导航的被试在脑成像中呈现出与从未卸载的控制组不同的神经通路招募模式。换言之，长期卸载在某种程度上改变了执行该认知功能所依赖的神经架构本身。

这正是成功性脆弱与纯卸载机制的分界线。在纯卸载框架下，被卸载的功能在需要时仍可通过复训重建——因为从理论上讲，保证该功能生长的情境基础仍在。而成功性脆弱所描述的，是卸载行为积累到某一特定阈值后，原来迫使该功能必须被调用的那个情境本身，已经被从根本上替代了。导航软件不只是让你不记路了，它确保了你不再有“不记路就会迷路”这一处境；这个处境，恰恰是空间记忆机能在代代人类身上得以生成的发生前提。处境被注销，机能随之无从生成；长期未生成，其神经基础便不再能以“功能尚未丧失、只待激活”的模式被唤醒。这便是成功性脆弱对认知卸载的补充：卸载描述了认知任务的委外，成功性脆弱描述了发生情境的置换。

判决性对比预测：这也是本文最核心的可裁决预测。如果正在全代偿环境中成长的从业者，在首次面对真实的不可撤回情境之前，仅需追加一系列高仿真但无不可逆后果的练习，便能在真实情境中表现出与在担责环境中成长的对照群体同等的初始独立裁断质量，则成功性脆弱可被纯认知卸载理论覆盖。如果上述复训尝试效果显著更低或无效，而必须将该从业者重新暴露于“错了真的会死”这一只属于肉身生命而非信息符号的条件时，其神经系统方能开始重建该裁断能力，则本文的分离线成立。

（六）与偏差正常化对质

戴安·沃恩在对挑战者号灾难的里程碑式研究中，提出了“偏差正常化”这一概念：组织内部对技术异常的接受标准，在一次又一次“未造成灾难”的成功经验后逐渐漂移，使得原本被视为不合规的状态被反复重新定义为可接受。这一概念是本文所面对的最近、也最强的近邻。一个严厉的批评者完全有理由说：你所谓的“成功性脆弱”，不过是偏差正常化的另一个名字。

二者的区分在于漂移的对象。偏差正常化所描述的，是组织行为偏离了原初的技术标准，而每一次未造成灾难的飞行都将这一偏离进一步“正常化”了。它的参照系是固定的——原初标准在那，只是人对它的遵守松动了。成功性脆弱所描述的，则是标准本身在历史成功经验中发生了位移：组织不是在偏离标准，而是在成功中重新定义了什么算作“安全”的标准。感知阈值抬升后，以前被认为值得警觉的信号，现在不再被视为“偏差”；代偿层闭合后，以前被视为完整裁断过程的一部分的独立判断，现在被视为可以被清单替代的冗余动作；叙事层加固后，以前被视为侥幸的未遂事件，现在被

读作“风险已被管控”的证据。在每一步位移中，组织都在依据既往的成功经验对自己进行理性的、正确的重新校准——因而事后找不到任何一个可以被指认为“违规”的行为。

这意味着一个关键的判别预测：在偏差正常化的典型案例中，由于标准本身并未改变，组织内部应当能够找到对漂移的“零星觉醒者”——那些知道标准在被偏离、并试图发出警告的人。挑战者号发射前夜，工程师博伊乔利试图阻止发射，正是因为他清楚地知道密封环烧蚀是一个偏离了设计标准的异常。而在成功性脆弱的纯范例中，由于参照系本身已经随成功历史发生了位移，组织内部将找不到任何一个能识别当前漂移的从业者——即使该从业者具备充分的批判能力。因为在新的参照系内，一切看起来都是正常的、合规的、有先例支撑的。换言之：偏差正常化的标志是“有人知道但阻止不了”；成功性脆弱的标志是“根本没人知道——包括最警觉的人”。

判决性对比预测：若在一次灾难后的调查中，能识别出组织内部曾有人依据原初技术标准将某一异常明确判定为“不可接受”，而该判定被管理层的风险接受逻辑所压倒，则属于偏差正常化的解释范围，成功性脆弱不适用。若调查发现，在灾难发生前，组织内部没有一个从业者——包括那些以严谨著称的一线专家——曾将任何一个事后看来明显的危险信号判定为“不可接受”，因为所有信号在当时的参照系内均被合理且合规地评估为正常或已受控，则成功性脆弱的解释力成立。这一判别是可操作的：只需在事后调查中检查安全档案、会议记录和证人证词，即可区分“被压制了的警报”与“从未被任何人感知为警报”。

（七）与漂向失败理论对质

西德尼·德克尔在其《漂向失败》一书中，以大量航空与医疗案例展示了系统如何在正常运行中逐渐向灾难漂移，而每一次局部决策在当时都是合理的、有依据的。德克尔的“漂移”概念与本文的“成功性脆弱”在经验指涉上有相当程度的吻合：二者都追踪了一条由局部合理决策累积而成的全局灾难路径。

然而，二者之间存在一个分析层级上的分工。德克尔的“漂移”是一个现象学概念：它描述系统如何一步一步走向失败，拆解了每一步决策背后的局部合理性。它回答的问题是：“灾难是怎么一步一步发生的？”成功性脆弱则是一个机制概念：它要解释的是，为什么在特定的组织条件下，这一漂移过程是不可逆的，并且是不可从内部被察觉的。它回答的问题是：“是什么使漂移在持续成功中变得不可逆？”

二者的关系可以这样概括：漂移是路径，成功性脆弱是使该路径滑向不可逆的引擎。一个系统可以在没有成功性脆弱的情况下发生漂移——例如，一个组织的标准在外部环境变化后未能及时更新，导致行为与新环境之间的错位；这种漂移可以被外部审计或一个外部事件所纠正。但是，当一个系统的漂移是在由成功–反馈稀缺引擎驱动的条件下发生时，漂移便不再是可纠正的偏差，而成为不可逆的退化——因为纠正所需的参照系本身，已经在成功的自我论证中被重新定义了。德克尔描述了漂移的解剖学，本文提供了漂移动力学的发生条件。

判决性对比预测：若一个系统的漂移能在一次外部审计或一次未遂事故后被有效逆转——从业者的警觉恢复到漂移前水平，被废弃的独立裁断通路重新被制度性地激活——则漂移属于德克尔的理论所覆盖的范围，成功性脆弱不适用。若即使引入强外部干预（如重大事故调查、监管高压、领导层更

替），系统的感知阈值、代偿层依赖深度和叙事自我论证力仍无法显著恢复——因为恢复所需的“不可撤回担责的发生条件”早已被制度安排所注销——则成功性脆弱的不可逆性断言成立。

（八）本节收束：一个系统的判决性对比表

下表将成功性脆弱与七个近邻的不可还原性收束在一起，以提供一张清晰的对照谱系。

近邻理论	核心变量	本概念提出的控制变量	判决性预测
正常事故理论	可识别故障的复杂互动	无故障的系统性退化	事后无任何可识别的失效点
高可靠性组织理论	正念管理原则的实践	成功绩效对发生基础的注销	满足全部HRO标准的系统，裁断力低于担责训练对照组
韧性工程	吸收与适应能力	适应能力发生条件的丧失	不恢复担责情境则韧性干预无法复原裁断力
去技能化理论	概念权被剥离	发生生态从未建立	新一代从业者需初次生成，非再培训可恢复
认知卸载理论	认知任务委外	发生情境的置换	无不可逆后果的模拟训练效果显著低于真实担责暴露
偏差正常化	对固定标准的偏离	标准本身的位移	事后无法找到任何曾将异常判为不可接受的内部人员
漂向失败理论	局部合理决策的累积漂移	漂移的不可逆性引擎	外部强干预后裁断力仍不可恢复

三、经验追踪：一个纵深的行业案例与跨领域痕迹

（一）一个纵深案例：商业航空业自动化演进中的成功性脆弱（1980–2010年代）

前文已指出，成功性脆弱作为一个机制假说，其核心约束在于：它必须能够在灾难发生前被观测到，而非仅能在灾难回顾中被“归因”。本节选择商业航空业作为一个纵向案例，追踪正反馈引擎在三十年时间跨度内于三个界面上的渐进显现。之所以选择航空业，不仅因为其安全数据的完整性与公开性，更因为它是一个“成功性脆弱”最不易被指控为危言耸听的领域——过去五十年，商业航空的安全绩效确实实现了令人敬畏的跃升。正是在这片最灿烂的成功的土壤上，引擎的运行痕迹最值得被仔细察看。

1. 背景：从手动操纵到飞行管理系统

20世纪70年代末至80年代初，波音757/767与空中客车A310的相继服役，标志着商业航空驾驶舱从“钢丝绳与液压”的手动操纵时代，大规模转入基于飞行管理系统（FMS）与电子飞行仪表系统的自动化时代。这一转型的初衷是降低机组工作负荷并减少因人为操作失误造成的事故。此后三十年间，自动油门、飞行指引仪、地形预警与自动着陆系统逐层叠加，飞行员在正常航班运行中亲手操纵飞机——尤其是在低空进近与着陆阶段——的时段，被逐年压缩至以秒计算。

2. 引擎三界面的展开

感知层：自动化的表现越可靠，机组对自动化失败的警觉就越弱。在80年代初自动化引入初期，飞行员对任何仪表不一致都抱有极大警惕，因为彼时对自动化的信任尚未建立。然而，随着一代代飞行管理系统在数百万次飞行中证明了其远低于人类的操作失误率，机组的预测模型被持续校准为“系统正常”这一高阶先验。2009年法国航空447号航班的一个细节提供了鲜明的注脚：当空速管结冰导致自动驾驶断开并触发失速警告时，机组未能及时识别飞机已进入失速——因为在他们的整个职业生涯中，飞机从未在正常法则保护下失速过。长达二十年的成功经验已将失速这一事件归入了“系统会自行防止其发生”的范畴，因而当它真实发生时，感知被延迟了致命的关键几秒。

操作层：代偿层界面在航空业的闭合过程，可以清晰地划分为三个阶段。第一阶段（80年代）：自动化作为辅助工具，飞行员保留在全航段手动操纵的完全能力，模拟机训练中包含大量手动飞行与系统失效应急处置。第二阶段（90年代至2000年代中）：航空公司为确保节油与合规，开始普遍推行“全程使用最高自动化等级”的政策；此时手动操纵虽仍被允许，但已受到软性约束——飞行员在职业生涯中积累的真实手动飞行时间急剧下降。第三阶段（2000年代后期至今）：多起事故和严重征候的调查指向了同一个现象——当自动化突然断开时，机组的手动操纵能力与对飞机当前能量状态的情景意识，已出现令人不安的代际性衰退。这不是任何一位飞行员个人的过错；这是在一代职业飞行员的全职业生涯中，发生独立裁断能力所需的发生条件——大量、无脚本、带有真实后果的手动操纵经验——被系统性地移除了他们的日常工作之外。当美国联邦航空管理局在2010年代后期开始推行“强化手动操纵训练”的倡议时，它所试图恢复的，正是被二十年的成功合规所闲置了的那条神经通路。

叙事层：航空业的安全叙事在三十年成功绩效的滋养下，变得异常强大：“航空是有史以来最安全的交通方式”这一陈述，是真实的，也是雄辩的。这套叙事的成效是，每一次未致灾的征候——比如一次因飞行员及时接管而避免的自动化错误——都被纳入了“系统的多层防护在奏效”的解释框架，而不是被读作“系统正在面临其操作者已越来越不熟悉的异常模式”的警告信号。当叙事层将每一件侥幸都编码为成功的证据时，它便完成了对其他两层位移的最终锁定。

3. 近距离审视一桩未遂事件

2008年1月，英国某航空公司的一架波音777在希思罗机场进近时，双发突然失去推力，飞机在跑道头前接地。事故调查发现，燃油系统内部结冰堵塞了燃油–油热交换器，而这一结冰机制在之前的飞行中被多次局部触发，但因每一次都未造成推力损失，因而被归入“已知并受控”的类别。这一部分是叙事层驯化的清晰痕迹：一个反复出现却从未致害的信号，被归档为“已知并受控”，于是它所指向的未知被消解了。 但这起事件的另一半，恰恰是本文必须诚实交出的一个阴性观测，且它比前半段更重要。在推力丧失后的数十秒内，机组并没有退回到系统所能提供的有限信息里：他们判断出以现有构型无法飞到跑道，于是把襟翼从30收到25以减小阻力、延长滑翔距离——这是一个没有任何检查单规定、也来不及查阅任何规程的临场裁断，事故调查报告对此予以肯定（AAIB，2010）。按本文的引擎模型，一个代偿层高度闭合的行业本不该在这个位置上长出这样一个动作。因此本案例对本文构成的是约束而非支持：它表明操作层的闭合即使在最深的自动化行业中也并非均质——它在个体与机组层面留有未闭合的缝隙，而这些缝隙的分布规律（是谁、在什么样的早期经历之后、还保有这类裁

断）本文尚无法解释。任何声称能够解释它的框架，都必须先解释这个动作是从哪里来的。本文把这一条列为自己最主要的未决问题。

必须澄清：航空业的三界面闭合并未完成。它的独特之处在于，一系列惨痛的教训——加上监管机构、工会与训练组织间持续的低度对抗——使得感知阈值在多次灾难后被临时拉回、代偿层被部分重启（如前述的强化手动训练）、叙事层被迫承认“自动化依赖”是一个真实的问题。但正因如此，航空业才成为一个绝佳的观测场：它展示了一台在周期性外部干预下被反复减速、却从未被真正关停的引擎的运转痕迹。引擎减速的证据，恰恰是引擎存在的证据。而那些未经历过类干预的其他高风险行业——如部分自动化高度闭环的金融交易系统——成功性脆弱可能早已完成更深层的闭合，只是尚未被灾难暴露。

（二）跨领域痕迹

除航空业外，正反馈引擎在其他高风险领域的运作痕迹同样清晰可见。以下三个片段简要展示了引擎在各自领域内的运作形态，以验证其跨领域的普遍性。

核工业领域。1979年三里岛事故中率先失效的那一类卸压阀，在此前的运行史中已经发生过卡开：同型机组在1977年就出现过同质事件，制造方内部也曾有工程师就此写过备忘。事后的总统调查委员会认定，这些既有经验并没有被转化为运行人员在现场可用的判据（Kemeny Commission, 1979）。这里的关键不是记录的缺失——记录是完备的——而是记录所能触发的系统性警觉，已在连续的“零严重事故”运行史中被降格为局部的操作确认：信号接收的硬件完好运转，而它本该指向的未知被叙事层吸收了。需要说明的是，本文无法从公开材料中核出此类信号在该电站的确切累积件数，因此不在此给出任何数量断言；本文所依据的只是调查委员会关于“既有经验未转化为现场判据”的这一项认定。

航天领域。1986年挑战者号航天飞机爆炸前的全部飞行历史中，相同型号固体火箭推进器的密封环烧蚀和热气泄漏已反复出现。每一次评估的结论都是“虽观察到损伤，但未导致任务失败”。每一次“完成任务”的判定，都在成功叙事中转化为“现有设计安全裕度充足”的新证据。发射前夜，工程师博伊乔利试图以个人判断阻止发射，但他无法提供被裁定为“足以改变发射指令”的数据化证据。在他的认知框架内，密封环烧蚀是一个明确的偏差——这正是偏差正常化的典型标记。但在他身处的决策系统的认知框架内，成功叙事已经将其重新定义为“已知风险，可接受”。这是一个处于偏差正常化与成功性脆弱交界处的案例：博伊乔利是最后一位抱持原初参照系的觉醒者，而整套决策系统已经在成功中位移至一个新的标准。

金融领域。2008年全球金融危机前的数年间，可量化的市场风险、信用风险、流动性风险均在全球各大金融机构的公开报告中处于被认为严控的低位。风险价值模型对高频低波动数据的依赖，使感知层对尾部依赖结构和流动性突然蒸发的可能性毫无警觉；以风险管理模型为核心的代偿层界面让交易员与决策者不必亲身承受巨额不可撤回裁定的身体性警觉；持续攀升的利润和逐年改善的风险报告则将任何外部危险信号驯化为“已被定价”的项目。正反馈引擎在海啸触及礁石之前，将整个系统牢牢锁在“完美”之中。

上述经验材料并非作为“验证”成功性脆弱的统计学证据——它们本身是灾难或侥幸后的回顾。它们的价值在于提供了引擎在不同领域各自独立运作的证据痕迹；而真正的检验，在于未来能够在大规模灾难发生之前运行本文所提出的前瞻性研究设计。

四、边界与反驳

（一）引擎在何处激活：三项边界条件

成功性脆弱并非所有专业系统的默认命运。它只有在同时满足以下三项边界的组织中，才会被充分激活。第一，代偿层取代而非辅助：系统的日常运行在结构上高度依赖制度化防护层为从业者提供决策依据，且从业者极少有正当制度许可绕过代偿层直接依据个人裁断采取行动。第二，可审计安全指标持续长时间改善，或至少未出现被当前指标可捕捉的持续恶化。第三，系统的危险清单中存在一类可被物理地定义、但其触发形态与处置条件完全落在当前任何规程设计之外的无先例事件，且此种事件一旦发生，其后果是根本性的。三项边界同时成立，正反馈引擎便获得充足的燃料与封闭的运行空间。

继而必须补充一条关键约束：机制断言本身的可观测性，并不依赖于灾难的发生。本文断言的是——在满足三项边界的系统中，从业者的初始独立裁断力将因发生条件的丧失而呈现可测量的退化。这一退化是否实际导致了灾难，是概率问题；但退化本身的存在与否，是机制问题。因此，该机制应当在尚未发生灾难的高绩效系统中，被独立的测量工具（如模拟无先例极端情境下的裁断质量评估）所捕获。这一观点将直接导向本节的阴性案例讨论与本文结论部分的前瞻性检验设计。

（二）阴性案例与前瞻性检验设计

本文的实证案例绝大部分取自已发生灾难或严重征候的高绩效系统。这一选择性偏差——只在脆弱性已显现的样本上验证了脆弱性机制——是当前实证基础的最薄弱一环，必须诚实地自陈并给出控制路径。

阴性案例的可能位置是：在部分尚未完全由制度化代偿层覆盖其核心操作的专业系统中——例如某些仍保留高强度真实担责训练的军事航空单位、某些仍以多年师徒制将新手逐步暴露于不可逆裁断情境的外科传统——即使其日常安全绩效同样优越，从业者在面对无先例事件的初始裁断质量，可能并未呈现本文所描述的那种代际衰减趋势。这些系统中，真实担责情境尚未被完全移除出日常训练的发生条件，因而正反馈引擎在操作界面上的闭合并未完成。

但这仅仅是定性指认。要使本文的机制断言能够被转化为一个可验证的前瞻性研究纲领，需要给出一个更具体的检验设计草图如下。

研究设计。以“代偿层覆盖度”为自变量，按操作标准化程度、从业者绕过代偿层的制度许可度、以及从业者在日常工作中亲身承担不可逆裁断的频率三个指标，将若干个具有同等高水平日常安全绩效的专业系统分为高代偿层组与低代偿层组。因变量为从业者在面对“无先例极端情境”时的初始独立裁断质量。该情境需满足三个标准：属于该系统领域内公认的危险，但从未被编入现有规程、检查清单或训练脚本中；具有真实的时间压力与无可回避的裁断要求；受试者的裁断行为与结果可被客观记录并评分。控制变量包括日常绩效水平、从业年限、近期操作频率等。

预测。高代偿层组从业者的初始独立裁断质量将显著低于低代偿层组从业者，且该差异在控制日常绩效水平后仍然显著。附加预测：若在高代偿层组内部，将从业者随机分配为两组——一组接受高仿真但无不可逆后果的模拟训练，另一组在真实且不可逆的担责情境中暴露一段时期——则后者的裁断质量提升将显著大于前者。这一交互预测，旨在剥离纯认知卸载效应与成功性脆弱机制的核心差异。

这一设计将本文的机制假说直接推向了可操作的实证检验场，使得“代偿层覆盖度”成为一个可被独立测量、并可产生二阶交互预测的控制变量。这就是本文的实证纲领。

（三）根本性的反驳与防御

本文必须回应的最核心反驳是：文中所描述的机制，最终难道不仍是已被裁定的“偏差正常化”的某种变体？本节将从概念与经验两个层面给予正式回应。

在概念层面，二者区分已在第二节详述，此处不赘。在经验层面，本文的判决性预测是可操作的：如果在一场灾难的调查中，能找到组织内部曾有任何人将那些关键异常判定为“不可接受”，则归入偏差正常化的谱系；如果找不到任何人，则成功性脆弱的解释力更强。挑战者号之所以是一个典型的偏差正常化案例，恰恰因为博伊乔利在场——他清醒地知道标准正在被违背，他的警报被记录在案。而在满足成功性脆弱三项边界条件的纯范例中——比如一个代偿层已高度闭合、且长期零事故运行的系统——本文预言：未来其事后调查将找不到这位博伊乔利。这不是因为那个从业者不勇敢，而是因为在那个系统的成功位移后的参照系内，他无从将异常识别为异常。换言之，偏差正常化的本体是“有警报但被压下”；成功性脆弱的本体是“警报从未生成”。

（四）关于“完全自动化系统”与“高绩效反例”的回应

两个补充性反驳需在此处得到处理。其一：如果一个专业系统全部由自动化决策组件构成，是否仍有成功性脆弱？本文的回答是：如果主体是人类从业者，那么成功性脆弱的生成，依赖于长期成功中不可逆担责情境的缺失对神经基础的效应。如果主体是完全无神经基础的AI系统，则本机制不适用——但那类系统将面临另一类同构的、以“训练数据分布锁定”为引擎的脆弱性，那是另一篇论文的论域。

其二：某些高绩效系统——如早期以色列空军的战斗机部队、某些外科精英团队——是否在数十年高绩效中保留了强独立裁断力？若然，本文的边界条件是否需要修正？本文的回应是：这些系统恰恰是本文边界条件第一项——“代偿层取代而非辅助”——并不完全成立的案例。在这些系统中，制度化代偿层要么从未覆盖核心裁断环节（空战中不可预编程的战术决策），要么被制度性地设计为“在常规训练中不被作为首要依靠”（外科师徒制中主刀医生作为最终责任人）。它们的存在，不是对本机制的证伪，而是对本机制边界条件的实证。一套理论只有当它精准地指出了“在何处不适用”时，方能在“在何处适用”上立稳。

五、结论

当组织的全部注意力都集中在压缩可识别的故障时，对未知风险之真正免疫力的发生基础，却在完美指标的静默中逐年丧失其可被重新激活的物质条件。这便是成功性脆弱的核心判断。

本文已论证：成功性脆弱不是对既有安全科学概念的替代，而是对一个集体盲区的机制性补充。安全科学已有的全部探测仪器，均部署在“失效”这一维度的不同位点上；而成功性脆弱所指向的那类危险，在日常运行中不会触发任何一部探测仪器的阈值。它不是故障链的一环，而是成功弧线的负侧影子。本文给出了这一机制在感知、操作与叙事三个界面上的运作原理及其时间–序列结构，通过与七个最强近邻的严肃对质析出了其控制变量——“标准本身的位移”，以一个纵深的行业案例与跨领域痕迹展示了引擎的实际运作，并自陈了实证盲区、给出了前瞻性检验设计。

最后，以一句严格的证伪条款收束全文：如果未来在任何完全满足本文所述三项边界条件的真实专业系统中，通过独立的模拟无先例极端情境测量，发现其从业者的初始独立裁断质量不低于、甚至优于在同等操作强度却在低代偿层环境中持续被暴露于真实担责压力的对照群体，则本文的核心机制——“成功–反馈稀缺”引擎驱动了不可逆的裁断力退化——即被推翻。在那一天之前，成功性脆弱是检验每一座由完美安全指标撑护起来的大厦时，不应被绕开的一个脆弱性假说。

附论零 · 站内划界：与《代偿性闭合》的分离线

本节处理一个比任何站外近邻都更近的对手：本文作者本人已经发表的《代偿性闭合：生成丰裕时代的能力萎缩机制》（2026年7月24日）。那篇论文已经立下了一条与本文高度同源的母命题——生成系统在个体尚未长出判断力的那个生成时间窗口内，就用流畅的输出把它盖住了；它也已经明确声明所诊断的“生成空缺”既不同于技能退化，也不同于认知卸载。如果本文只是把这条母命题换一个场景复述一遍，那么本文的不可还原性主张就不成立。因此这条分离线必须先于所有站外对质被划清。

这一条对本文尤其严厉，因为本文所使用的核心术语之一“代偿层闭合”，在字面上就是作者本人前作标题的上移。必须明说：**“闭合”这个词是从《代偿性闭合》搬来的**，本文不主张它是一个新造的词。本文所主张的是，把它从个体—工具层搬到组织—制度层之后，有两样东西是搬不过来、必须重新造的，而这两样东西构成本文的全部增量。

第一样是**时序**。前作的三道屏障是并置的：它们各自失效，失效的叠加逼出回路，前作明确写道三者的耦合“不是偶然的同时在场”，但它没有给出三者之间谁先动、谁被谁驱动。在个体层面这是可以接受的，因为三道屏障作用于同一个人的同一次调用。在组织层面则不然：感知层、操作层与叙事层分属不同的载体（个体的注意力、工作流程的界面、管理层的自我理解），它们之间的传递需要具体的因果通道。本文因此必须给出并且给出了一条时序——感知层先动，因为它最直接暴露于反馈稀缺且不需要任何人做决策；感知钝化截断了启动独立通路的内驱力，操作层继之闭合；闭合产生的完美绩效成为叙事层所需的证据，叙事层最后收紧并反过来加固感知阈值。这条时序是可证伪的：它预测在一个正在滑入的组织中，**弱信号报告的降级早于手动裁断频次的下降，而后者又早于安全叙事的定型**。三者的时间戳都留在档案里（安全委员会会议记录、操作日志、年度安全报告的措辞），一次纵向档案研究即可查。若发现叙事层最先定型，本文的时序即错。

第二样是**控制变量**。前作的控制变量是屏障的失效程度；本文的控制变量是“标准本身的位移，而非对标准的偏离”。这个变量在个体层面没有对应物——一个人不会“重新定义什么算判断”，他只是不判断了；而一个组织会，并且组织重新定义之后，每一个成员的每一步都仍然是合规的、可辩

护的、正确的。本文最锋利的那条判决性预测——事后调查将找不到那位吹哨人——只有在组织层面才有意义，前作既提不出也不需要它。

最后必须交代一处本文相对于前作的**退让**。前作论证了闭合回路的一个特殊后果：低质调用反向稀释生成系统本身的锋度，从而使回路在技术侧也自我加深。本文所讨论的组织系统没有这个通道——安全规程不会因为被平庸地执行而自己变钝。因此本文的引擎比前作的回路少一条腿，其自我维持完全依赖人这一侧的三层咬合。这也意味着，本文的引擎在原理上比前作的回路**更容易被外部干预打断**——航空业三十年的经验正是如此：周期性的惨痛教训确实把它减速过多次。本文不主张组织层的退化比个体层更不可逆；本文只主张，在没有那类外部教训的行业里，它没有内部的刹车。

附论一 · 站外近邻补切（编辑增补）

原稿已就本领域内最常被援引的几家做过对质。按全球库口径重扫一遍之后，仍有若干家未上桌，而它们与本文判据的距离并不比已上桌的那些远。下列补切按“越近越先”排列；每一家给一条分离线与一条可裁决的对照预测。

零、模板层的对质：“越成功越失败”这个形式本身早有名字。在逐家对质之前，先做一道更靠前的检查：把本文的命题去掉一切组织与安全领域的名词，只留“什么东西对什么东西做了什么，导致什么”，得到的纯形式是——**一个系统因持续成功而丧失应对新情形的能力**。这是一个反转形式，而重命名最容易藏在模板层而不是内容层：这个形式在组织研究里已经被命名过至少四次，每一次都有确切的出处。不点名它们而径直命名，等于把一个已有三十到六十年历史的结构重新发明一遍。原稿一家也没有点到，这是它最实在的一处缺口。

（一）成功陷阱（Levinthal & March, 1993）。《学习的近视》指出组织学习有几重近视，其中之一是：在既有做法上的成功会提高对该做法的投入，从而进一步提高其成功率，把探索挤出去——成功自我强化为一个陷阱。这是四家里与本文最近的一家，因为它的因变量也是应对新情形的能力。需要说明的是，正文所引的“能力陷阱”出自 Levitt 与 March（1988），而“成功陷阱”这个更贴本文的命名出自 Levinthal 与 March（1993），二者不可互替。 分离线在**是否知情**。成功陷阱是一个学习的正反馈：组织的每一步都是对回报的合理响应，它**能够说出**自己在放弃什么（探索的期望回报更不确定），只是不去做。本文所述的组织不是权衡后不探索——感知阈值已经把那些本可以指出方向的弱信号滤掉了，它**说不出**自己在放弃什么。 **判决性对照预测**：即附论二第三条。成功陷阱预测组织答得出并能说出理由；本文预测答不出，且被追问时会援引安全绩效作为不必回答此问的理由。

（二）伊卡洛斯悖论（Miller, 1990）。米勒论证企业常常**恰恰因为**其最大的长处而衰败：驱动成功的那一项特质被推向极端，沿几条轨迹走向失败——工匠型走向偏执的技术至上，开拓型走向漫无节制的扩张。 分离线在**有没有一样东西被推到极端**。伊卡洛斯的机制是**放大**，其可观测标志是某项特质在事后可被清晰指认为“他们太执着于这个了”，而且这个指认在事发之前对外部观察者往往已经可见。本文的机制是**标准本身的位移**：没有任何一项特质被推向极端，每个人的每一步都仍然温和、合规、可辩护；正因如此，它不产生任何信号。 **判决性对照预测**：在事后调查中，伊卡洛斯预测能指出至少一项被推到极端的组织特质，且该特质通常在事发前的行业评论里已被提及；本文预测指

不出——事后所能找到的只是一系列各自正常的判断。这一条与本文那条“找不到吹哨人”的预测方向一致而内容不同：吹哨人是**人**，被推到极端的特质是**属性**，两者可以分别查。

（三）核心能力的刚性面（Leonard-Barton，1992）。该文指出核心能力有其反面：同一套使组织擅长既有业务的能力，会在新产品开发中变成阻碍变革的刚性。 分离线在**能力是在还是不在**。刚性理论的因变量是**已有能力抵抗新任务**——能力完好，只是被用错了地方、或不肯让位。本文的因变量是**能力本身未曾生成或已经退化**。二者可用一个很干净的设计分开：给一项与既有能力完全无关、也不触动任何既有路径的全新任务。 **判决性对照预测**：刚性理论预测在这类无冲突的新任务上表现良好（刚性只在与既有能力冲突处发作）；本文预测表现仍差，因为裁断力本身不在了。若观察到表现良好，本文的诊断错，该组织属刚性而非退化。

（四）内卷化（Geertz，1963）。格尔茨用它描述爪哇农业的一种形态：在既有系统内持续追加劳动投入，产生越来越精细的内部分化，而不发生突破。 分离线在**投入曲线的方向**。内卷化的可观测标志是**投入量在持续增加**——它首先是一个关于劳动吸纳的判断。本文的引擎不要求投入增加：一个安全投入平稳、甚至因绩效良好而被削减的组织，照样会滑入本文所述的三层闭环。 **判决性对照预测**：查事发前十年的安全投入曲线。内卷化预测单调上升；本文预测持平或下降，而弱信号报告的降级仍在同期发生。两条曲线的方向不同，且都留在预算与安全委员会的记录里。

这一节承认了什么。上述四家共同占据了“成功导致失能”这个形式的绝大部分，本文在这个形式上没有任何新东西。本文相对于它们的增量只在两处：一是**控制变量**——标准本身的位移，而不是对标准的偏离、不是特质的放大、也不是投入的追加；二是由它导出的那条可执行判据——**事后调查将找不到那位吹哨人**。四家里没有一家做得出这条预测，因为它们的机制都允许、甚至要求存在一个看得见问题的内部人：成功陷阱里有人算过账，伊卡洛斯里有人早已提醒过那项特质，刚性里有人抱怨新路走不通，内卷化里有人看着投入年年在加。本文的位移则不给任何人这个位置。

一、明斯基的金融不稳定假说（Minsky，1986，1992）。本文以2008年危机作为跨领域痕迹之一，却未提明斯基，这是原稿最刺眼的缺口——因为“长期的平稳本身孕育了脆弱”这句话，在经济学里就是他的。他的命题是：持久的繁荣使债务结构从对冲型逐步滑向投机型与庞氏型，系统的脆弱性因此在没有任何人犯错的情况下内生增长。 分离线在**脆弱性长在哪里**。明斯基的脆弱长在**资产负债表**上：它是一个可以被外部观察者直接测量的财务结构性质，与从业者是否还会裁断毫无关系；本文的脆弱长在**人的裁断力**上，是一个认知—操作—叙事的过程。二者可以完全脱耦。 **判决性对照预测**：考察一个杠杆结构与期限错配均未恶化、但代偿层高度闭合的机构（例如一个高度自动化、以规则引擎驱动的合规或风控部门）。明斯基的框架预测其不脆——因为财务结构健康；本文预测其在面对一类无先例的、落在规则引擎训练分布之外的异常时，初始裁断质量仍显著低于对照。这一条把两个同名的“成功导致脆弱”彻底分开。

二、马奇的能力陷阱与探索/利用（Levitt & March, 1988; March, 1991）。能力陷阱是组织学习领域对本文命题最直接的先占：组织因在既有做法上不断成功而不断加码，从而系统性地不再探索替代方案。 分离线在**知情与否**。能力陷阱是一个**权衡**的结果：组织知道存在别的做法，只是因为既有做法的即时回报更确定而选择不去。它在原理上是可以被一次算清收益的决策所纠正的。本文所描述的组织不是权衡后不去探索——它**不知道有什么可探索的**，因为感知阈值已经把那些本可以指出方向的弱信号滤掉了。 **判决性对照预测**：直接问组织的中层与一线：“你们知道自己正在放弃哪些替代方

案吗？”能力陷阱预测答得出，并能说出放弃的理由；本文预测答不出，且被追问时会援引安全绩效作为不需要回答此问题的理由。这一条只需一次半结构访谈。

三、斯努克的实践漂移（Snook，2000）。本文对质了德克尔的漂向失败，却漏了比它更早也更近的这一家。斯努克在对黑鹰误击事件的研究中提出，局部实践会在长期无事故的条件下与全局规程逐步脱钩，直到某一刻规程与实践之间的落差致命。需要特别提醒的是，该概念的中译常作“操作漂移”，与本文的“代偿层闭合”并非同义，不可混用。 分离线在**落差是否可测**。斯努克的漂移，其可观测标志正是实践与规程之间存在一个可测量的落差——把规程文本与实际操作记录对读，落差就在那里。本文的位移是规程与判据**本身随成功史一起移动**，因此对读的结果是二者高度一致，落差测不出来。 **判决性对照预测**：这是本文最容易执行的一条档案研究。在一场灾难之后，把事发前十年的规程文本与实际操作记录逐年对读。斯努克预测落差随时间单调扩大；本文预测落差始终很小，而**规程文本自身的判据在逐年放宽**（同一类信号的报告门槛、同一类异常的处置等级在文本中被逐版下调）。两条曲线的形状不同，且都留在档案里。

四、塔勒布的脆弱与反脆弱（Taleb, 2012）。读者见到“成功性脆弱”这个名字，第一反应必定是塔勒布，因此必须处理。分离线是干净的：塔勒布的脆弱是一个**数学性质**——收益函数对波动的凹性，它与组织是否有人还会裁断无关，也不需要任何关于认知、制度或叙事的假设。塔勒布确实谈到过长期平静掩盖尾部风险，但他把不可见性归因于叙事谬误这一**一般性的人类认知病**，并不追问它在具体组织中由什么机制逐层生产出来。本文的全部增量恰在这里：给出不可见性的生产机制与时序。**判据**：若一个组织的收益结构呈凸性（即从波动中获益）却仍出现本文所述的三层闭合与裁断力退化，则脆弱性的数学定义与本文的发生学定义确属两回事，本文的命名有独立存在的理由；若二者总是同时出现，本文应当退回为塔勒布框架下的一个组织学注脚。

五、维尔达夫斯基的预期与韧性（Wildavsky，1988）。一条短切。他关于“预期”（事先堵住已知风险）与“韧性”（保留应对未知的余力）此消彼长的论证，在结构上与本文的边界条件第二、三项接近。分离线在于：他讨论的是资源在两种策略间的**分配**，是一个可以被决策者自觉权衡的问题；本文主张三层闭环之后，决策者已经失去了识别“我们正在把余力换成预期”的位置——分配问题变成了不可见的问题。

附论二 · 可裁决预测（编辑增补）

正文已给出三项边界条件、一条正式证伪条款与一个前瞻性研究设计。以下四条是本次补切新增的、与特定近邻直接对撞的可裁决预测，且四条全部可在现有档案上执行，不必等待灾难。

1. 落差不扩大，而判据在放宽（最容易执行的一条）。把事发前十年的规程文本与实际操作记录逐年对读。斯努克的实践漂移预测二者落差随时间单调扩大；本文预测落差始终很小，而规程文本自身的判据在逐年放宽——同一类信号的报告门槛、同一类异常的处置等级在文本中被逐版下调。两条曲线形状不同，且都留在档案里。**若落差单调扩大而判据未放宽**，则属斯努克的解释范围，本文不适用。

2. 财务健康而裁断仍脆。考察一个杠杆结构与期限错配均未恶化、但代偿层高度闭合的机构。明斯基的框架预测其不脆；本文预测其面对落在规则引擎分布之外的异常时，初始裁断质量仍显著低于对照。**若二者总是同涨同落**，本文与明斯基的假说无从分开。

3. 答不出放弃了什么。直接询问组织中层与一线：“你们知道自己正在放弃哪些替代方案吗？”马奇的能力陷阱预测答得出并能说出理由；本文预测答不出，且被追问时会援引安全绩效作为不需要回答此问题的理由。**若答得出**，则组织是在权衡后不去探索，属能力陷阱，本文的感知层机制多余。

4. 三层的先后次序留在档案里。本文的同源性时序主张是可查的：它预测弱信号报告的降级早于手动裁断频次的下降，而后者又早于安全叙事的定型。三者的时间戳分别落在安全委员会会议记录、操作日志与年度安全报告的措辞里。**若发现叙事层最先定型**，本文的时序结构即错，三条原理将退回为三个并置的相关现象。

编辑校订说明

本次编辑校订共六处，其中三处涉及可核验事实，必须公开说明，其中第一处尤为重要：①第三节原文称2008年希思罗那起未遂事件的调查有一项“附属发现”，即机组放弃直接感觉、几乎完全依赖飞行管理系统。经核，公开的调查报告所记录的与此相反：机组在数十秒内做出了把襟翼由30收至25以延长滑翔的临场裁断，报告对此予以肯定。该段已整体改写——不再作为支持本文的正面观测，而作为本文必须诚实交出的一个阴性观测，并据此把“操作层闭合并非均质”列为本文的主要未决问题。②三里岛一节原文称“事故前数年累积数百起完全同质的小型异常信号、无一遗漏”，该数量断言无法从公开材料核出，已删去数字，改为总统调查委员会确曾认定的那一项内容（既有经验未被转化为运行人员在现场可用的判据），并注明本文不给出任何数量断言。③博帕尔一例被置于金融段落之末，且该事件属防护层被人为拆除，恰好落在本文自设的第一项边界条件之外，属反例误用，已删。④“本文未能在初稿中将其纳入对质”属草稿痕迹，已删。⑤“成功性弱性”错字，已改。⑥摘要末尾一处病句，已理顺。另：原稿未附参考文献表，而正文点名了十余家理论与四起事故调查，现由编辑据正文所引逐条补齐；凡本次无法核实出处者，已在正文中相应降级表述，不列入表内。
又：本文的纯形式是“越成功越失败”这一类反转模板，而该模板本身在组织研究里早有命名。原稿与初次增补均未点名其正主，属重命名风险最高的一处遗漏；现于附论一之前补入“模板层的对质”一节，逐一处理成功陷阱、伊卡洛斯悖论、核心能力的刚性面与内卷化四家，并明确承认本文在这个形式上没有新东西、增量只在控制变量与那条“找不到吹哨人”的判据上。

参考文献

AAIB (2010). Report on the accident to Boeing 777-236ER, G-YMMM, at London Heathrow Airport on 17 January 2008. Aircraft Accident Report 1/2010. Air Accidents Investigation Branch.

BEA (2012). Final Report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France, flight AF 447. Bureau d'Enquêtes et d'Analyses.

Braverman, H. (1974). Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century. Monthly Review Press.

Dekker, S. (2011). Drift into Failure: From Hunting Broken Components to Understanding Complex Systems. Ashgate.

Geertz, C. (1963). Agricultural Involution: The Processes of Ecological Change in Indonesia. University of California Press.

Hollnagel, E. (2014). Safety-I and Safety-II: The Past and Future of Safety Management. Ashgate.

Hollnagel, E., Woods, D. D., & Leveson, N. (Eds.) (2006). Resilience Engineering: Concepts and Precepts. Ashgate.

Leonard-Barton, D. (1992). Core capabilities and core rigidities: A paradox in managing new product development. Strategic Management Journal, 13(S1), 111–125.

Levinthal, D. A., & March, J. G. (1993). The myopia of learning. Strategic Management Journal, 14(S2), 95–112.

Levitt, B., & March, J. G. (1988). Organizational learning. Annual Review of Sociology, 14, 319–340.

Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S. J., & Frith, C. D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. Proceedings of the National Academy of Sciences, 97(8), 4398–4403.

March, J. G. (1991). Exploration and exploitation in organizational learning. Organization Science, 2(1), 71–87.

Miller, D. (1990). The Icarus Paradox: How Exceptional Companies Bring About Their Own Downfall. HarperBusiness.

Minsky, H. P. (1986). Stabilizing an Unstable Economy. Yale University Press.

Minsky, H. P. (1992). The Financial Instability Hypothesis. Levy Economics Institute Working Paper No. 74.

Perrow, C. (1984). Normal Accidents: Living with High-Risk Technologies. Basic Books.

President's Commission on the Accident at Three Mile Island (1979). The Need for Change: The Legacy of TMI (Kemeny Commission Report). U.S. Government Printing Office.

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

Snook, S. A. (2000). Friendly Fire: The Accidental Shootdown of U.S. Black Hawks over Northern Iraq. Princeton University Press.

Taleb, N. N. (2012). Antifragile: Things That Gain from Disorder. Random House.

Vaughan, D. (1996). The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA. University of Chicago Press.

Weick, K. E. (1993). The collapse of sensemaking in organizations: The Mann Gulch disaster. Administrative Science Quarterly, 38(4), 628–652.

Weick, K. E., & Sutcliffe, K. M. (2007). Managing the Unexpected: Resilient Performance in an Age of Uncertainty (2nd ed.). Jossey-Bass.

Wildavsky, A. (1988). Searching for Safety. Transaction Books.

鲍锦朝（2026）. 代偿性闭合：生成丰裕时代的能力萎缩机制. SDE Universes 学员专栏，2026年7月24日.