第八章　越审越不能改

《越顺利，越没有你的位置》· 王德生 + Claude 编 · 本章在全书中为第 215–239 页 · ISBN 979-8-90690-018-0

> 本章原题《越审越不能改：人工监督如何同时生产控制证据与控制失效》，作者　少敏，原文见 https://sdeuniverses.com/students/shao-min/review-forecloses-change/

摘 要

生成式人工智能治理已经不再把“有人看过”简单等同于有效监督：现有研究与欧盟《人工智能法》均区分监测、判断与可行干预。尚未被充分处理的困难在于，这些条件会在复核进行中共同变化。复核一方面提高错误识别率，并生成签名、异议、轮次、批准与日志等“人类参与证据”；另一方面，若焦点方案仍被允许进入工单、接口、预算、合同、排期和团队预期，复核的中间状态也可能被下游读取为继续投入的许可。组织于是同时获得更多控制证据和更少现实控制：它越来越能证明人参与过、知道过、签署过，却越来越难把人的异议转成另一条可执行路径。

本章以“反事实承载力”测量合格异议到来时，组织是否仍能把当前处置转向至少一条合格替代路径。它不是备选想法的数量，而是替代证据、承接权限、资源、可承担的转换成本与剩余时间共同维持的行动能力。由此，监督被拆成三类不能互相替代的结果：认识增益、控制证据与路径可改性。本章保留“控制反转”的强判决：同样增加复核，在冻结焦点方案新增专用依赖时提高有效控制，在允许依赖继续累积时反而降低有效控制；同时，两种条件都会生成更多复核轨迹。其机制是“临时合法性接力”——尚未批准的中间状态被下游当作继续投入的通行信号，使监督在留下更多可归责痕迹的同时，消耗自身赖以生效的反事实基础。

文章把有效监督、可争议性、路径依赖与实物期权作为机制近邻，把道德缓冲区、可追溯性与审计轨迹作为责任近邻。前一组已说明干预需要现实路径，后一组已说明有限控制的人可能承受过多责任；本章不重复这两项诊断，而提出二者的内生连接：复核过程本身可以同时生产用于归责的参与证据与造成失控的路径依赖。为使该连接可被推翻，本章给出复核强度×依赖冻结实验，以及第二阶段的档案盲评：评审者随机看到传统复核档案或加入反事实承载力时间曲线的能力档案。理论要求四项并存——识错率上升、复核轨迹增厚、替代路径衰减、实际控制下降；冻结依赖后只有控制效应翻正，轨迹仍继续增厚。若联合预测不成立，若传统档案与能力档案不改变责任分配，或既有近邻能够完整吸收结果，新增责任主张必须删除，论文退回动态可行性扩展。

关键词：生成式人工智能；人工监督；反事实承载力；控制证据；责任归属；路径依赖

When More Review Leaves Less to Change: How Oversight Produces Evidence of Control While Eroding Control Itself

Abstract

AI governance no longer simply equates the presence of a human reviewer with effective oversight. Existing scholarship and the EU AI Act distinguish monitoring, judgment, and feasible intervention. The unresolved difficulty is that these conditions can change during review itself. Review may improve error detection and generate signatures, objections, stage records, approvals, and logs that evidence human involvement. Yet, if intermediate states are read downstream as permission to continue investing in the focal option, the same process may erode the alternatives needed for intervention. An organization can therefore accumulate more evidence that a human participated, knew, and signed while retaining less capacity for that human's objection to redirect the current path.

The article measures counterfactual carrying capacity: an organization’ s capacity, upon receiving a qualified challenge, to redirect the current disposition into at least one admissible and executable alternative. It separates three outcomes that cannot substitute for one another: epistemic gain, evidence of human control, and path alterability. A control reversal occurs when added review increases effective control under a dependency freeze but decreases it when focal-path dependencies continue to accumulate, even though both conditions generate thicker review traces. The proposed mechanism is a provisional-legitimacy relay: an intermediate status that formally means “still under review” functions operationally as permission to continue investing, so oversight leaves more attributable traces while eroding the counterfactual basis on which intervention depends.

The article treats effective oversight, contestability, path dependence, and real options as mechanism neighbors, and moral crumple zones, traceability, and audit trails as responsibility neighbors. The first group shows that intervention requires feasible paths; the second shows that people with limited control may bear disproportionate responsibility. The proposed remainder is their endogenous connection: review can jointly produce attribution evidence and the path dependence that reduces causal control. A preregistered review-intensity × dependency-freeze experiment is paired with a second-stage dossier experiment in which evaluators receive either a conventional review record or a capacity-aware record that includes the time path of feasible alternatives. Support requires a joint pattern: epistemic gain and review traces rise while path alterability and effective control fall; under a dependency freeze, only the control effect changes sign. If that conjunction fails, if the two dossiers do not change responsibility allocation, or if existing neighbors fully absorb the result, the responsibility claim must be deleted and the theory downgraded.

Keywords: generative AI; human oversight; counterfactual carrying capacity; evidence of control; responsibility attribution; path dependence

一、监督为何可能在成功识错时失去作用

人工监督在制度文本中通常以一个令人安心的顺序出现：系统先给出输出，人再监测、理解、复核、决定是否干预。欧盟《人工智能法》第14条要求高风险人工智能系统能够被自然人有效监督，并使监督者能够理解系统能力与局限、警惕自动化偏误、忽略或推翻输出、干预或停止系统\[1\]。最新的有效人类监督框架也把监督写成“监测—风险评估—决定—干预”的控制回路，并明确区分信息条件、监督能力与干预可行性\[2\]。这些进展已经超越“界面上放一个人”的薄弱治理观。

困难不在现有理论完全没有区分，而在组织实践常用第三类结果替代前两类。第一类是认识上的成功：监督者发现错误、理解风险并形成合格异议。第二类是行动上的成功：该异议能够改变当前处置，而不只是生成说明、延迟、补偿或下一轮学习。第三类是证据上的成功：流程留下了足以证明某人被通知、参与复核、提出意见或签署批准的轨迹。认识成功回答“人是否看见”，行动成功回答“现实是否还能转弯”，证据成功回答“事后能否证明人曾在场”。三者可能同向，也可能分岔。若复核提高识别质量、增厚参与记录，却同时使替代路径失去资源和时间，制度便会用越来越强的在场证据，支撑对越来越弱的因果控制者的责任结算。

设想一份AI生成的软件变更方案。组织安排三轮人工复核：架构、信息安全和发布委员会依次检查。每一轮都发现了更多问题，也留下“可继续评估”的中间状态。与此同时，开发分支、测试脚本、部署窗口、客户通知和回滚计划开始围绕这份方案建立。第三轮出现决定性反证时，委员会完全理解风险，也拥有拒绝权限；但替代实现没有分支、没有测试资源、没有发布窗口，拒绝只会造成停服或违约。此时，监督在识别意义上更加成功，在记录意义上也更加完整，在改变结果的意义上却更加失败。第三轮复核者的签名、会议发言和风险知情全部可见；允许焦点路径在前三轮继续长厚的分布式行动反而不容易被压缩成一个可归责主体。

反过来，若同样三轮复核期间禁止新增方案专用依赖，并为至少一个替代方案保留证据预算、承接人和切换窗口，更多复核就可能同时提高识别与控制。差异不在监督者是否更聪明，而在复核过程中组织是否保存了异议可以进入的现实分岔。

本章由此提出一个条件性判断：

人工监督不是外加于决定过程的静态资源。监督期间若焦点路径继续获得专用依赖，监督可能一边提高识错能力和控制证据，一边消耗自身赖以改变结果的反事实基础；事后归责因而可能沿着最清楚的记录回流，而不是沿着最大的边际控制分布。

这一判断不主张监督通常有害，也不以取消人工复核为政策结论。它拒绝两个替代：既不允许用“留下记录”替代“保留控制”，也不允许用“承担责任”倒推“曾经具有足够控制”。它要求分别测量认识质量、控制证据、路径可改性与责任归属。

二、现有研究已经走到哪里

（一）有效监督已经包含“能否干预”，但尚未把监督写成可行性的生产者

有效人类监督研究已经明确，监测不等于控制。Gaube等人的跨学科框架把监测理解为证据积累，把干预理解为人对系统施加因果影响；干预不仅需要权限和技能，还取决于行动是否在当前语境中可用、可行，以及时间、努力、经济与声誉成本是否允许\[2\]。这与本章的出发点高度接近，也排除了一个虚假创新靶子：学界并未普遍认为“有人看过”就等于监督有效。

但该框架主要把干预可行性作为监督架构需要记录和评估的条件。其控制回路承认任务层与监督层相互作用，却尚未提出一个更强的动态命题：监督活动本身可能系统性地降低下一时刻的干预可行性，而且同一项“增加复核”操作会因依赖是否冻结而改变效应符号。本章研究的正是这一内生反馈。

Laux把监督理解为制度化不信任，强调监督角色不能只承担责任，还需要不同类型的质疑能力与民主治理位置\[3\]。这一思路说明监督不是个体注意力技巧，而是组织结构。本章进一步追问：制度化的不信任是否同时制度化了“等待监督完成期间，谁可以继续投入”？若该问题没有被写进流程，不信任可能在程序上增加、在行动上失效。

（二）可争议性与意义上的人类控制已经要求“能够改变”，但没有测量改变能力如何在争议期间衰减

意义上的人类控制以tracking与tracing要求系统响应人的相关理由，并能把结果追溯到理解系统且能够承担责任的主体\[4\]。可争议人工智能则把挑战、回应、重新决定、补救与生命周期治理纳入社会技术系统设计\[5\]\[6\]。Alfrink等人尤其指出，末端出现一个人并不足够；控制者应有权力和能力实际改变先前决定，争议机制也不应把负担推给个体\[5\]。

这些理论已经堵住了“有申诉入口便有控制”的简单说法。本章的剩余不是再发明一种权利，而是把权利的现实承载条件改成时间变量：在挑战被提出、审理和回应的过程中，改变当前处置所需的路径可能继续衰减。系统可以越来越会听、越来越会解释、越来越能记录责任，同时越来越不能改变这一轮结果。程序性可争议度与行动性可争议度因而可能背离。

（三）路径依赖已经解释选择空间收窄，但没有把保护性程序作为收窄的条件性加速器

组织路径依赖理论用协调效应、互补效应、学习效应和适应性预期解释行动走廊如何自我强化并进入锁定\[7\]\[8\]。Petermann、Schreyögg与Fürstenau的模拟研究进一步表明，层级权力可以延缓路径形成，却未必阻止自我强化过程最终压过正式权威\[9\]。这直接支持v8曾提出的“有权而无路”，也同时削弱了“因果截止点”作为独立理论的原创性。

新问题不再是锁定是否存在，而是监督如何改变锁定速度。若三轮复核只是多花时间，那么现象属于普通时延；若每一轮中间状态都被下游当作临时通行证，使专用依赖只围绕焦点方案累积，那么监督不是锁定之外的旁观者，而成为自我强化链的一部分。只有后者成立，本章才有独立增量。

（四）实物期权已经说明保留选择有价值，但可能把选择基础错误地当成监督之外的对象

实物期权理论把分阶段投资、等待信息和保留放弃权视为不确定条件下的灵活性来源\[10\]。然而Adner与Levinthal指出，组织过程会破坏实物期权所依赖的清晰阶段、退出标准和可放弃性；搜索灵活性本身甚至可能削弱放弃灵活性\[11\]。这与本章非常接近：等待更多信息并不自动保存未来选择。

但“监督消耗反事实基础”仍比“期权会到期”更窄也更危险。它要求观察：原本以保存判断质量为目的的复核，是否通过中间状态的组织传播，内生改变了被评估方案与替代方案的相对可执行性。若期权模型在不加入这一传播机制时即可完整预测数据，本章不应保留新概念。

（五）自动化偏误与人机组合研究说明“有人参与”不保证更好，但没有给出路径层的翻转机制

自动化研究长期发现脱离回路、接管困难、过度依赖与判断偏差\[12\]\[13\]。Green与Chen的行为实验表明，算法进入决策回路后，人对建议的使用可能产生意外且不公平的交互后果\[14\]；Buçinca等人发现认知强迫设计可以减少对AI建议的过度依赖\[15\]。Vaccaro、Almaatouq与Malone对106项实验、370个效应量的预注册元分析则显示，人机组合平均而言可能劣于人或AI中的最佳一方，且决策任务尤其容易出现损失\[16\]。

这些证据足以否定“人工参与必然改善结果”，却不能直接证明本章的路径机制。自动化偏误主要解释人为何错误相信或使用建议；本章预测，即使参与者准确识别错误、不信任AI、没有心理承诺，专用依赖的非对称累积仍可使异议无法落地。因此，认知变量必须作为竞争中介而非理论的默认解释。

（六）道德缓冲区已经诊断“低控制、高责任”，但没有解释归责证据与控制衰减如何被同一流程共同生产

Elish提出“道德缓冲区”，说明复杂自动化系统中的最近人类操作者可能像汽车缓冲区一样吸收系统性失败的责任，而拥有更大设计与组织控制的行动者退入背景\[20\]。这一理论已经覆盖本章命题五的一半：责任归因与实际控制可以不一致。因此，“批准者被多归责”本身不构成新发现。

本章增加的是一个发生机制和一个可分离预测。最终复核者之所以容易成为责任承载点，不只因为其靠近事故或具有正式角色，还因为复核流程主动生成了关于其知情、参与和签署的高密度证据；与此同时，中间复核状态允许其他节点继续添加专用依赖，使其边际控制下降。若在控制角色接近性、正式权限和结果知识后，复核轨迹密度不再影响第三方归责，责任增量应被道德缓冲区完全吸收。

（七）可追溯性与审计轨迹能够支持问责，但“记录了控制”不能直接证明“保存了控制”

Kroll把可追溯性理解为连接系统构造、运行机制、目的与治理判断的原则，使透明信息能够服务问责\[21\]。Power则说明审计轨迹不仅记录组织活动，还会塑造行动者持续生产可审计账户的倾向，形成审计制度自我扩张的微观基础\[22\]。欧盟《人工智能法》第12条的自动日志义务与第14条的人类监督要求，也使记录与控制在同一监管结构中相邻\[1\]。

这些研究说明轨迹有治理价值，不能被本章贬为纯粹表演。问题在于传统轨迹更擅长回答“谁在何时看到了什么、做了什么”，未必回答“那一时点仍有哪条替代路径能够承接其异议”。若日志只记录人的参与，而不记录替代路径何时失去证据、权限、资源与时间，它可以提高归责可见性而不提高因果控制的可见性。本章因而不反对可追溯性，而提出一种更严格的充分性条件：任何以复核轨迹支撑责任分配的制度，都必须同时报告当时的反事实承载力及其下降由哪些节点造成。

三、从备选方案到反事实承载力

（一）概念定义

对决策事件i、焦点方案x_i与时点t，令$\mathcal{P}_i(t)$为组织能够陈述的非焦点路径集合。一个可想象的办法不等于一个能承接异议的路径。对每条替代路径p，定义五项连续条件：

$q_{ip}(t)$：证据充分性；

$a_{ip}(t)$：是否存在有权接收并推进该路径的主体；

$r_{ip}(t)$：时间、预算、接口、人员与专业资源的可用程度；

$b_{ip}(t)$：在事前声明的权利、安全和成本边界内进行切换的可承担程度；

$d_{ip}(t)$：在后果不可接受地固化前完成切换的时间充分性。

单一路径的承载值取最弱条件：

$k_{ip}(t) = \min$

事件i在时点t的反事实承载力定义为最强非焦点路径的承载值：

$K_i(t) = \max_{p \in \mathcal{P}_i(t),\, p \neq x_i} k_{ip}(t)$

$K_i(t)$取0至1。它使用“最弱项—最强路径”结构，而不是把五项简单相加：一条替代路径若没有承接人，更多预算不能补回权限；若已经错过安全切换窗口，更完整的证据也不能追溯改变当前处置。多条替代路径中只需至少一条仍然足以承接异议，因此事件层取最大值。

二元研究可事前设定风险相称阈值κ_i：当$K_i(t) \geqslant \kappa_i$，至少一条替代路径被视为可执行；当$K_i(t) \<\kappa_i$，异议可能仍能触发解释、补偿或未来学习，但不能改变当前处置。阈值必须在观察结果前冻结，并报告敏感性区间。

（二）它不是什么

反事实承载力不是选择数量。十条无人承接的建议不如一条有证据、有预算、能按时执行的替代路径。它也不是一般可逆性：物理上可以撤回的行动，可能因合同、排期和权利边界而无法承受；已经发生部分投入的方案，也可能因备用资源被保护而仍可改变。

它不是监督者的主观能动感。一个人可以确信自己有权拒绝，却没有任何路径接收拒绝；也可以感到压力很大，但受保护的替代方案事实上随时可启动。主观控制感是需要测量的结果之一，不能代替$K_i(t)$。

它也不是路径依赖的同义词。路径依赖描述行动走廊如何收窄；$K_i(t)$是某一事件在某一时点还能否把合格异议转为现实路径的测量对象。若没有输出级事件、替代路径和干预时点，不能从“组织有历史”直接推出$K_i(t)$下降。

（三）因果截止点与批准错位退居派生量

v8把因果截止点作为主概念。本章将其降为$K_i(t)$的派生边界：

$\lambda_i = \sup$

令ρ_i为正式批准时点，则$G_i=\rho_i-\lambda_i$仍可用于描述批准发生时路径是否已经死亡。但λ_i和G_i不再承担理论原创性；它们只是把反事实承载力的时间变化转成可报告的事件量。若$K_i(t)$连续衰减且不存在稳定阈值，λ_i应改为区间或生存曲线，不得保留一个伪精确的“截止点”。

四、监督如何消耗自身的有效条件

（一）临时合法性接力

正式批准之前，组织常有“继续评估”“原则上无阻”“等待补件”“进入下一关”之类中间状态。它们在规范意义上并非批准，却可能被不同下游角色读取为行动信号：工程可以继续写接口，采购可以预留供应商，法务可以围绕现有文本改条款，运营可以占用发布窗口。每个角色都只做一项可撤回的小动作；这些动作彼此互补后，替代路径的切换成本却可能陡增。

本章把这种跨角色传播称为临时合法性接力：监督流程中的非终局状态，从“尚未否决”转化为“可以继续围绕焦点路径投入”的操作许可。它不是法律上的授权，也不要求监督者主观认可方案。它是一种状态语义的外溢。

若中间复核完全私密，下游看不到状态，接力应消失；若下游收到相同时间长度的普通进度通知，却没有“已通过一轮检查”的含义，接力也应显著减弱。这两个对照使机制区别于单纯时间流逝。

（二）专用依赖继续与替代资源暂停

临时合法性接力产生两种不对称。其一，焦点路径继续获得专用依赖：接口、测试、排期、预算、合同和认知参照围绕它增长。其二，替代路径通常被暂停，因为组织把“先完成当前复核”当作避免重复劳动的效率原则。焦点方案在等待中继续成为现实，替代方案在等待中停止成为可能。

这种不对称比一般沉没成本更精确。关键不是总投入增加，而是投入的专用性及其在焦点与替代路径之间的分布。若新增资源可被所有方案共享，或替代路径获得等量维护，$K_i(t)$不应显著下降。

（三）合理异议成本被私有化

当焦点路径依赖积累后，提出异议的人不仅要证明方案有问题，还要承担重建替代证据、协调返工、解释延期和面对同事损失的成本。原本属于系统的“保持可改”成本，被转移给最后提出异议的个人。形式上，异议权仍在；经济上，异议者必须独自购买一条已被组织停止维护的路径。

可争议AI研究已经警告申诉负担向个体转移\[5\]。本章将其前移到批准前工作流：即使没有受影响主体发起的事后申诉，内部复核者也可能成为拒绝成本的最终承受者。若组织建立公共的替代预算、反报复安排和明确接收人，合理异议成本不再私有化，控制反转应减弱。

（四）控制证据增厚与边际控制衰减

每一轮复核都会产生材料：谁收到通知、谁打开文件、谁提出问题、谁签署意见、谁批准继续。这些材料有正当用途，但它们的生成速度通常快于替代路径状态的记录速度。制度于是更容易看到“最后谁留下了名字”，不容易看到“更早谁允许替代路径失去资源”。当归责依赖可见轨迹，而轨迹只覆盖人类复核、不覆盖依赖增长时，责任会沿证据密度回流到最终复核者。

这一过程与异议成本私有化相互加强。替代路径越弱，最终复核者拒绝的现实成本越高；其签署继续的概率越高，留下的归责记录也越完整。复核轨迹不是控制衰减的唯一原因，但可能同时成为临时合法性的载体和事后归责的证据。由此，同一材料在事前推动路径长厚，在事后证明人类在场。

五、认识增益、控制证据与控制效力的三通道模型

（一）有效控制不是复核次数

令R表示复核强度，可由独立轮数、信息源数量或检查深度操作化；$E_i(R)$表示监督者识别相关错误并形成合格异议的概率；$T_i(R)$表示可归属于人类复核者的轨迹密度，包括通知、阅读、异议、签名与批准记录；A_i表示其调用干预的有效权限；$K_i(R,Z)$表示复核结束时的反事实承载力；Z表示复核期间的依赖治理，Z=1为冻结焦点方案新增专用依赖并维护至少一条替代路径，Z=0为依赖继续。

本章把一次监督改变当前处置的概率写为：

$$H_i(R,Z) = E_i(R) \times A_i \times K_i(R,Z)$$

乘法不是为了宣称三个变量彼此统计独立，而是表达一个必要条件结构：没有识别，就没有合格干预；没有权限，异议无法被调用；没有承载力，权限只能产生补救或解释。经验模型可以使用广义线性形式和交互项，不必机械套用乘法函数。

（二）控制反转条件

通常预期$\partial E/\partial R\>0$：更多复核提高识别概率。依赖冻结时，复核不应系统性消耗K，因而：

$$\partial H/\partial R \mid Z=1 \geqslant 0$$

依赖继续时，更多复核轮次可能通过临时合法性接力和更长的专用投入窗口降低K。由乘积求导：

$$\partial H/\partial R = A(K\cdot \partial E/\partial R + E\cdot \partial K/\partial R)$$

当反事实承载力的相对损失超过认识能力的相对增益时：

$$-(1/K)\cdot \partial K/\partial R \,\backslash> (1/E)\cdot \partial E/\partial R$$

即使监督者更会识错，整体控制仍随复核增加而下降。本章把下列联合状态称为控制反转：

$$\partial H/\partial R \,\backslash|\, Z=1 \,\backslash> 0 \text{ 且 } \partial H/\partial R \,\backslash|\, Z=0 \,\backslash< 0$$

新理论的判决点不是存在一个R的负效应，而是同一复核操作随依赖治理改变效应符号。若复核在两种条件下都改善控制，依赖继续并未消耗反事实基础；若在两种条件下都损害控制，结果更可能来自疲劳、信息过载或低质量复核。只有翻转才支持认识—路径双通道结构。

（三）控制证据与因果控制的分岔条件

在常规数字工作流中，复核轮次增加通常使T上升：

$$\partial T/\partial R \,\backslash> 0$$

令B_i表示第三方对最终复核者分配的责任强度。B不是本章对其应受谴责程度的规范判决，而是可实验测量的归责结果。若责任评估主要读取传统复核档案，T、E、正式角色和最终签名会提高B；若档案同时呈现K的时间曲线、依赖添加者与替代资源暂停者，B应更接近各节点对路径可改性的边际影响。

本章所要求的强分岔发生在依赖继续条件：

$$\partial E/\partial R \,\backslash> 0,\quad \partial T/\partial R \,\backslash> 0,\quad \partial K/\partial R \,\backslash< 0,\quad \partial H/\partial R \,\backslash< 0,\quad \partial B_{trad}/\partial R \,\backslash> 0$$

也就是说，识错更多、记录更多、归责更多与能改更少同时发生。依赖冻结后，T仍随复核增厚，但H的效应翻为正。这个“轨迹不翻、控制翻转”的差异是新增责任论证的判决点：若T与H总是同向，或加入K轨迹后责任分配不变，控制证据与因果控制的分离没有独立经验价值。

（四）一种允许失败的数值示意

假设一次复核时E=0.70，三次复核时E=0.88。依赖冻结时K保持0.80，则H从0.56增至0.704；依赖继续时，若K从0.70降至0.35，H反而从0.49降至0.308。该例只说明模型可以同时容纳“知道更多”与“改变更少”，不是经验结果。若真实数据中的K损失不足以越过上述不等式，控制反转不成立。

六、七项可区分命题

命题一：认识增益—路径损失并存

增加复核应提高错误识别率或合格异议率；在依赖继续条件下，同一操作同时降低复核结束时的K。只观察最终采纳率不足以支持本章，必须同时测得一正一负两条中介路径。

命题二：复核强度×依赖冻结的符号翻转

高强度复核在依赖冻结时提高合格异议改变当前处置的概率，在依赖继续时降低该概率。主要统计检验是交互项及两个条件内方向，而不是复核主效应。

命题三：临时合法性的可见性是必要中介

在总时长、复核内容和参与者负担相同时，公开中间复核状态比私密中间复核状态产生更多焦点方案专用依赖、更低K和更低的晚期切换率。若私密复核产生同等效应，临时合法性接力不是必要机制。

命题四：效应不以信任AI为必要条件

在控制AI信任、建议采纳倾向、识别准确率、主观承诺和重复暴露后，公开复核×依赖继续仍应通过客观专用依赖降低K。若加入这些认知变量后路径效应消失，理论应退回自动化偏误或承诺升级。

命题五：控制证据与因果控制反向移动

随着复核轮次增加，关于最终复核者知情、参与和签署的T上升；在依赖继续条件下，其改变当前路径的边际能力却下降。向第三方只展示传统复核档案时，对最终复核者的归责B应上升。若T与H始终同向，或归责不受轨迹密度影响，本章对责任错位没有增量。

命题六：能力档案应改变责任分配

在事件事实完全相同的条件下，看到传统复核档案的评估者应更多归责最终复核者；看到能力档案的评估者还会看到K何时下降、谁添加专用依赖、谁暂停替代资源，其责任分配应从“最后签名者”转向对路径衰减具有更大边际影响的组织节点。若两种档案不改变分配，增加K轨迹对问责没有价值。

命题七：AI标签不是必要原因，完成外形与接口覆盖才是

AI候选预期更容易触发控制反转，是因为生成速度、完成外形和跨接口覆盖使临时合法性更快转成专用依赖。将AI与人类候选在质量、完成度、接口覆盖和下游传播速度上匹配后，单纯AI标签效应应显著缩小。若标签仍有独立效应，需要另行解释信任、权威或厌恶；若人类候选产生同等效应，理论应扩展为一般组织监督，不能凭AI场景主张独占性。

七、最强近邻能否替换核心命题

最强近邻	已经解释的部分	若要完全替换本章，必须同时推出	分离实验
2026有效人类监督框架	监测与干预分离；权限、资源、时间、成本决定干预可行性	监督轮次本身通过下游状态传播改变干预可行性，并在冻结条件下发生符号翻转	固定人员、权限和信息，随机化中间状态可见性与依赖冻结

欧盟AI法第14条	人应能理解、推翻、逆转或停止AI输出	合规复核越完整，当前处置反而越难逆转；停止系统不等于存在替代路径	保持法定权限与停止按钮不变，测量K与实际转向
可争议AI	系统应开放、响应挑战并支持重新决定或补救	程序性可争议度提高与行动性可争议度下降并存	分开记录解释、补偿、未来学习与当前路径改变
意义上的人类控制	系统应追踪人的理由，责任应可追溯	对理由的识别改善与理由改变当前结果的能力下降并存	同时测E、A、K与当前处置改变
组织路径依赖	自我强化使行动走廊收窄，层级未必阻止锁定	保护性复核作为条件性加速器，并由状态可见性触发	私密复核、公开复核、普通时间通知三组对照
实物期权	保留选择具有价值，组织过程会破坏放弃灵活性	用来等待信息的监督程序内生消耗被等待的选择基础	固定等待时间与信息量，只改变中间复核是否许可专用投入
控制时延／信息过载	延迟、疲劳和复杂性降低干预质量	同样延迟下，冻结使复核效应转正，公开状态使效应转负	时间匹配、任务量匹配、私密复核控制
自动化偏误／承诺升级	人过度依赖建议，或因投入而继续原方案	不信任AI且没有个人承诺时，客观依赖仍使异议失效	测量认知中介；由其他角色自动产生依赖
道德缓冲区	低控制的人类操作者可能吸收分布式系统的责任	监督流程共同生产对最终复核者的归责证据与其控制能力的衰减	固定角色接近性和权限，随机化轨迹密度与能力档案
可追溯性	连接系统构造、运行、目的与决策过程，以支持问责	传统参与轨迹增厚可与反事实承载力下降并存；能力轨迹改变责任分配	同一事件随机展示传统档案或含K曲线的能力档案
审计轨迹	可审计账户具有自我复制和塑造组织行动的能力	同一复核轨迹事前充当投入许可、事后充当归责证据	追踪状态传播、依赖附着和事后档案使用的时间顺序

表 1　最强近邻、必须同时推出的内容与分离实验

两种压缩下的不可还原性检查

机制压缩：人工监督的效力取决于监督期间是否仍保存把合格异议变成另一条现实路径的能力；复核记录只能证明人曾在场，不能单独证明其当时仍有足够控制。

前半句仍可被“干预必须可行”大幅吸收，后半句也可被道德缓冲区与可追溯性批评分别吸收。把两个近邻并列不产生新理论，单凭机制压缩仍不足以主张高不可还原性。

并存预测压缩：依赖继续时，复核同时提高识错率、复核轨迹和对最终复核者的归责，却降低替代路径与实际控制；冻结依赖后，实际控制由负转正，而复核轨迹仍继续增厚；补入K时间曲线后，责任分配从最后签名者转向造成路径衰减的节点。

后一压缩留下两层关系剩余。第一层是：复核不是只受锁定影响，而会通过中间状态的可见传播改变锁定速度，并在冻结新增专用依赖后发生符号翻转。第二层是：同一过程还生产对最终复核者的高密度参与证据，使归责可见性与边际控制反向移动；将K时间曲线加入档案后，责任归属应重新分

布。路径依赖、监督框架、实物期权、道德缓冲区与可追溯性各自覆盖其中一部分，但任何替代都必须同时推出“轨迹不翻、控制翻转、归责再分配”这一联合预测。若敌意检索或数据发现该同构结构已被提出，责任层增量必须删除。

八、证据现在支持到哪一层

本章区分三层证据，避免把研究设计提前写成发现。

第一层是条件证据。欧盟AI法要求监督者能够推翻、逆转或停止输出\[1\]；有效监督框架把干预可行性、时间和成本列为核心条件\[2\]；可争议AI要求控制者有能力实际改变先前决定\[5\]。这说明“看见”与“能改”不可合并，但不证明监督会消耗能改的条件。

第二层是机制邻接证据。组织路径依赖研究说明互补与协调会收窄行动空间\[7\]\[8\]，层级权力可能只延缓而不能阻止锁定\[9\]；实物期权批评说明组织搜索过程会破坏退出灵活性\[11\]。这使“复核期间的路径衰减”具有既有机制基础，但仍不证明临时合法性接力。

第三层是结果异质性证据。人机组合元分析显示，人工参与并不保证优于最佳单方，决策任务尤其可能出现损失\[16\]；行为实验也表明算法建议进入人的决策过程后会产生复杂交互\[14\]\[15\]。这为拒绝单调性先验提供独立证据，却不能识别本章提出的路径中介。

第四层是责任邻接证据。道德缓冲区说明有限控制的人类可能吸收分布式系统责任\[20\]；可追溯性研究说明参与和决策轨迹可以支持问责\[21\]；审计轨迹研究说明记录实践能够塑造组织行为并自我扩张\[22\]。三者共同使“记录更多、归责更多、控制更少”成为有机制邻接的候选，但没有直接证明复核轨迹在事前促进依赖、事后支撑归责，更没有证明加入K时间曲线会改变责任分配。

因此，本章当前仍是理论与测量候选。独立文献两次改变并约束了理论：它先迫使本章放弃“有权无路”作为原创主张，把机制赌注缩到监督内生改变干预可行性；又迫使本章承认“低控制、高责任”已被命名，把责任赌注缩到复核轨迹、路径衰减与归责重分配的联合生成。两项核心交互均无本章所需的直接数据，不得写成已验证组织规律。

九、预注册实验：让新理论与近邻同场竞争

（一）主实验：2×2设计

在模拟软件发布与采购决策中，随机分配：

复核强度R：一次终局复核／三次阶段复核；

依赖治理Z：依赖继续／依赖冻结。

所有组接收同一焦点方案、相同初始证据、相同正式权限和相同总时长。低复核组用与任务无关但负担匹配的活动填充时间，排除单纯等待差异。高复核组每轮形成“继续评估”状态；依赖继续条件把该状态公开给下游模拟角色，并允许产生候选专用接口、排期和资源预留；冻结条件禁止任何方案专用依赖，且维护一条最低合格替代路径。

在终局批准前注入同一条经预试确认的合格反证。主要结果是参与者能否把当前处置切换到合格替代路径；不是是否点击反对。共同主要中介是E（是否识别反证）与K（反证到来时替代路径的承载值）。

支持性模式必须同时满足：

高复核相对低复核提高E；

在冻结条件下，高复核提高H；

在依赖继续条件下，高复核降低H；

R×Z交互方向与预注册一致；

K的变化中介H的翻转，而主观AI信任不能完全吸收该中介。

若只有交互显著但两个简单效应不发生方向翻转，只能称为强度调节，不能称为控制反转。

（二）机制拆解：公开、私密与普通通知

第二项实验保持复核内容、轮次、时长和下游资源完全相同，只改变中间状态的传播：

公开复核：下游看到“已通过本轮检查，可继续评估”；

私密复核：复核结果只对监督者可见；

普通通知：下游只看到“流程仍在进行”，不含检查通过语义。

若临时合法性接力成立，公开复核应产生更多专用依赖和更低K；私密复核与普通通知的差异应较小。若三组相同，效应来自时间或重复暴露，而不是状态语义外溢。

（三）认知近邻控制

每轮测量AI信任、建议采纳倾向、风险识别、个人承诺、沉没成本感、疲劳和主观控制感。依赖由模拟的其他部门自动附着，参与者本人不作早期资源投入，以降低自我辩护解释。若控制这些变量后R×Z通过K的间接效应消失，本章接受自动化偏误或承诺升级解释。

（四）AI与人类候选的匹配复制

第三项实验加入候选来源：AI／人类专家。两类材料在字数、质量、完成度、接口覆盖、置信表达与格式上盲化匹配。若控制反转只随完成度和接口覆盖变化，不随来源标签变化，论文把AI定位为高频发生条件而非本体边界；若AI标签仍有剩余效应，再检验算法权威、厌恶与责任转移。

（五）两阶段责任档案实验

主实验的每个事件在结果揭示后生成两种事实等价、信息结构不同的档案。传统复核档案包含复核轮次、通知、异议文本、正式权限、签名与最终批准；能力档案在其上增加K五项条件的时间曲

线、每项专用依赖的添加者、替代资源的暂停者，以及各节点采取相反行动时对H的预注册边际变化。两种档案不得改变事故后果、任何行动者的知识、角色或实际行为。

由不知道研究假设的第二批评估者分配描述性因果责任、可避免性责任、组织改进责任与惩罚性责任。主要检验不是“能力档案让所有人少担责”，而是责任分布是否改变：传统档案应更集中于最终复核者；能力档案应把因果与改进责任部分移向允许专用依赖继续、暂停替代资源或定义中间状态传播的节点。惩罚性责任涉及规范判断，单独报告，不由边际控制机械决定。

为排除“信息越多越分散”的平凡解释，设置等字数安慰剂档案，增加与因果路径无关的技术细节；并设置只增加一般组织结构图、不增加K时间信息的对照。只有K轨迹与节点贡献信息产生定向重分配，才支持能力档案主张。若三种档案效果相同，或重分配仅由篇幅与复杂度造成，命题六删除。

（六）统计与判决门

主要模型使用二项回归预测当前处置是否被合格异议改变，含R、Z、R×Z、场域及预注册协变量。E与K作为平行中介，报告组间差异和不确定区间。预测比较采用嵌套模型与样本外Brier分数\[17\]\[18\]：

M0：任务难度、方案质量、人员经验、AI来源；

M1：复核时长、信息量、疲劳、信任、自动化偏误、承诺、一般依赖量；

M2：在M1上加入K、临时状态可见性、专用依赖不对称与R×Z。

M2相对M1的样本外Brier改善若不足0.01、95%区间跨0、留一场域外推方向不一致，或K中介不稳定，不能主张独立预测价值。

责任档案实验采用组成数据模型或多项模型分析责任份额，并预注册最终复核者份额与路径衰减节点份额的方向。能力档案相对传统档案必须同时降低“最后签名者”的因果责任份额、提高至少一个经事件账本确认的路径衰减节点份额；若只有总体责任下降而没有定向转移，不支持本章的归责机制。档案类型对责任分布的样本外预测改善也须超过安慰剂档案，并在软件发布与采购两个场域方向一致。

（七）功效分析与样本承诺

本章不从Vaccaro等人的人机组合元分析直接借用交互效应量。该研究发现人机组合相对最佳单方的总体效应为负，且不同任务与设计之间存在高度异质性\[16\]；这些结果支持“人类参与并非单调有益”的背景判断，却没有估计本章所需的“复核强度×依赖冻结”二项交互。把其Hedges' g转换为本章的交互参数，会把不同结果、不同对照和不同机制强行同一化。

在缺少直接先导数据时，样本量改由事前最小有意义差异与蒙特卡洛模拟共同确定\[19\]。主要检验对象是饱和二项Logit模型中的R×Z交互项，双侧α=0.05。基准情景把低复核条件下的路径改变率设为0.50；把具有理论意义的最小翻转定义为：高复核在依赖继续时把该概率降至0.40，在依赖冻

结时把它升至0.60。也就是说，两个简单效应分别为−10与+10个百分点，概率尺度上的差中之差为20个百分点。

采用固定随机种子20260809、50,000次重复和独立二项单元计数，结果如下：

情景	两个简单效应	每格可分析样本	总可分析样本	交互检验功效	方向翻转率
大翻转	±12个百分点	140	560	0.811	0.952
最小有意义翻转	±10个百分点	200	800	0.811	0.951
小翻转	±8个百分点	320	1280	0.816	0.954
基线率0.35，最小翻转	±10个百分点	200	800	0.850	0.962
基线率0.65，最小翻转	±10个百分点	200	800	0.848	0.962

表 2　R×Z 交互项的功效模拟（固定随机种子 20260809，50,000 次重复）

零效应校准在每格200人时的 I 类错误率为0.050，说明模拟判决与预设α一致。基准方案因此要求每格200个可分析样本，共800人；按10%损耗率，最低招募量为892人，实施时取整为900人。若预试发现基线率不在0.35—0.65之间、操纵未产生预设强度、或有效损耗超过10%，不得在看过主结果后临时改写效应门槛；应在揭盲前重跑模拟、冻结新版方案并说明变更。

该样本承诺只保证对交互项约80%的检验功效。理论判决还要求两个简单效应的估计方向分别为负与正，但不要求它们各自在同一研究中独立达到p\<0.05；在基准样本下，两者同时显著的模拟概率只有0.269。若研究者把“两条简单效应均显著”升级为共同主要终点，则需每格约520个可分析样本、总计2080人；按10%损耗率招募2312人，此时两者同时显著的模拟概率约为0.811。两种成功标准必须在预注册时二选一，不得在结果出来后切换。

责任档案实验另行招募1,200名盲评者，每种档案400人。主要二元结果为“最终复核者是否获得最大因果责任份额”；以传统档案组比例0.50、能力档案组下降10个百分点作为最小有意义差异，双侧α=0.05时两组比较约有80%功效。安慰剂组用于检验额外篇幅与复杂度，不与主实验样本互换。若预试显示基线远离0.50，须在揭示组间结果前重算并冻结，不得把连续责任份额与二元最大份额事后互换为主要终点。

十、从实验到组织影子账本

（一）最小事件结构

组织研究不能只读取审批表。每项焦点产出至少记录：

事件	最小字段	用途
candidate_created		确定焦点方案起点

	candidate_id、来源、时间、内容哈希、完成度、接口覆盖	
review_state_changed	review_id、轮次、状态语义、可见对象、时间	识别临时合法性传播
control_trace_created	通知、阅读、异议、签名、批准、可见对象、时间	测量T及传统归责证据
dependency_attached	对象、专用性、创建者、成本、可撤回性、candidate_id	测量焦点依赖增长
alternative_maintained	path_id、证据、承接人、预算、接口、截止时间	计算K的五项条件
qualified_challenge	理由、证据等级、目标、时点	形成统一干预探针
challenge_outcome	当前路径改变／补救／未来学习／无行动	区分行动性与程序性回应
cost_allocated	返工、延迟、机会成本、承担主体	测量异议成本私有化
dependency_permissioned	授权节点、适用对象、依据、时间、可撤回条件	识别谁允许复核期间继续长厚路径
formal_approval	批准者、条件、时间、可撤回窗口	与派生截止边界比较
effect_committed	现实后果、可逆边界、时间	校验当前处置

表 3　组织影子账本的最小事件结构

（二）编码规则

q、a、r、b、d分别按0、0.5、1三级编码，并保留原始证据引用。0表示不满足最低条件；0.5表示条件存在但有未决障碍；1表示在事前边界内可直接调用。K取各路径最小值后再取路径间最大值。研究同时报告连续K与按κ二分的结果，避免阈值选择决定结论。

“专用依赖”必须满足反事实删除测试：若移除焦点方案，该投入不能无显著返工地转用于主要替代路径。共享培训、通用基础设施和与方案无关的合规检查不计入。中间复核状态的可见性由实际接收日志判断，不能因制度文件写着“仅供评审”就假定下游没有读取。

两名编码者先独立编码不少于30项事件。K五项条件的逐项一致率须达到80%，Cohen's κ须达到0.70；任一关键字段未达标，先修订手册并重新编码，不进入理论检验。若K只能在知道最终结果后通过主观故事恢复，构念不可识别。

（三）两周影子运行

选择一个低至中风险、已有正式批准机制的流程，影子记录不少于100项连续产出，不改变真实放行，也不进入个人绩效。第一周验证编码；第二周记录复核轮次、状态传播、依赖增长、替代维护和挑战效力。最低经验产物不是“发现控制反转”，而是四条可检查曲线：E随复核变化、K随复核变化、H随复核和冻结条件变化、责任归因与边际控制变化。

十一、治理含义：让复核暂停成为真的暂停

若控制反转得到支持，治理重点不能只增加审批人或检查表，而要改变复核状态的组织含义。

第一，建立依赖冻结窗。焦点方案进入高风险复核后，不允许新增不可共享的接口、合同、预算和发布承诺；确需继续的，必须记录例外理由和由谁承担未来切换成本。

第二，把“继续评估”与“允许投入”拆成两个状态。前者是认识进程，后者是资源授权；不得由同一枚状态自动触发。中间复核结果默认私密，只有明确的可撤回资源授权才向下游传播。

第三，维护最低可行替代路径。高风险事件至少为一条非焦点方案保留证据、接收者、基础资源和时间窗。它不是要求组织永久平行建设，而是让正式拒绝在关键窗口内仍有承接对象。

第四，把合理异议成本公共化。若异议符合预先证据门槛，由组织而非异议者个人承担必要的验证、返工和协调成本；否则所谓反报复条款只能保护发言，不能保护异议的现实效力。

第五，审计“复核产生了什么”，而不只审计“复核发现了什么”。每轮复核后同时报告新增信息与新增专用依赖。一个复核流程若识别率上升、K持续下降，应被视为双通道冲突，而不是用更厚的会议记录证明治理完善。

第六，把控制轨迹升级为能力轨迹。签名、通知与异议记录继续保留，但不得单独用于证明“人类保持控制”；同一档案还要显示当时至少一条替代路径的K、K下降的时间、添加专用依赖与暂停替代资源的节点。责任评估应读取“谁留下了记录”与“谁改变了可行动空间”两条时间线，禁止用最终签名替代边际控制。

这些建议不是普遍最优规则。紧急救援、不可并行维护的基础设施或极低风险高频任务，冻结成本可能高于收益。组织应按权利影响、可逆性、替代维护成本和时间敏感性事前设定适用范围。

十二、边界、反噬与理论死亡条件

（一）适用边界

本章适用于生成式AI产出作为候选协调物进入多角色工作流，并且复核期间仍可能发生下游资源配置的场景。它不覆盖纯个人、即时、低后果的内容生成；不覆盖不存在任何合格替代路径的紧急处置；不把安全上必须迅速封闭的路径写成治理失败；也不主张所有决策都应维持多个方案。

（二）最可能的反噬

依赖冻结可能造成信息贫乏：某些风险只有在真实集成中才会显露。替代路径维护也会消耗稀缺资源，拖慢所有方案，甚至让没有责任的团队承担“保持可改”的成本。中间状态完全私密可能阻碍必要协作。因而，本章不能把K最大化当作单一目标；治理需要在学习价值、处置速度、权利风险和路径可改性之间明确权衡。

另一个反噬是组织策略化。若团队知道K成为绩效指标，可能制造名义替代方案、虚假承接人或不可用预算，使账本看似健康。反事实删除测试、真实挑战探针和随机抽查必须优先于字段完整率。

（三）十条降级与删除条件

若增加复核在依赖冻结与依赖继续条件下不出现预注册方向的符号翻转，删除“控制反转”，改为一般调节效应。

若公开、私密与普通通知的差异不显著，删除“临时合法性接力”，转向时间、疲劳或重复暴露解释。

若信任、自动化偏误、个人承诺或沉没成本完全吸收路径效应，理论降级为认知偏误／承诺升级应用。

若一般依赖量已经完整预测结果，而专用性、不对称性和状态传播没有增量，理论降级为组织路径依赖应用。

若K不能达到80%一致率与$\kappa \geqslant 0.70$，或只能事后凭结果恢复，反事实承载力降级为启发式检查表。

若M2相对M1的样本外Brier改善不足0.01、区间跨0或跨场域方向不稳，停止独立预测主张。

若依赖冻结只增加延期与成本，未改善合格异议落地、错误放行或权利结果，撤回主要治理建议。

若敌意检索发现已有理论已经同构提出并验证“复核识别增益与路径损失并存，且随依赖冻结发生符号翻转”，取消新命名并归入该理论。

若在控制正式角色、知识与行为后，复核轨迹密度T不提高对最终复核者的归责，删除“控制证据增厚推动责任回流”的主张，责任错位归入道德缓冲区。

若能力档案与传统档案不产生定向责任重分配，或其效果不超过等字数安慰剂档案，删除能力轨迹治理建议，不把K测量扩张为问责基础设施。

（四）本章自己的未完成状态

本章要求监督保存改变结果的条件，但本章尚未运行主实验，也没有权把研究设计追认成经验结论。它目前完成的是：确定新构念、给出内生机制、组织最强近邻、提出区分性预测、提供编码入口并预先写明概念死亡后的处置。它没有完成的是：跨编码者重放、预注册数据、跨场域复制、责任档案实验和独立盲评。

因此，本章自身仍是一条受保护的候选路径，而不是已经批准的理论。后续证据有权删除“临时合法性接力”“控制反转”乃至“反事实承载力”这些命名；若删除后只剩“有效监督需要可行干预”，既有框架已经足够，本章应停止传播额外术语。

十三、结论

生成式AI监督的关键困难，不只是机器会不会犯错、人能不能发现错误，也不是审批是否发生在执行之前。更隐蔽的风险是：组织在等待监督完成时，可能继续围绕被监督的方案建设现实，同时围绕监督者建设责任证据。于是，复核每增加一轮，异议更精确、记录更完整、焦点路径也更厚；最后的批准者留下更多理由与签名，却只剩更少能够承接拒绝的路径。

反事实承载力把这一问题变成可测对象：合格异议到来时，是否仍有至少一条具备证据、承接权、资源、可承担成本与足够时间的替代路径。控制反转给出第一个不受作者保护的判决：同样增加复核，若依赖冻结，控制应提高；若依赖继续，控制可能下降。能力档案给出第二个判决：当K的时间曲线与路径衰减节点进入责任材料，责任分配应离开最后签名者，转向对可行动空间具有更大边际影响的节点。没有符号翻转，机制理论降级；没有定向重分配，责任理论删除。

这两项判断若被跨场域证据支持，人工监督的基本单位将从“一个人、一个按钮、一次复核”改为三条同时审计的时间线：人知道了什么，制度留下了什么控制证据，现实还保存了什么替代路径。治理也必须从要求人能够说“不”，推进到确保“不”说出口时仍有一条路接住它，并确保事后责任不会只沿最清楚的签名回流。

附论一 · 最近邻补切（编辑增补）

原稿第二节已经系统处理了七类近邻，并在第七节主动做了两轮不可还原性压缩。以下四家仍未出场，其中前两家与本章的机制最近。补切不改变作者的核心判断。

一、Perrow 的正常事故与紧耦合

已说到哪一步。正常事故理论把系统沿两轴分类：交互复杂性与耦合紧密度。紧耦合的定义几乎就是本章的 K 趋近于零——过程一旦启动就不能中止，替代路径必须事先设计好，缓冲、冗余与替换品无法临时凑出，恢复只能靠预置而不能靠临场发明。Perrow 还给出一条与本章方向一致的判断：给紧耦合系统追加监督层，可能同时增加交互复杂性，从而使干预更难而不是更易。

分离线。正常事故理论把耦合当作系统的结构属性——由技术与工艺决定，在事故分析中作为给定条件；本章把它改写成事件层的时间变量：同一个组织、同一套技术，K 在一次复核过程中逐轮下降，且下降速度取决于中间状态是否向下游传播。这是一个可被随机化的操作（公开／私密／普通通知三组对照），而耦合紧密度在 Perrow 那里不可随机化。

判决性对照。在耦合紧密度与交互复杂性匹配的两组事件之间，若依赖冻结仍使复核强度的效应符号翻转，本章获得独立内容；若在控制耦合紧密度后翻转消失，本章应并入正常事故理论，「反事实承载力」降为紧耦合的一个事件级读数。

这位祖宗反过来削弱本章的一条。Perrow 的结论比本章悲观：在紧耦合且高交互复杂性的系统里，他认为增加组织手段（包括本章第十一节提出的冻结窗、状态拆分、替代路径维护）本身会成为新的复杂性来源，从而制造新的失效模式。本章第十二节已经写到「依赖冻结可能造成信息贫乏」，

但没有写到更硬的一条：冻结窗、例外审批、承接人指派与能力档案这四项治理动作，各自都是新的状态语义，因而各自都可能成为下一轮临时合法性接力的载体。本章的治理建议应当接受这一自反检验。

二、Vaughan 的越轨正常化

已说到哪一步。对挑战者号发射决策的经典研究给出了一条与本章高度同构的机制：异常信号被逐轮纳入既有的可接受风险范围，每一轮都留下完整、合规、可审计的工程评审记录；正是这些记录使继续推进成为合规行为。事故不是规则被违反的结果，而是规则被逐步执行的结果。

分离线。越轨正常化解释的是判据本身漂移——什么算异常的标准被一轮轮放宽，因而识别能力下降；本章明确要求相反的形态：识错率上升（∂E/∂R\>0）而控制下降，即监督者越来越会发现问题，异议越来越精确，失效发生在路径侧而不是判据侧。这条区分是可判决的：越轨正常化预测异议数量与质量随轮次下降，本章预测二者上升。

判决性对照。逐轮记录合格异议的数量与证据等级。若随复核轮次上升而异议质量下降，越轨正常化足以解释，本章的双通道结构不成立；若异议质量上升而 K 下降，本章获得支持。这一条应当写进原稿第九节主实验的共同主要中介——现有设计只测 E 是否识别反证，未测 E 随轮次的轨迹，因而无法把自己与这位最近的祖宗分开。

三、Feldman & Pentland 的常规二重性（ostensive／performative）

组织常规有两个面：规范面（这条流程「应该」是什么，写在制度文本里）与执行面（人们实际做了什么）。本章的「临时合法性接力」在结构上正是这一区分的一个特例：「继续评估」在规范面上不是批准，在执行面上却是资源授权。分离线：常规二重性是一般性描述，不预测哪一面会赢；本章给出了赢的条件——中间状态是否向下游可见，并据此提出可随机化的对照。建议：原稿第十一节第二条（把「继续评估」与「允许投入」拆成两个状态）如果引入这一对术语，可以直接说明为什么单靠制度文本声明「仅供评审」无效——原稿第十节编码规则里那句「不能因制度文件写着『仅供评审』就假定下游没有读取」，正是执行面优先于规范面的一次应用。

四、Turner 的灾难孵化期

灾难在正常运行期间孵化：警告信号存在，却被既有的信息处理安排逐一吸收、分派与归档，直到触发事件到来。分离线：孵化期理论中，信号被吸收意味着没有被看见；本章的靶形态是信号被完整看见、被记录、被签署，失效只发生在可行动空间。两者可在同一批档案上分开：查异议是否进入正式记录即可。

附论二 · 站内划界（编辑增补）

本章与作者已发的三篇同线论文距离很近，原稿未作互引，以下分工由编辑部补写。

与《错误没有主人：签名、证据与追责为何不能相互替代》

这是本章最近的站内近邻，两文共享同一个反直觉起点：组织可以同时拥有更完整的记录与更弱的实际保障。分离线在失效落点：那篇的三本账（授权／证据／追责）之间不可代偿，失效发生在登记侧——一项关键主张从未被任何人以与风险相称的方式验证，而授权闭合被当成了证据闭合；本章的失效发生在行动侧——主张已被验证、错误已被识别、异议已经成立，可是没有一条替代路径能接住它。两者可以一真一假：一个组织的证据账可以非常干净（每一项关键主张都有独立核验），而在第三轮复核结束时 K 已降到零；反过来，一个组织可以保有充裕的替代路径，却从未验证过任何一项关键主张。

更精确的一处：那篇指出「证据账会退化为第四张签名表」，本章指出同一份复核轨迹「事前充当投入许可、事后充当归责证据」。前者是账本之间的功能替代，后者是同一份材料在时间上的两种用途。合起来给出一条两文都没有单独说出的预测：当证据账退化为签名表时，它同时也在加速 K 的下降，因为签名式材料的生成速度快于替代路径状态的记录速度。

与《最后一次点击之前：人机决策中的选项承接、路径效力与责任分配》

那篇的核心命名是「选项承接」，本章 K 的五项条件之一（a：是否存在有权接收并推进该路径的主体）正是承接。分离线在被承接物与时点：那篇问一个被提出的选项是否被决策链承接，时点在决定之前；本章问一条替代路径在合格异议到来时是否仍具备证据、权限、资源、可承担成本与剩余时间，时点在复核进行之中，且强调这五项会在等待期间同时衰减。那篇的因变量是选项是否进入下一环节，本章的因变量是当前处置是否被转向。

与《当过去不再代表现在：生成式记忆系统中的身份修订、效力复发与履行审计》

那篇审计的是已被撤回的内容是否仍在充当输出的理由；本章审计的是已被识别的异议是否还有落地的通道。前者测撤回后的残余，后者测异议到来时的承载；一个是效力没有按时消失，一个是效力无处安放。两者的时间方向相反，可在同一批日志上分别读出。

与本站阳涌《耦合可逆性极化》（同日上站）

两文都在问「事到临头还有没有第二条路」，但扰动来源相反：那篇的扰动来自外部条件改变（模型被撤除、替换、出错或任务结构迁移），因变量是恢复；本章的扰动来自内部的合格异议，因变量是转向，且路径的衰减恰恰由监督程序自身在等待期间制造。形式上也有一处可比：那篇的四维可逆性剖面拒绝把不同失效机制压成一个数，本章的 K 则先在五项条件内取最弱、再在多条路径间取最强——两种聚合规则可以互相借用，是两文之间最实的一处接口。

附论三 · 编辑校订说明

本站在不改动作者判断、论证结构与任何数字的前提下，做了以下技术性处理：

一、表格归位。原稿的三张表在文档中脱离了正文位置，本站按其内容分别还原到第七节（最强近邻与分离实验，十一行）、第九节第七小节（功效模拟结果，五行）与第十节第一小节（最小事件结构字段表，十一行），单元格内容逐字未改。

二、形式记号。正文中的定义式、控制反转条件与偏导表达式按原稿逐字保留，仅改为独立居中排版以便阅读；下标字符（如 k_{ip}、K_i、E_i、T_i）沿用原稿写法。

三、未作任何内容改动。本章改姓扫描为零（无本站方法论术语泄漏），文献 22 条编号与正文标记一一对应，未发现杜撰的卷期或页码；本站未逐条复核 DOI。

四、一处建议已写入附论一第二条：原稿第九节主实验的共同主要中介只测「是否识别反证」，未测识错质量随复核轮次的轨迹，因而在设计上还不能把自己与越轨正常化分开。这是一处可在预注册阶段低成本补齐的缺口。

五、定位提示。本章自述版本 v9.2，属「理论建构、形式模型与双阶段预注册实验方案」，主实验尚未运行；作者已在第十二节第四小节声明「本章自身仍是一条受保护的候选路径，而不是已经批准的理论」，并写明后续证据有权删除其三项命名。请读者按此定位阅读。

本分册取自《越顺利，越没有你的位置》（德麦国际专著第 84 号）。全书 343 页，含前言、导读、导论、四编十一章、枢纽章、合章、结语、参考书目与三则附录；本章的判据与划界须与枢纽章、合章一并读。全书网页版：sdeuniverses.com/books/m/84/text/