学科通融 · 之二十四

谁替你按住那颗按钮

论一次批准算不算控制，不看批准者懂多少，而看他不按的时候代价先落在谁身上

人机协作治理 × 核武器指挥控制 × 支付结算法

作者 王德生 + Claude

栏目 学科通融 · 站外碰撞

篇幅

发表 2026年8月1日

摘　要

一份批准之所以算数，通常被归给两件事之一：批准者理解了决定结局的那一步，或者动手的权限被分割给了互不知情的多人。本文论证两者都不够，而且它们漏掉的是同一样东西。**为一个人保留干预窗口，等于把这段时间里的不可逆转移给了别人；转移本身不是病，不记账才是。**　本文把这笔被转移而未被记账的不可逆称为逆债，并给出一个不含任何情态词的问法：**如果这个人从不使用他的干预权，代价先落在谁身上？**　这个问法在急诊、福利审核、代码回滚、跨境支付与核发射五个场景里给出互不相同且可核对的答案，并解释了一件长期没有解释的事：为什么同样的「人在回路」在软件工程里管用、在行政审批里几乎必然退化成盖章。本文给出二维辨别表，指认最危险的一格是「窗口在、代价不在」——它能通过全部形式审查。两个领域反过来迫使命题收窄：核指挥体系说明有些逆债是故意欠下的，支付结算法说明「没有窗口、也没有人承担」在很多系统里是健康常态而非溃缩，因而本文不含「越可逆越好」这一价值排序。

这一篇由哪三个领域的争论撞成

三边对「一个不可撤回的后果落定之前，控制应当以什么形式存在」给出互相排斥的答案：人机协作治理说以理解的形式，且窗口越晚关越好；核指挥体系说以分割的形式，而且必须防止任何一个位置掌握全局；支付结算法说根本不该有窗口，可撤销性本身就是风险来源。**本篇的两处收窄由后两边逼出：核指挥说明有些转移是故意欠下并另有承接的，结算法说明「没有窗口、也没有人承担」在很多系统里是健康常态——因而本文不含「越可逆越好」这一价值排序。**

人机协作治理 · 保住人的理解与干预窗口

Santoni de Sio & van den Hoven, Meaningful Human Control over Autonomous Systems（Frontiers in Robotics and AI, 2018）；及其近年的两条推进：最低理解阈值（arXiv:2602.00854，预印本）与责任的溃缩（ESTS, 2019）

要求系统对人的理由保持响应、结果可追溯到有能力理解的人；近年的推进指出长期依赖会造成「辅助下表现上升而内部模型退化」，因而主张在无实时辅助的条件下检验能否恢复关键推理。共同的处方是：让人懂得更多、说明来得更早。

https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2018.00015/full

核武器指挥控制 · 用分割替代理解

Feaver, Guarding the Guardians: Civilian Control of Nuclear Weapons in the United States（Cornell University Press, 1992）

「需要时必须能用，不需要时必须绝对不能用」这对反向要求，催生了必须由两人各自独立完成的动作结构与把解锁编码放在执行现场之外的硬件设计。安全不建立在执行者懂得更多之上，而建立在没有任何一个位置拥有全局之上；该体系明确知道这会拖慢反应，并在别处以预先授权与继任安排承接这份代价。

https://www.cornellpress.cornell.edu/book/9780801480645/guarding-the-guardians/

支付结算法 · 用立法消灭可撤销性

Directive 98/26/EC on settlement finality（欧盟，1998）；及其国际标准配套 CPMI-IOSCO, Principles for Financial Market Infrastructures（BIS, 2012）原则八

一九七四年那次银行关闭暴露了「一腿已付、另一腿未付」的跨时区敞口。此后的制度方向是把不可撤回的时点尽量提前并用法律钉死：结算指令一旦进入系统即不得撤销，即使参与者随后破产、即使破产程序具有溯及效力。**终局性是一条立法，不是一个物理事实——这正是本文推翻三家共有前提所用的材料。**

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31998L0026

目　　录

一、导论：三个不该同时为真的说法

三场争论，分属人机协作治理、核武器指挥控制与支付结算法。三场都有厚实的支撑，三场都在回答同一个问题，而它们的答案不能同时成立。

第一场在人机协作治理。 自动系统能够给出完整的诊断、判决草稿、代码与投放方案之后，各国的合规框架几乎一致地退到同一条底线上：保留一个人，让他在结果生效前批准。围绕这条底线的争论在近几年变得尖锐。一派指出，批准如果只是点击，它提供的是责任的外观而不是控制，因而主张提高批准者的理解——要求系统给出与证据相连的说明，要求批准者能够在没有系统的条件下重建关键的那一步。另一派指出，要求人重做机器的全部推理既不现实，也会用信息把真正重要的分歧淹掉。争点是：**批准要成为控制，人必须懂到什么程度？**

第二场在核武器指挥控制。 这个领域对同一个问题给出的答案是反过来的。为了让一次不可撤回的动作不可能由任何单独一个人发起，工程与制度两侧都朝着同一个方向做了几十年：动作被拆成必须由两个人同时完成的两半，两半之间用物理装置隔开，任何一半单独完成什么也不会发生；解锁所需的编码由使用现场之外的层级持有；在设计上，现场的执行者不需要、也不应当掌握整条链的全貌。这里的安全不建立在执行者理解得更多之上，恰恰建立在没有任何一个位置的人拥有全局之上。争点是：**要防住一次不该发生的不可逆动作，靠的是让人懂，还是靠让人各自只懂一段？**

第三场在支付结算法。 这里的答案又换了一个方向。一九七四年，一家德国银行在营业日中途被关闭，它在欧洲时段已经收到了对手方付来的马克，而它应当在美国时段付出的美元还没有付出。问题不在于谁欺骗了谁，而在于一笔交易的两条腿落在不同的时区里，一条腿已经不可撤回、另一条腿还可以撤回。此后几十年，这个领域的全部制度努力都朝着一个方向：把不可撤回的时点**尽可能提前**，并且用法律把它钉死。欧洲在一九九八年通过一部专门的指令，规定结算指令一旦进入系统就不得撤销，即使其后当事人破产、即使破产程序在法律上具有溯及效力，也不得回溯撤销。这条规则的目的写得很直白：可撤销性本身是系统性风险的来源。争点是：**保留撤回的余地，是安全，还是危险？**

三场争论互不引用，却在同一件事上互相顶撞：**一个不可撤回的后果落定之前，控制应当以什么形式存在？**

第一场说：**以理解的形式。** 批准者必须能独立重建那一步，否则批准无效。第二场说：**以分割的形式，而且必须防止理解。** 让任何单个位置都无法凑齐全局，是这套体系的设计目标而不是缺陷。第三场说：**根本不该有窗口。** 撤回余地留得越久，下游承受的不确定越大，正确的做法是把它尽早关死。

三条不能同真。若理解是控制的条件，则核指挥体系里那些被刻意剥夺全局视野的岗位就是失控的——而那套体系是人类为不可逆后果所建的最严的一套，几十年没有出过一次由该体系本身失效导致的误发。若分割加不理解就够了，则支付领域大可以把撤回权分给两方各持一半，而这个领域的判断恰恰相反：撤回权无论分给谁,只要它还在,风险就在。若撤回余地一律有害,则第一场那条底线就该被整个取消——不必留人批准,直接把系统调到最优即可,而没有一个监管者敢这么写。

本文认为，僵局来自一个三方共有、却从未被单独说出的前提：**「什么时候变得不可撤回」是任务本身的属性，「谁在这之前有权动手」是可以在这个给定时点之上另行安排的事。** 第一场在这个给定时点之前塞进理解，第二场在它之前塞进分割，第三场主张把它往前挪。三家争的是往这个时点里放什么、把它挪到哪儿，没有人问这个时点是从哪来的。

而支付结算法自己提供了推翻这个前提的材料，只是它没有往这个方向用：**结算的终局性是一条法律，不是一个物理事实。** 那部指令所做的事，是把一个本来存在的可撤回性用立法消灭掉。既然不可撤回的时点可以被制度制造出来，它就不是任务的属性；既然它不是任务的属性，那么在它之前塞什么，与它落在哪里，就不是两件可以分开处理的事。

顺着这条线往下走，三场争论会合并成一个此前没有被单独提出的问题，而这个问题的答案与理解程度和权限分割都无关。

二、三家各自说到了哪一步

人机协作治理这一支，二〇一八年有一篇被反复引用的哲学论文提出，一个自动系统若要处在有意义的人类控制之下，至少需要两件事：系统的运行应当对相关的人类理由和环境事实保持响应；结果应当能够沿着设计与运行的链条追溯到至少一个有能力理解其道德含义的人。这两条后来成了很多监管文本的骨架。[1]

近年的推进主要发生在两个方向。一个方向是往下追问人这一侧到底要保住什么。二〇二六年初的一篇立场论文提出，长期依赖之后会出现一种它称为「能力—理解落差」的现象：在辅助之下的表现持续提升，而使用者的内部模型持续退化，最终监督沦为程序性动作。该文由此提出一条最低理解阈值，并把它拆成三个方面——核验能力、能保住理解的交互方式、以及支撑治理的制度性脚手架；它明确主张应当在没有实时辅助的条件下检验使用者能否恢复关键推理。[2] 另一个方向是往回追问责任落在哪里。有研究者用一个来自汽车工程的说法描述这样一种反复出现的格局：在复杂自动系统里，人类操作员实际拥有的控制非常有限，却在系统失效之后吸收了主要的道德与法律责任。[3]

这两个方向合起来给出的处方是清楚的：让人懂得更多，让说明来得更早，让责任链条更清晰。它们共同的假设是，控制的短缺是**认知与信息的短缺**。

核武器指挥控制这一支，走的是相反的路。这里的核心问题被表述成一对必须同时满足、方向相反的要求：需要时必须能用，不需要时必须绝对不能用。围绕这对要求形成的制度里，有两件东西最能说明问题。一件是要求任何一次不可撤回的动作都必须由两个人各自独立完成自己那一半，两人互相监督，任一半单独完成不产生任何效果；这条规则的适用范围一直延伸到日常的维护与检查。另一件是把解锁装置做进硬件：没有从上级链条送达的正确编码，现场无论如何操作都无法完成动作，而编码的持有位置刻意与执行位置分离。[4]

这套设计里没有一处依赖执行者理解全局。相反，全局知识在这里被当作风险：一个能独立凑齐全部条件的位置，就是一个可以被策反、被胁迫、被误判所利用的位置。同一文献传统里还记录了这套设计的代价

——它使得在真正需要动作的时刻，反应会变慢、链条会更脆，因而必须在别处补上预先授权与继任安排，否则这套防误发的体系会把整个威慑功能拖垮。**这个代价在该领域是被公开讨论、被明确安排、被专门制度承接的**。这一点在后面很关键。

支付结算法这一支，问题的形状又不同。一九七四年那次银行关闭之后，跨时区结算中「一腿已付、另一腿未付」的敞口成了这个领域的标准案例。随后几十年的制度演进有两条主线。一条是技术性的：把两条腿绑在一起，做到一条腿付出当且仅当另一条腿同时付出，从而在结构上消灭那段敞口。另一条是法律性的：直接立法规定，结算指令一旦按系统规则进入系统，就不得由参与者撤销，也不得因参与者随后进入破产程序而被回溯撤销——即使当地破产法本来赋予破产程序溯及既往的效力。[5] 与之配套的国际标准更进一步，要求这类基础设施明确规定终局性发生的确切时点，并尽可能把这个时点提前到当日之内而不是日终。[6]

这里的判断与前两支都不同：可撤回性不是安全余量，它是负债。只要一笔交易还可能被撤回，所有依赖这笔交易的下游参与者就必须自己扛着一份无法关闭的敞口。这个领域给出的解法不是让谁更懂，也不是把撤回权分给更多人，而是**把撤回权取消掉，并且明确规定取消发生在哪一秒**。

三、冲突落在哪一点上

把三支的处方并排放，冲突的形状就出来了。

治理这一支要延长一个东西：从分歧可见到后果不可撤回之间的那段时间。它的全部工具——更早的说明、更强的核验能力、不实时依赖辅助的训练——都是为了让人能在这段时间里进来。

核指挥这一支要缩小另一个东西：单个位置所能触及的范围。它不延长时间，它切断路径。而且它明确拒绝延长理解——理解的扩散在这里是威胁模型的一部分。

结算这一支要消灭治理那一支想延长的那个东西。在它看来，那段「还能撤回」的时间不是控制的容身之处，而是风险的容身之处；它的成就恰恰是把这段时间压到零并写进法条。

三家的目标两两不兼容，而且不是程度上的分歧。治理那一支若在支付系统里推行，会要求在结算指令生效前保留一个可撤销窗口，而这正是那部指令要废除的东西。结算那一支若在临床里推行，会要求把用药的不可撤回点尽量提前，而这与临床的方向相反。核指挥那一支若在临床里推行，会要求医生不掌握用药的全部理由，这在临床上荒谬。

但把三家的**材料**而不是**处方**并排放，会出现另一样东西。

结算这一支说清了一件前两支没有说的事：**那段可撤回的时间，不是白留的**。 一九七四年那次事故的结构是，付出马克的一方已经不可撤回，而应当付出美元的一方还处在可撤回状态。可撤回性没有消失，它只是全部集中到了一侧；集中在这一侧的可撤回性，在另一侧表现为一份无法关闭、无法定价、无法定位何时结束的敞口。这个领域用了几十年才把这件事讲清楚，而它讲清楚的方式是：**一方的余地，就是另一方的敞口**。

核指挥这一支说清了另一件事：**这种转移可以是故意的，而且可以被明码记账。** 两人规则把「一个人独自发起」这条路径切断，代价是「该动而未动」的风险上升。这个领域没有假装这个代价不存在，它把代价明确算到威慑体系头上，并在别处建了预先授权与继任安排来承接它。转移是被承认的、被安排的、有承接方的。

治理这一支恰恰在这一点上是空的。它主张延长干预窗口，却从不问这段被延长的时间里，不可撤回性去了哪里。一个急诊医生被保留了十秒钟的干预余地，这十秒钟里病情不会停住；一份福利申请被保留了人工复核的环节，等待期间申请人的房租不会停住。窗口是给批准者的，代价先落在别人身上。而这份代价没有出现在任何合规要件里——所有现行的检查表都在问「有没有人批准」「有没有提供说明」「批准者是否受过培训」，没有一份在问「他不批准的时候，谁先付账」。

三家合起来，只能得出一条三家各自都得不出、而且三家都不会同意的判断。

四、逆债

先把这条判断写完整。

一次批准之所以构成控制，不是因为批准者理解了决定结局的那一步，也不是因为动手的权限被分割给了互不知情的多人，而是因为——这次批准之后剩下的那份可撤回的余地，握在一个同时承担「不使用它」的代价的人手里。

与之配套的是一件此前没有被单独命名的东西。为一个人保留一段干预窗口，等于在这段时间里把不可撤回性从他身上移开，移到别处；被移过去的那一份，在别处以「无法关闭、无法定位何时结束的敞口」的形态存在。**这份被转移而没有被记账的不可撤回性，本文称之为逆债。**

三个词需要拆开说。

「转移」而不是「消失」。 保留窗口不产生额外的安全，它只是把不可撤回从一处挪到另一处。急诊里那十秒钟，医生获得了可撤回性，患者的病理过程获得了十秒钟不可撤回的进展。行政复核里那三周，审核员获得了可撤回性，申请人获得了三周不可撤回的等待。这不是修辞。判别办法很具体：**撤走这个窗口，谁的敞口会因此关闭？** 如果撤走窗口之后有人的敞口关闭了，那这个窗口就是从那个人身上借来的。

「未被记账」而不是「转移」本身。 转移不是病。核指挥体系是本文所知记账记得最干净的一处：它明确知道两人规则会拖慢反应，明确把这个代价算给威慑功能，并明确建了别的制度去承接。支付结算体系记的是另一本账：它索性不许有余地，从而不欠这笔债。**病灶只出现在第三种情况——余地被保留了，代价被转移了，而承接方没有名字。**

「债」而不是「成本」。 用债这个词，是因为它有三个成本没有的性质。第一，它有期限而期限不由持有人决定：申请人不能决定复核什么时候结束。第二，它会累积：一个长期低干预率的复核环节，欠下的

不是一次等待，是每一个申请人的每一次等待。第三，也是最要紧的一点，**它可以在账面完全干净的情况下持续累积**——所有的批准都签了，所有的说明都出具了，所有的培训都完成了，而债一直在长。

这条判断为什么三家都不会同意，值得单独说一句。治理那一支不会同意，因为它的整个纲领是「加窗口」，而本文说加窗口在代价不内部化时会把系统从一种病推到另一种病，而不是推向健康。核指挥那一支不会同意，因为它的解法是分割，而本文说分割本身既不产生也不消灭这笔债，它只改变债的形状。结算那一支不会同意，因为它的立场是可撤回性一律有害，而本文说可撤回性的害处不在它存在，在它的持有人不付账。

而三家的材料合起来只能得出它：结算法证明了转移的存在与它的破坏力，核指挥证明了转移可以被记账并且记了账就不出事，治理那一支提供了大量转移未被记账的现场。缺任何一家，这条判断都立不住——只有结算法，会得出「取消一切窗口」；只有核指挥，会得出「分割就够了」；只有治理那一支，会得出「加窗口加理解」。

逆债有三个签名，三个都可以在不访谈任何人的情况下读出来。

第一个签名：干预率随时间单调下降，并趋于一个很小的常数。 这条曲线的形状不由人决定。每一次干预都要付出成本——多花的时间、与上游的冲突、被要求解释的风险；每一次不干预不付出任何成本。在这样的成本结构下，任何一个理性的占位者都会向下走，而且走法与他的职业道德无关。**最能说明问题的一点是：换一批人，曲线会重新走一遍同样的形状。** 这条可以在任何积累了几年日志的复核环节上直接验证，不需要新的数据采集。

第二个签名：形式完整度不下降，甚至上升。 债在累积的同时，批准记录、说明文档、培训证明都保持完整，因为这些东西的产出成本低而收益明确（它们能通过审查）。于是出现一种特有的组合：所有可审查的指标都在改善，而这个位置实际起的作用在归零。这个组合本身就是诊断依据——**指标改善与作用萎缩同时出现，是右上格独有的形态；** 在左下格里，形式完整度会因为占位者的挫败而下降，在左上格里，形式与实质同向。

第三个签名：短期内表现为效率提升。 复核环节的干预率从一成二降到三分，在任何一张运营报表上都是好消息：处理时长下降、返工下降、争议下降、成本下降。正在积累的那一份不出现在报表里，因为它不在这张报表的主体身上。**这解释了为什么逆债几乎从不被主动发现——发现它需要看一张属于别人的账。**

三个签名合起来还给出一件事：逆债不会自己暴露。它不像一次事故那样有时间戳，也不像一次违规那样有条文可依。它只有在两种情况下才会被看见——要么有人去看那张属于别人的账，要么等到累积的敞口在别处以一次集中的失败爆出来，而那次失败通常会被归因到别的东西上，因为它与任何一次具体的不干预都对不上号。

还需要澄清一件容易混淆的事。逆债不是通常所说的成本外部化。外部化说的是代价落到交易之外的第三方，处方是把第三方拉进交易、明晰权利、允许议价。逆债说的是代价落在**同一条责任链内部**的一方身上，而且落在一个他无法定位的时刻。急诊里的患者不在交易之外，他是这条链的目的；他的问题不是没有权利，是他的权利在代价发生的那一刻无法行使。这个差别会在后面变成一条可以判决的分离线。

五、两根轴，四种局面

只用一根轴分不开这些局面。「还剩多少可撤回的余地」这一根轴上，急诊医生的十秒与软件工程师回滚前的两小时会落在同一侧，而它们的结局相反。要把它们分开，需要第二根轴。

第一根轴：这次批准之后，系统里**是否还剩下可用的可撤回余地**。可用包含三件事——余地在时间上还没有关闭，握有它的人有权使用，且他有能力知道该在什么时候使用。三者缺一，余地就只是名义上的。

第二根轴：这份余地的持有人，**是否同时承担「不使用它」的代价**。这里的承担指的是代价先落在他身上，而不是事后可能被追究。事后追究是另一件事，它发生在代价已经落定之后，且通常无法定位到具体的哪一次不使用。

	持有人承担不使用的代价	**持有人不承担不使用的代价**
批准后仍有可用余地	**实质控制** 余地在他手里，不用它的后果也先落在他手里。干预率与干预质量同向。这一格稀缺且昂贵。	**逆债** 窗口在，代价不在。全部形式审查都能通过：有人批准、有说明、有培训、有责任链。干预率长期趋零而无人察觉。
批准后已无可用余地	**溃缩** 没有余地却要担责。事故之后责任向最近的那个人集中，而他当时什么也做不了。	**已终局** 余地没有，也没有人被要求承担。这不是病，是很多系统的正常状态与设计目标。

四格里有两格需要立刻说清楚，因为它们最容易被误读。

右下格不是病。 一笔支付完成结算之后，没有人还握着撤回的余地，也没有人被要求为「没有撤回」承担什么——因为本来就不该撤回。这一格在支付、在登记、在已经生效的判决、在已经发射的物体上都是常态，而且是几十年制度努力争取来的结果。本文不主张越可撤回越好。任何把这一格一律读成失控的框架，都会在第一次接触结算法时崩掉。

左下格与右上格的区别，是这张表最要紧的一处。 左下格已经有名字，它描述的是一个人被要求为自己无力改变的结果负责。右上格没有名字，而它更常见、更难被发现，也更难被修正。因为左下格会在事故之后暴露——追责会追到一个显然做不到的人身上，这种荒谬本身是信号；而右上格在事故之前和之后都不产生信号。事故之后，人们会发现批准者当时确实有权干预、确实有时间干预、也确实受过培训，于是结论会是「他没有尽责」，而不是「这个位置从一开始就不会有人干预」。

两格之间还有一个可以直接观测的分别：**左下格的干预率可以很高，右上格的干预率一定很低。** 在没有余地的位置上，人会做很多动作——追问、记录、上报、要求复核——只是这些动作改变不了结果。在有余地而不承担代价的位置上，人做的动作会随时间单调减少，因为每一次干预都要付出成本（时间、冲突、被质疑），而不干预不付出成本。这条差别不需要访谈，只需要看时序上的干预率曲线。

右上格还有一个特征使它格外难被发现：**它在短期内表现为效率提升。** 一个复核环节的干预率从百分之十二降到百分之三，在任何一张运营报表上都是好消息——处理时长下降、争议下降、成本下降。系统正在积累的东西不出现在报表里，因为它是别人的敞口。

六、一个不含情态词的问法

判断一个位置落在哪一格，可以不用「应当」「有意义」「实质性」这类词。只需要问一句：

如果这个人从不使用他的干预权，代价先落在谁身上？

「先」这个字承担了全部重量。它问的不是最终谁负责，而是时序上谁第一个承受。事后的责任分配几乎总能追到批准者，那条路径不提供任何区分能力。而代价第一个落在谁身上，是可以在具体流程里查出来的。

把这个问法在五个场景里跑一遍，答案互不相同，且都不是从命题里推出来的。

急诊。 一名医生面对迅速恶化的患者，系统给出建议后只剩很短的时间。他若不使用干预权，代价先落在患者的病理过程上。医生事后可能被追究，但那次「没有干预」在时序上无法被定位——没有人能指出是哪一秒的不作为导致了什么。判定：右上格，逆债。这与原来那种「时间不够所以人不该在环里」的说法结论相近，但理由不同，而且不同的理由给出不同的处方。按时间不足的说法，处方是缩小系统可以自主处理的病例范围；按本文，缩小范围只解决了一半——另一半是让「不干预」的代价在时序上可定位，比如强制记录每一次未干预的当时状态，使它成为一个可被回溯到具体时刻的事件。这条处方在时间上不花成本，因而在急诊里可行，而缩小范围要花成本。

福利审核。 审核员对系统给出的拒绝建议点确认。他若从不干预，代价先落在申请人的等待与被拒上。审核员本人不承受任何东西——干预要写理由、要承担被上级质疑的风险、要拉长处理时长（而处理时长通常是他的考核项）。判定：右上格，而且是本文所见最纯粹的一例。这里可以顺带解释一件长期没有好解释的事：为什么行政领域的人工复核几乎必然退化成盖章，而这个退化与审核员的素质、培训、职业道德都关系不大。它不是人的问题，是这一格的问题。

代码回滚。 一名工程师在不可逆的数据迁移开始前有回滚窗口。他若不回滚，代价先落在他自己所在的团队——半夜被叫醒的是他，写事故报告的是他，下一个迭代被推迟的是他的迭代。判定：右上格与左上格之间，偏左上。这给出一条可以直接检验的跨领域预测：**「人在回路」的实际有效性排序应当是软件工程高于临床、临床高于行政审批，而这个排序不由技术复杂度、从业者受教育程度或监管强度解释，只由代价归属解释。** 若实证排序与此不同，且差异能被上述三个因素解释，本文这一条即失效。

跨境支付。 结算指令进入系统后，没有人握着撤回余地。若问「不撤回的代价先落在谁身上」，答案是没有代价——本来就不该撤回。判定：右下格，健康。这一例的价值在于它是本文的边界样本：一个完全没有人在回路、完全不可撤回的系统，按本文判定为健康，而按治理那一支的默认排序会被判为最危险。两者预测相反，可以分开检验：本文预测这类系统的失效模式集中在**进入之前**的准入与校验环节，治理那一支预测失效模式集中在缺少人工干预点上。

核发射链条。 两人规则之下，任一位置若不完成自己那一半，代价先落在威慑功能上——一个无法在需要时动作的威慑体系等于没有威慑。这个代价既不落在执行者身上，也无法在时序上定位。按本文判定：右上格。**而这个领域自己知道这一点，并且已经在别处补了账**——预先授权、继任安排、多重独立链路，都是为了承接这笔债。这就引出后面要说的第一处收窄。

这个问法还有一个附带的好处：它把一个通常需要专家判断的问题，变成了一个流程分析问题。判断「批准者是否真正理解了关键推断」需要领域专家、需要设计测验、需要控制表演性回答；判断「他不干预时代价先落在谁身上」只需要读流程文件和考核办法。前者昂贵、主观、可被表演；后者便宜、客观、难以伪装——因为代价的流向写在别人的合同、别人的考核和别人的等待里，不由被评估者自己陈述。

它也解释了一件在合规实践里长期被当作执行问题的事。监管者反复抱怨「人在回路」在落地时退化成盖章，通常的归因是机构不重视、人员不专业、培训不到位。按本文，这不是执行问题：只要代价不落在批准者身上，退化就是这个结构的稳定态，重视、专业与培训都改不了稳定态的位置，只能改到达稳定态的速度。**一个每年重新培训的机构，与一个从不培训的机构，最终会落在同一个干预率上，只是前者慢一些**。 这是一条可以直接检验的、代价很低的预测。

七、逐领域兑现

下面十二处兑现里，前十处是命题的应用，后两处反过来改了命题本身。

一、临床用药审核。 医院药师对处方拥有否决权。若他从不否决，代价先落在患者。但药师制度里有一件东西改变了格局：药师对因未拦截而发生的用药错误负有独立的专业责任，而这份责任在很多司法辖区是可以被单独追究的，不与处方医师混同。这使代价在时序上部分可定位——事故复盘会问「这张处方经过了谁的审核」。这解释了为什么药师否决权比行政复核有效得多，尽管两者在流程图上是同一种结构。

二、法律文书送达前的复核。 一份行政处罚决定在正式送达前可以被撤回。复核者若不撤回，代价先落在相对人的权利上。此处的判定与福利审核相同。但法律领域提供了一个现成的、已经被验证的修法方向：把「未经实质复核而送达」本身规定为程序违法，使得不复核在时序上产生一个可定位的瑕疵点。这不是要求复核者更懂，是把不作为变成一个有位置的事件。

三、自动驾驶的接管请求。 系统请求接管，驾驶者有数秒。他若不接管，代价先落在他自己身上——这是本文所举例子里代价内部化程度最高的一处。按本文预测，同等技术条件下，驾驶者的实际接管率应当高于其他领域中批准者的干预率；而现有的接管研究关注的多是接管的**质量与时延**，很少报告在明确请求下的**不接管率**。这是一个现成的、数据可能已经躺在那里的检验点。

四、航空的自动化脱离。 与自动驾驶结构相同，代价高度内部化，但加了一层：机组是两人，而两人共同承受同一份代价。这解释了为什么航空的双人制与核指挥的两人规则表面相似而机能不同——前者是两个都承担代价的人，后者是两个都不承担代价的人。

五、内容审核。 审核员对自动标记的结果做人工复核。若不推翻，代价先落在被误删的发布者身上；若推翻错了，代价落在平台的合规风险上，而这份风险会被计入审核员的考核。**代价在两个方向上不对称**：

错误保留会被追究，错误删除不会。本文预测这类位置的干预会系统性偏向一个方向，且偏向的幅度与两侧代价的可定位性之差成正比——这是一条可以用平台自己的日志直接检验的预测。

六、科研的伦理审查。 审查委员会若从不否决，代价先落在受试者身上，且在时序上极难定位。这一格与行政审核同构。已有的修补方向——事后不良事件必须回溯到具体的审查意见——正是把代价可定位化的动作。

七、供应链的停线权。 生产线上任何工位都可以拉绳停线。若不拉，代价先落在下游的返工与最终的缺陷上。这套制度的关键设计常被误读成「授权」，而它真正做对的是另一件事：停线的代价被明确算给班组而不是拉绳的个人，且拉绳之后立刻有人到场，使得「拉绳」的成本低于「不拉绳」的成本。这是本文左上格已知的最成熟的一个实现。**本站此前一篇讨论结算权归属的文章曾用这套制度说明「谁来执行结算」**；此处的用法不同：那篇问的是终止的权力落在系统内还是系统外，本文问的是这份权力的持有人是否付不用它的账。两者可以分开——一个把停线权完全留在班组内部、却让停线的代价全部由下一班承担的工厂，那篇会判为健康，本文预测拉绳率会随时间趋零。

八、临床试验的中止规则。 独立的数据监查委员会按事先写死的规则决定是否中止试验。这里的设计很特别：委员会成员看不到全部信息，且不承担试验失败的代价。按本文，这应当是右上格。但它运转良好，原因在于第二根轴被一件事替换掉了——**中止规则是事先写死的，触发条件到了必须中止，不留裁量。** 这提示了一条本文之外的解法：当代价无法内部化时，可以用取消裁量来替代。这条解法的代价是丧失应变能力，因而只适用于失败模式可以被事先穷举的场合。

九、信贷审批。 信贷员对模型的拒绝建议若从不推翻，代价先落在被拒的申请人身上；若推翻错了，坏账会计入他的考核。与内容审核同构，且不对称方向相同。本文预测推翻率会低于最优水平，且这个偏差在坏账追责周期越短的机构里越大。

十、编辑对稿件的处理。 若从不改动作者交来的稿子，代价先落在读者身上，而读者不会向编辑追究。这一处放在这里，是因为本文自己就站在这一格上，后面会回到它。

十一、核指挥体系反过来加的约束。 前面已经判定两人规则处在右上格：持有人不承担不动作的代价。但这套体系是人类为不可撤回后果所建的最严的一套，几十年没有出过由该体系本身失效导致的误发。若本文坚持右上格一律是病，就会得出一个荒谬的结论。**因而命题必须收窄：转移本身不是病，未被记账的转移才是。** 核指挥体系欠下的这笔债是故意欠的，欠给谁是明确的（威慑功能），并且在别处专门建了制度去承接（预先授权与继任安排）。这条收窄把本文从一个更强、也更容易被反例击倒的断言，退成一个更准的断言，代价是它不能再单凭「代价没有内部化」判定一个系统有病——还必须查有没有别处的承接方。**这是本文第一处主动削弱。**

十二、支付结算法反过来加的约束。 前面已经把右下格判为健康常态。这条约束比第一条更重，因为它拿掉的是一整套价值排序。治理那一支的默认排序是：可撤回优于不可撤回，有人在环优于无人在环。结算法说的恰好相反：在这个领域，几十年的努力方向是**取消可撤回性**，而且这个方向被证明是对的。**因而本文不含「越可逆越好」这一排序，也不能把「无人在环」当作失控的同义词。** 本文只主张一件更窄的事：**凡是保留了余地的地方，就要问这份余地的持有人付不付账。** 至于要不要保留余地，本文不回答

——那是另一个问题，且不同领域的答案不同。**这是本文第二处主动削弱，也是更彻底的一处：它把本文的适用范围从「所有人机系统」缩到「已经决定要保留干预余地的人机系统」**。

八、怎样把一个位置移出右上格

诊断出一个位置处在右上格，接下来的动作不是加培训、不是加说明、也不是换人。这三样在右上格里都无效，而且它们无效的方式是可以预先说出来的：培训提高的是理解，而这个位置上的人不缺理解，缺的是使用理解的理由；说明提高的是信息，而这个位置上的人不缺信息；换人会让曲线重新走一遍同样的形状，因为形状不由人决定。

真正能动的有四条路，代价依次上升。

第一条：把不使用变成一个有位置的事件。 这是最便宜的一条，因为它不改变任何人的利害，只改变记录方式。现在的日志普遍只记「批准了什么」，不记「在什么状态下没有干预」。把后者变成一条必须落笔的记录——未干预时的当时读数、当时可见的替代方案、当时的时间余量——不需要多花时间，却把一件原本弥散在时间里的不作为变成了一个可以被回溯到具体时刻的事件。它不会自动让代价落到批准者身上，但它把代价从「不可定位」变成「可定位」，而不可定位正是逆债之所以能无限累积的原因。行政程序法里那条把「未经实质复核而送达」规定为程序违法的思路，做的就是这件事。

第二条：把代价在时间上拉近。 很多位置上的代价并非不可归属，只是归属得太晚——晚到与那一次不作为之间的联系已经断了。缩短这个距离往往比改变归属容易。信贷环节的坏账追责周期若从三年缩到一年，本文预测推翻率会上升；不是因为责任变重了，是因为责任与动作之间的距离变短了。这条路的风险在于，拉近的若只是一侧的代价，会把偏向拉得更歪——内容审核里错误保留的代价本来就比错误删除的代价近得多，再拉近它只会让误删更多。**拉近必须两侧同时拉，否则不如不拉。**

第三条：换一个已经承担代价的人来握这份余地。 这是组织层的改动，成本高但有效。航空的双人制之所以与两人规则机能不同，就在这里：两个都在同一架飞机上的人共同承担同一份代价。同样的道理可以解释一个反直觉的现象——把某项复核权从一个专职复核岗移交给后续环节的执行者，往往会提高复核质量，尽管执行者的专业训练更少。他更少懂，但他要在下一步吃到这份后果。这条路的边界很清楚：它只在代价确实会流到某个具体位置时可用，对威慑功能这类弥散的代价无效。

第四条：取消裁量。 当代价在原理上无法归属到任何一个具体位置时，只剩这一条。事先把触发条件写死，条件到了必须动作，不留判断空间。临床试验的独立数据监查委员会走的就是这条路：委员不承担试验失败的代价，也看不到全部信息，但中止规则是事先写死的。这条路买来的是可靠，卖掉的是应变——凡是失败模式无法事先穷举的场合，写死的规则会在第一个未预料的情形面前失效，而且失效时没有人有权变通。**因而它只适合那些「宁可错停也不可漏停」的场合。**

四条路里前两条不动组织，后两条动组织。一个现实的判断是：绝大多数被诊断为右上格的位置最终只会走第一条，因为它不需要任何人让渡什么。而只走第一条是不够的——它把债变得可见，并不偿还它。可见本身有价值，但把可见误当作已解决，是这套判断最容易被误用的方式。

还有一件事需要说在前面。上面四条都有一个共同的前提：这个位置**应当**有人干预。如果一个位置根本不需要人，最干净的处理是把它取消，而不是给它配代价。支付结算走的正是这条——它没有去改造撤回权的归属，它取消了撤回权。**取消一个没有人会用的按钮，比给这个按钮配一套问责制度要诚实得多，也便宜得多。** 现行合规框架的一个系统性偏差在于，它几乎从不允许「这里不需要人」这个答案，于是每一处都留下一个按钮，而每一个不必要的按钮都是一笔新债。

九、与最近的几种说法的分界

一个新说法若不能在同一个案例上给出与旧说法不同的预测，它就只是换了个名字。下面八条每一条都落到具体的可观测差异上，且每一条的判决点各不相同。

一、与有意义的人类控制。 那套框架要求系统对人的理由保持响应，并要求结果能追溯到一个有能力理解的人。它管的是追踪与追溯，本文管的是代价的时序归属。判决点：取两个系统，都能把结果追溯到某个具名的人，都提供同等质量的说明，只有一个把「不干预」的代价在时序上落到批准者自己身上。追溯框架预测两者的控制水平相同；本文预测两者的干预率显著不同，且差异随时间拉大。[1]

二、与最低理解阈值。 那篇论文测的是人这一侧的认知存量，并主张在没有实时辅助的条件下检验能否恢复关键推理。**顺带更正一处：该文的三个方面是核验能力、能保住理解的交互方式、以及支撑治理的制度性脚手架，而不是核验、重建与边界意识——后一种概括在中文讨论里已经流传，与原文不符。** 判决点：在周期性测验中理解水平相同的两组，若代价归属不同，本文预测实际干预率不同；理解阈值那一支预测相同。更强的一条：本文预测理解水平会**跟随**代价归属变化而不是相反——一个长期不承担不干预代价的位置，其占位者的理解会自行退化，因为理解在这个位置上没有用处。这条把因果方向掉了个个儿，可以用纵向数据分开。[2]

三、与责任的溃缩。 那个说法描述的是没有实际控制却吸收主要责任的格局，即本文的左下格。判决点在时间上：溃缩那一支的读数出现在事故之后（责任向最近的人集中），本文的读数出现在事故之前（该位置的干预率已长期趋零，且可从日志直接读出）。两者可以在同一个系统上同时检验：若一个系统事故后责任集中而事故前干预率并不低，则是溃缩不是逆债；若事故前干预率长期趋零而事故后责任反而分散，则是逆债不是溃缩。[3]

四、与道德风险与委托代理。 这是本文最近的一位经济学占位者，必须正面说清。委托代理里的核心困难是**行动不可观测**：代理人可以偷懒而委托人看不见，因而处方是监控、是把报酬与可观测结果挂钩。本文说的困难不同：**批准者的行动完全可观测**——每一次点击、每一次未点击都有日志——困难在于代价发生在别处、别时，观测性解决不了它。判决点很干脆：加强监控是道德风险的标准处方；本文预测在右上格里加强监控不改善干预率，只增加防御性文书，而改变代价归属会改善。这两条处方可以在同一个复核环节上做对照，成本很低。[7]

五、与外部性。 外部性说的是代价落到交易之外的第三方，处方是明晰权利并允许议价。本文说的代价落在同一条责任链内部，而且落在一个他无法行使权利的时刻。判决点：明晰权利并允许议价，是外部性

理论的处方；本文预测这条处方在急诊里完全无效，因为承受代价的那一方在代价发生的那几十秒里没有议价能力，也不知道自己正在承受什么。反过来，在代价发生后仍有议价空间的场合（比如信贷被拒后的申诉），两条理论的处方会收敛，因而那个场合不适合用来分辨二者。[8]

六、与结算风险。 这是本文最近的一位一比一威胁：那个领域早就把「一侧已不可撤回、另一侧仍可撤回」识别成风险源。分离线在机制上：结算风险的机制是**同步失败**，两条腿落在不同时刻；处方是把它们绑在一起同时完成，做到了就消灭了风险。本文的机制是**权利与代价的持有人分离**，即使完全同步也照样成立。判决点：同步交收之后，结算风险这一支预测该问题被消灭；本文预测仍留下一个它管不到的东西——谁承担「本可结算而未发起结算」的机会成本。这个残余在同步交收系统的运行数据里应当仍然可见。[5][6]

七、与两人规则。 那套设计预测分割权限会降低未授权动作的风险，这一点已被几十年的运行支持。本文预测它同时抬高「该动而未动」的风险，且这份风险必须由别处的制度承接，否则整套体系的功能会退化。判决点：这两个预测指向两个不同的读数——前者是误发次数，后者是响应时延与继任安排的完备程度。若能找到一个长期运行的两人规则体系，其防误发有效而完全没有任何补偿性安排、功能却不退化，本文这一条即被驳。[4]

八、与警报疲劳。 这是最容易与本文混淆的一个，因为两者的读数看起来一样：响应率随时间下降。分离线在成因上：疲劳的成因是刺激频率与假阳性比例，处方是降频、提高特异度。本文的成因是代价归属，与频率无关。判决点很清楚：把警报频率降到十分之一，疲劳理论预测响应质量提升；本文预测在代价完全外部化的位置上，响应率会回升一阵然后重新趋零，因为回落的驱动力没有变。这条对照在医院的警报系统上有现成数据可做。

十、三家各自为什么到不了这一条

值得单独说一句：为什么三家各自走了几十年，都没有走到这一条上。

治理那一支到不了，是因为它的问题意识来自事故之后的追责。 从追责往回看，最刺眼的是「签字的人不懂」，因而全部注意力落在认知上。而代价归属这件事在追责视角里是不可见的——追责天然是事后的、总量的，它问「这次谁负责」，不问「不作为的代价在时序上先落在谁身上」。

核指挥那一支到不了，是因为它的威胁模型里只有一种错误。 它要防的是不该发生的动作，因而它把全部结构做成路径切断。「该动而未动」的风险它知道，也补了，但它把这件事当作工程上的补丁而不是理论问题，因为它的对象只有一件事、只有一个后果。没有对象的多样性，就不会去问「这类代价一般落在谁身上」。

结算那一支到不了，是因为它把问题解决得太干净了。 它选择的是取消可撤回性，而不是管理可撤回性。一旦选了取消，「持有人付不付账」这个问题就不再出现——没有持有人。它对转移的认识极深，深到足以推出本文的一半，但它从不需要走完另一半。

三家各自的盲点合起来，恰好覆盖了这条判断所需的全部材料。

十一、推论、适用边界与反噬

第一条推论：加一个按钮很容易，让按钮的持有人付账很难，因而系统会自然地漂向右上格。 这不是懈怠，是成本结构。加按钮的成本一次性且可见，改代价归属的成本持续且要动考核、动责任分配、动组织边界。任何只要求「有人在环」的合规框架，都会在几年内把被监管的系统整体推到右上格，而每一次审查都会通过。

第二条推论：干预率的时序曲线是最便宜的诊断工具。 它不需要访谈、不需要问卷、不需要理解水平测验，只需要日志。右上格的签名是干预率随时间单调下降并趋于一个很小的常数，而这个下降与人员更替无关——换一批人，曲线会重复同样的形状。这条可以在任何已有几年日志的复核系统上直接跑。

适用边界一：本文只管已经决定要保留干预余地的系统。 要不要保留余地，是另一个问题，且答案因领域而异；支付结算给出的答案是不保留，而它是对的。

适用边界二：代价可定位性有下限。 有些代价在原理上无法定位到某一次不作为——威慑功能就是这样一个。对这类代价，本文的处方（内部化）执行不了，只能退到临床试验中止规则那一路：取消裁量，用事先写死的触发条件替代人的判断。这条替代方案有它自己的代价，即丧失应变能力。

适用边界三：本文不处理代价在两个方向上都存在的情形的最优点。 内容审核与信贷审批里，干预与不干预各有代价，本文只能预测偏向的方向，不能给出应当偏到哪里。那需要一个本文没有的、关于两侧代价相对权重的判断。

反噬预言。 若「不干预的代价先落在谁」成为一项合规要件，最可能发生的不是代价被真正内部化，而是**代价被形式上内部化**：机构会设立一项针对未干预的考核指标，然后批准者会开始做防御性干预——干预那些成本低、看起来像干预、而实际不改变结果的项。这会让干预率曲线回升，同时干预质量下降，于是本文自己提出的最便宜的诊断工具会被废掉。**这是本文预见到而无法防住的一件事。** 唯一的部分对策是把干预率与干预的实质效果分开测——但实质效果的测量成本高，而这恰恰是当初大家只测干预率的原因。这个循环没有出口。

还有一个交出去的洞。 本文假定代价的持有人是可以被指认的。在多层外包、跨机构的链条里，代价可能落在一个由多方共同构成、谁也不完整承担的位置上。这时第二根轴取不到值，四格表退化成一根轴，而本文没有办法处理这种情形。

十二、与同期几篇的分界

本文与同期发表的另外六篇处理的是同一片地面，因而必须逐一说清分界，否则会互相冒充。

与「谁能改下一轮的标准」那一篇。 那篇问的是结果出现之后，谁有权修改下一轮的问题与评价尺度。本文问的是不修改的代价先落在谁身上。判决点：取一个把标准修改权完全交给使用者、而修改与不修改都不对使用者产生任何代价的系统。那篇预测认知健康提升，本文预测修改率会随时间趋零，与授权无关。

与「机器答案到达之前留一份独立记录」那一篇。 那篇要求前置留痕，本文要求后置付账。判决点：一个底稿制度完备、而不写底稿或写得敷衍都不产生任何代价的环境里，本文预测底稿的实质内容会随时间单调劣化到只剩格式，而形式完整度不下降——于是所有基于完整度的审查都会通过。

与「学习的单位是规则的前后差」那一篇。 那篇的机制在个人的判断结构里，本文的机制在代价的分配结构里。判决点：同一批学习者，规则差记录质量相同，若「不改写规则」的代价一组落在自己身上（下一次同类题会重考）、一组不落，本文预测迁移成绩仍有差异；那篇预测无差异。

与「中断之后能不能接手」那一篇。 那篇测的是能力，本文测的是动力。判决点：接续能力测验分数相同的两组，代价归属不同者的实际接手率不同。这一条尤其要紧，因为能力测验会在右上格里给出很好的分数——那个位置的人完全有能力接手，只是没有理由接手。

与「谁先提出、谁先否决」那一篇。 那篇把否决权的成立条件写成「具有拒绝效力」。本文认为拒绝效力只是第一根轴，有效力不等于有动力。判决点：在首否决权制度之上再加一层代价内部化，那篇预测校验能力的沉淀不变（因为否决权已经成立），本文预测沉淀才开始出现。

与「闭合之后还能不能回来」那一篇。 那篇问的是二次进入的通道是否还在，本文问的是不进入的代价落在谁身上。判决点：一个留痕完备、触发规则明确、局部可修改的系统，若二次进入的成本全部由发起者承担而收益归他人，那篇预测长期纠错能力良好，本文预测二次进入率趋零，通道空置。两篇的处方也因此不同：那篇修通道，本文修账。

另外还有两处与本站早前的文章接壤，一并说清。**与那篇讨论闭合之后是否还留着一道门的文章**：它问的是闭合之后开放度还剩多少，本文问的是这份开放度的持有人付不付账；一个门大开而无人愿意推的系统，那篇判为健康，本文判为逆债。**与那篇讨论终止权归属的文章**：它问终止权在系统内还是系统外，本文问终止权的持有人是否承担不终止的代价；一个终止权完全在内、而不终止的代价全部由下游承担的系统，那篇判为健康，本文预测终止动作会趋零。

十三、本文自己落在哪一格

这篇文章保留了一份可撤回的余地：读者可以不接受它。若读者从不使用这份余地，代价先落在谁身上？落在读者——被一个不成立的框架占用的注意力、被它引导偏的判断。而作者不承受任何东西。按本文自己的表，这篇文章处在右上格。

更具体的一处：本文把赌注写死在二〇三一年。在此之前的五年里，本文享有「尚未被证伪」的地位，而承担这五年的是引用它的人。这是一笔标准的逆债，而且是本文自己在正文里制造的。

能做的对冲有限，但不是没有。第一，把最便宜的那条检验写清楚到可以被别人直接跑——干预率的时序曲线，任何有几年日志的机构今天就能跑，不需要作者参与。第二，明写本文自己预期会失败的地方（第十一节的反噬与那个交出去的洞），使失败不必依赖作者承认。第三，本文的两处收窄是在写作过程中被材料逼出来的，两处都削弱了原来更强的断言，两处都写在正文里而不是致谢里。

这三条对冲不改变本文所处的格。它们只是把这笔债的期限和金额写在了明处。

十四、可以判本文错的六种结果

第一，最便宜、今天就能做的一条。 在同一任务、同一界面、同一说明质量下，只改变一件事：不干预造成的错误，一组由被试自己承担（下一轮任务加重、或直接扣减报酬），另一组由他人承担。本文预测代价内部化组的干预率显著更高，且假阳性干预不同比例上升。若两组干预率无差异，或内部化组的上升全部来自假阳性，本文的核心机制不成立。

第二，跨领域排序。 「人在回路」的实际有效性应当呈现软件工程高于临床、临床高于行政审批的排序，且这个排序在控制技术复杂度、从业者教育水平与监管强度之后仍然成立。若实证排序不同，或差异可被上述三者解释，本文的跨域主张失效。

第三，与问责加压的分离。 给行政复核环节加一项「因未干预造成的错误计入考核」，本文预测实质干预率上升；而一般的问责加压理论预测上升的主要是防御性动作而非实质干预。两者预测不同，可以用干预后的结果改变率分开。若上升的全部是防御性动作，本文退化为问责理论的一个特例。

第四，打第七节第十一处收窄。 若能找到一个长期运行的两人规则体系，其防误发功能有效、而完全没有任何补偿性的预先授权或继任安排、功能也不退化，则「转移必须在某处被承接」这一条被驳，本文的记账要求随之失去根据。

第五，打第七节第十二处收窄。 若某个市场基础设施把结算终局性推迟，而系统性风险没有上升、下游敞口没有扩大，则「一方的余地就是另一方的敞口」这条转移律不成立，本文的整个第一根轴失去支撑。

第六，本文自陈最软的一处。 也许两根轴其实是一根：代价归属决定了余地是否真实可用，因而第一根轴上的「可用」已经把第二根轴吃掉了。要把它们分开，必须找到一个案例——余地真实存在且可用，代价完全外部化，而干预仍然频繁发生。本文目前找不到这样的案例。找不到不等于不存在，但若长期找不到，正确的做法是把两根轴合并，而不是继续维持四格表。

正向读数。 若本文成立，应当能观察到：在代价被内部化的位置上，干预率不随时间下降；在代价外部化的位置上，干预率随时间单调下降且与人员更替无关；把一个位置的代价归属改掉之后，干预率的变化应当早于任何理解水平的变化出现。第三条是最有力的，因为它规定了两个变量的先后顺序，而顺序比相关更难被巧合满足。

写死日期的赌注。 到二〇三一年十二月三十一日，至少应当有一个具有实际约束力的高风险自动系统监管框架，把「未干预的代价归属」写进人在回路的合规要件，而不再只要求「有人批准」「提供说明」「完成培训」。这里的「写进」必须是可执行的审核条款，只要求机构自行说明代价如何分配不算命中。若到该日期没有任何具有规范效力的制度采用这类要求，且干预率的时序曲线在多个领域被证明与代价归属无关，则本文关于评价重心将从「谁批准」转向「不批准的代价先落在谁」的预测即告失败。

结　论

一个不可撤回的后果落定之前，控制以什么形式存在，三个领域给出了三个不相容的答案：以理解的形式、以分割的形式、以及根本不该有余地。三个答案都对，但都只对了一段，因为它们共享一个未被说出的前提——不可撤回的时点是任务给定的，而在它之前放什么是可以另行安排的。支付结算法自己证伪了这个前提：终局性是一条立法，不是一个事实。

时点既然可以被制造，那么保留余地就不是在给定的时间里塞进控制，而是把不可撤回性从一处挪到另一处。挪走的那一份不会消失，它在别处以一份无法关闭、无法定位何时结束的敞口存在。挪动本身不是问题——核指挥体系挪得最狠，也记得最清楚，并且在别处建了承接的制度；支付结算索性不挪，因为它不留余地。问题只出在第三种情形：余地留了，账没记，承接方没有名字。

于是判断一次批准算不算控制，不必去问批准者懂多少，也不必去数有几个人分持权限。只需要问一句不含任何情态词的话：**如果这个人从不使用他的干预权，代价先落在谁身上。**　　落在他自己身上，那是控制；先落在一个无法定位这一刻的人身上，那就是一笔正在累积的债，而账面上一切正常。

最需要警惕的从来不是那些明显失控的系统。是那些每一项检查都通过、每一次批准都签了字、每一份说明都出具了、而那只手已经很多年没有真正按下去的系统。

注　释

1. 本文所说的「代价先落在谁身上」是时序判断，不是责任判断。事后的法律与道德责任如何分配，是另一个问题，本文不处理。
2. 「可用的余地」包含时间未闭合、有权使用、有能力知道何时使用三件事；三者缺一，第一根轴取低值。
3. 本文引用的核指挥体系材料均来自公开出版的学术与政策文献，不涉及任何具体国家的现行操作细节。
4. 第六节五个场景的判定是按公开可查的制度结构作出的，不针对任何具体机构；同一制度在不同机构的实现差异可能改变判定。
5. 本文提出的四格表尚未经过信度与效度检验，是研究设计的起点，不是成熟的评估工具。

参考文献

1. Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. Frontiers in Robotics and AI, 5, Article 15. frontiersin.org
2. Lin, F., Ge, Q., Xu, L., Li, P., Gao, X., Xing, S., Yamada, K., Zhang, Z., Zhang, H., & Tu, Z. (2026). Position: Human-Centric AI Requires a Minimum Viable Level of Human

Understanding. arXiv:2602.00854（预印本，引用前请核对版本）。 arxiv.org/abs/2602.00854

3. Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. Engaging Science, Technology, and Society, 5, 40-60. estsjournal.org
4. Feaver, P. D. (1992). Guarding the Guardians: Civilian Control of Nuclear Weapons in the United States. Cornell University Press.（两人规则与准许行动链的制度史与代价分析）
5. European Parliament and Council. (1998). Directive 98/26/EC on settlement finality in payment and securities settlement systems. eur-lex.europa.eu
6. Committee on Payment and Settlement Systems & IOSCO. (2012). Principles for Financial Market Infrastructures. Bank for International Settlements.（终局性时点的明确化要求见原则八） bis.org
7. Holmström, B. (1979). Moral hazard and observability. The Bell Journal of Economics, 10(1), 74-91.
8. Coase, R. H. (1960). The problem of social cost. The Journal of Law and Economics, 3, 1-44.
9. Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775-779.
10. Committee on Payment and Settlement Systems. (1996). Settlement Risk in Foreign Exchange Transactions. Bank for International Settlements.（一九七四年那次银行关闭所暴露的跨时区敞口的标准分析）

人机分工声明

本文由王德生提出研究方向、承重判断与最终发表责任；Claude 完成三个领域的选源与冲突定位、公开文献检索、命题成形、二维表构造、逐领域兑现、分界与证伪设计以及正文写作。第七节第十一、十二两处收窄是在写作过程中被材料逼出来的，两处都削弱了初始版本更强的断言，全过程保留在正文中。文中涉及的预印本已在文献表内标明状态，正式引用前应重新核对版本。本文未标注任何评分——按本站惯例，参与改写者不为自己改写的文本打分。

证据边界 三家的转述均依据公开出版物与现行法律文本的核心结论，外链可自行核对；核指挥一侧只使用公开学术与政策文献，不涉及任何国家的现行操作细节。第十四节六条判据本文**均未执行**，已在正文中明写；其中第一条与第二条所需的数据在多数机构的既有日志里已经存在。

字数 纯汉字约 1.8 万 **版本** v1.0 · 2026-08-01

本文由三个分属不同学科、且互相冲突的理论体系碰撞而成。 全部来源均为站外公开文献，可自行核对。 作者：王德生 + Claude · SDE Universes · 学科通融