

观点与分析

编辑日期: 2026年10月31日

迟到的签名：机器作出的决定，究竟属于谁？

> 2020 年夏天，一项考试成绩看起来像是对教育问题给出的数学答案。可是，当人们寻找是谁“决定”了这个分数时，找到的却是一连串人：每个人掌握一部分权力，没有人独自拥有完整图景。这不只是电脑算错的故事，而是机构拆分了决定，随后忘记重新汇集责任。

学生没有在试卷上写出的分数

2020 年 8 月 13 日，英格兰公布高级程度考试 A-level 成绩。这项考试会影响数以千计学生能否升入大学，但那个夏天并没有举行考试。新冠疫情促使英国政府于 3 月 18 日取消原定测试，随后建立替代机制：依据学校估分、校内学生排序与统计处理给出成绩，并尝试让总体结果分布与往年保持可比。[1][2]

学生看到的是一个没有半个字母余地的明确等级，抵达它的路径却被分散开来：教师提交估分并排列学生，监管机构设计标准化方法，考试委员会依正式要求发出结果，政府部门确定公共政策目标。问题并不是机器人坐在阅卷员的位置上；根本没有一张写完等待批改的试卷。国家试图为一件未发生的教育事件创造替代品。

仅四天后，即 8 月 17 日，规则改变了。在相关情形中，最终成绩改为学校估分与计算成绩两者中的较高者。考试监管机构 Ofqual 此后对一个特定样本作出分析：样本为英格兰 18 岁学生，包括 246,110 名学生的 457,420 项科目报考记录。计算成绩约有 59% 的报考记录与学校估分一致，39% 低于学校估分，约 2% 高于学校估分。这里的比例属于*科目报考记录*，不是学生人数，因此不能准确地说“39% 的学生被降分”。[3]

计算成绩不同于教师估分，不能证明教师一定正确，也不能证明电脑在每个案例中都对学生不公。但它证明了另一件事：这个系统能够发出对个人影响重大的判断，并在几天内由一项机构决定改变判断依据。那么，第一个决定属于谁？谁承担其后果？又是谁拥有在它到达学生之前阻止它的权力？

2020 年 9 月 2 日，Ofqual 主席 Roger Taylor 致信英国议会教育委员会，表示监管机构完全接受自己应承担的责任份额。他说明，在部长指示的框架内，政策选择是采用经过统计标准化的教师估分，而监管机构负责执行。他随后得出一个重要结论：仅仅改进算法，不足以让结果变得可接受；数据边界和整个流程的性质，也是问题的一部分。4

把这件事读成“算法丑闻”然后离开很容易，但这个名称奇怪地替人类卸下了责任。方程不会发出部长指示，不会选择把公平定义为两年成绩分布相似，也不会决定当个人结果偏离学生预期时，可接受的伤害有多大。这些都是先于计算发生的人类决定。

机器可以执行选择，却不能抹去选择的人类历史。

一个词里藏着四种权力

机构常说“决定已经作出”，仿佛这个动作只有一个清楚的主人。但机构决定并不是一个点，而是在不同权利之间移动：提议行动的权利、批准权衡标准的权利、确认结果的权利，以及停止或纠正结果的权利。

Eugene F. Fama 与 Michael C. Jensen 在 1983 年发表的《所有权与控制权的分离》（Separation of Ownership and Control）中，把决策管理（Decision Management）与决策控制（Decision Control）区分开来。前者包括提出选项和执行决定，后者包括批准选项和监督表现。他们讨论的是机构结构以及经理人与所有者的利益冲突，并非人工智能；但这一区分有助于看见现代界面用一个按钮展示结果时所掩盖的东西。5

设想一个虚构的银行案例：模型预测贷款申请人违约的概率。技术团队产生估计，风险团队选择划分批准与拒绝的阈值，管理层批准可接受损失规模，运营系统发出通知，客服人员解释结果，而持牌银行实体是与客户建立监管和合同关系的一方。拒绝决定究竟属于谁？

对于两个不同问题，可以有两个都正确的答案。外部法律责任主体由适用制度、合同关系和事实决定，不会因为购买供应商模型就自动消失。内部因果能力主体则可能是批准某个变量、设定阈值，或命令系统在没有复核时运行的人。两个描述不能相互替代；严肃调查需要同时找到它们。

层次	所回答的问题	必需权力	不能单独证明这项权力的事物
确定目标	系统被允许优化什么利益?	批准目标、约束和不可交换权利	购买“高精度模型”
构造手段	哪些数据、假设和成本把目标转成计算?	检验模型、替代方案与适用边界	向用户展示置信分数
使结果生效	建议何时成为产生影响的决定?	批准阈值、权限和自动化边界	以部门名义自动电子签名
纠正与救济	谁能停止、撤销或重新评价影响?	可执行权限、时间、资源与申诉路径	一个收件人无权改变任何事的异议邮箱

这不是某个国家新的法律责任清单，而是一张追踪责任的运营地图。失败发生在四种权力被拆开却没有明确连接时：审查模型准确性的人无权改变目标；签署目标的人不理解方程在做什么；面对客户的人只能朗读预先替他写好的理由。

书面权力的主人，未必拥有实际权力

Philippe Aghion 与 Jean Tirole 1997 年发表了《组织中的正式权力与实际权力》（Formal and Real Authority in Organizations）。与本文相关的思想既简单又深刻：拥有正式决定权，并不等于拥有真实选择替代方案的能力，尤其当下属掌握上级无法及时消化的信息与细节时。6

复杂模型会放大这种分离。风险经理可能有权批准政策，却不知道移除一个变量会如何影响数千个案例；数据科学家知道变量的影响，却无权判断从信贷或法律角度看，这种影响是否可以接受。会议上，“我们批准了模型”听起来很确定，它却可能掩盖三个不同动作：看过演示、理解风险、接受后果。完成第一项的人，不一定完成了第二或第三项。

委托—代理问题（Principal – Agent Problem）随后出现。董事会把执行委托给管理层，管理层把技术任务交给团队，团队又可能把系统托管或运行交给供应商。每一环都从自己的位置追求合理目标：按时交付、降低违约、缩短答复时间或提高批准率。但局部目标彼此一致，并不保证最终决定符合机构建立时要服务的利益。

在这条链上，正式权限的主人可能最后才知道自己的权限允许了什么。这未必源于阴谋或怠惰，而是知识、时间与影响工具分布不均的预期结果。机构如果不主动设计通道，把实际权力还给承担责任的人，“人工复核”就可能变成一枚在回头成本已经很高时才抵达的签名。

方程正确，问题却问错了

回到英格兰。让 2020 年成绩分布与往年保持稳定，并不是荒唐目标。如果直接采用未经校准的学校估分，不同学校的宽松程度可能不同，成绩可能膨胀，学生届次之间的比较也可能受影响。Ofqual 在记录中明确提出了这些担忧。另一方面，统计上统一分布，也会把学校之间的一部分差异传导到某个学生的

等级上，而他没有机会通过真实考试证明自己的水平。14

因此，模型可能与既定目标完全一致，而目标本身却不足以支撑一项可接受的决定。模型内部的计算正确与把计算用于判断个人是否正当之间的差别，不是处理器速度的差别，而是最初交给处理器的问题不同。

这件事也显露出技术团队内部很难看见的一种风险：设计者用结果分布是否规整衡量成功；政府用教育过程是否持续衡量成功；大学用申请人是否具备入学资格衡量成功；学生则用自己是否有公平机会展示所学衡量成功。它们不是同一件事的不同计量单位，而是可能冲突的利益与权利。单一目标函数不能自动调和它们；必须由政治或机构作出明确决定，排列优先级并说明牺牲边界。

危机之后，英国统计监管办公室（Office for Statistics Regulation）于 2021 年 3 月 2 日发布复盘。主要教训包括：区分委托模型并确定其目的的机构与模型开发者的责任；检验模型是否适合目的、能否获得公众接受；并确定异议的依据和时机。建议不是让国家停止使用模型，而是不要再把公共决定缩减成看似纯技术的练习。7

模型在贷款申请中究竟决定了什么？

仅为说明，设想一家银行使用模型预测还款概率，记作 $p(x)$，其中 x 是申请可用信息的集合。再设 B 为贷款偿还时的额外收益，L 为违约损失，C 为服务成本。在一个简化、风险中性的模型里，接受规则可写成：

$$p(x)B - (1 - p(x))L - C > 0.$$

这不是真实银行规则，也不是信贷建议。实际贷款更复杂，决定还受流动性、资本、监管、平等待遇、模型风险与客户权利约束。方程只揭示一件事：模型或许产生了预测，却未必确定 B、L、C 的数值，也没有决定机构是否接受以这种方式表达目标。即使把“预期利润”选为唯一目标，本身也是可以被质疑的决定。

“预测还款”这句话还隐藏着认识论问题。预测不是承诺；还款概率不是履约的道德能力，也未必等于依法取得融资的资格。如果模型使用了缺乏代表性的数据，或是在旧标准下形成的信用历史，数字就可能携带此前决定留下的影响。发现这些限制并不能免除我们作出决定，却会改变我们能对决定依据声称什么。

设想某客户的估计还款概率为 0.79，而经理选择阈值 0.80。数据分析师不能只对客户说 0.79 是“模型的命令”；有人作出过人类决定，让 0.79 与 0.80 之间的差别成为产生后果的边界。资本成本或风险预算变化时，阈值可能移动，模型和客户数据却都不变。机构有时只需改变政策旋钮，就制造出新的决定，随后又把它误归因于系统的“智能”。

这一区分对每一方都有意义：贷款申请人需要可以检查和纠正的理由；风险员工可能被要求解释一项自己未参与设计的选择；股东需要知道谁改变了风险偏好；监管者也不能只在模型登记册里找到一个名字就满足。

“我们不理解黑箱”为什么不能成为借口？

美国信贷规则提供了一个有用区别，但只在其司法范围内。现行公布版本的 B 条例（Regulation B）§ 1002.9 规定了向信贷申请人通知不利决定并说明具体理由的要求；条件和程序会随申请类型而异。官方解释还指出，“未达到评分系统的合格分数”一类表述本身不够，原因必须与实际使用的因素相关。8

美国消费者金融保护局（Consumer Financial Protection Bureau, CFPB）曾在 2022 年发布通函，强调算法复杂并不能成为不说明理由的借口。但该通函及 2023 年的一项相关通函，已于 2025 年 5 月 12 日撤回。因此，不能把它过去的解释立场说成该机构的现行政策。这里重要的是区分撤回文件与仍有效的规则文本，而不是对美国境外银行的义务作出法律结论。9

有人也许会说：我们有变量重要性归因等模型解释方法，所以知道拒绝原因。但解释计算预测，不一定就是解释机构决定。前者告诉我们，按某种分析方法，哪些输入对数字有贡献；后者还必须说明银行为什么选择这个数字作为标准、客户数据如何呈现，以及哪项政策让结果生效。近似解释方法还可能对同一模型给出不同解释，也不总代表每个变量与真实结果之间存在因果关系。

反过来，系统也可能公开全部计算理由，决定仍然不可接受。谁说拥有历史数据就有权把它用于所有目的？谁批准把某一类错误拒绝的全部代价交给客户？透明是一扇重要窗口，却不会自动把不正当决定变成正当决定。知道门为何关闭，并不等于拥有开门的钥匙。

机构为什么不该只雇一个“负责批准的人”

许多方案承诺在决策系统中加入人在回路（Human-in-the-Loop）审查。这在适当条件下是有用工具，但这个名称没有说明坐在回路中的人拥有什么权力。他看得到相关事实吗？知道模型边界吗？能够不受隐性惩罚地反对建议吗？有足够时间吗？在他完成评估前，结果会暂停生效吗？

机构添加一个写着“批准”的小框，并不会自动产生监督。面对数百项相似建议，员工可能更像一台盖章机，而不是判断者。这里的管理悖论表面温和，后果却昂贵：机构把人放进系统以证明人仍存在，随后又把他的工作设计成无法与系统意见不同。

美国国家标准与技术研究院（NIST）2023 年发布的人工智能风险管理框架（AI Risk Management Framework）因此用实际责任描述问题：GOVERN 2.1 要求记录角色和沟通线，GOVERN 2.3 则把有关系统开发、部署风险的决定责任交给高级领导层。该框架是自愿性框架，不是一部能够赋予人们原本不拥有之权力的法律。10

欧盟《人工智能法》（EU AI Act）第 14 条要求，高风险系统的人类监督应当能够依具体情境理解系统能力与边界、忽略或推翻其输出，并介入停止系统。但高风险系统义务分阶段适用；按欧盟委员会在 2026 年修订后的说明，部分附件三用途的条款自 2027 年 12 月 2 日起适用，嵌入受监管产品的系统条款自 2028 年 8 月 2 日起适用。不能把这段文字转给一家沙特银行，仿佛它现在就受这项欧洲义务约束，也不能假定每一款模型都适用每一条款。11

越过地理边界后，设计上的含义仍然成立：人类监督不是流程图上员工所在的位置，而是在反对尚有意义时，真正有能力提出反对。

沙特：从分配责任到检验责任

《沙特中央银行监管金融机构主要治理原则》于 2021 年 6 月 30 日发布，第二版更新于 2024 年 5 月 8 日。按央行规则记录，它目前有效；该文件为董事会、执行管理层、委员会、监督和风险管理的职责划分提供了既有监管基础。但它本身并不是“一部专门规定人工智能模型决定由谁签字的规则”。如果要把它应用于算法系统，就必须把原则转成可检验的权限与流程，同时不能虚构文本中并不存在的义务。12

一家沙特银行的董事会可以在拟议内部审查中询问“人工智能作出了多少比例的决定”。这个问题听起来很现代，却不完整。更准确的问题是：模型输出在哪个阶段变成改变权利或利益的影响？谁批准了这个阶段？获得授权的员工介入前，可能发生什么损失？客户要求纠正信息或质疑计算时，接收请求的人有权改变结果，还是只有权把它转发到另一个邮箱？

如果风险委员会批准一个模型，产品部门随后在没有适当批准的情况下改变接受阈值，统计准确性测试本身不会发现决策权（Decision Rights）的问题。如果外部供应商可以在缺乏适当合同和技术控制时更换生产模型版本，那么拒绝信上的银行名称也无法解释实际控制权去了哪里。如果系统从新互动中学习，其行为可能已经改变，合规文件夹里签署的却仍是旧版本；那张纸对一个已经不存在的世界完全放心。

国际清算银行金融稳定研究所（Financial Stability Institute, BIS）2024 年 12 月发布的研究，指出金融业使用人工智能时面临治理、模型管理与第三方依赖等挑战。它没有评价每一个项目或沙特机构，却说明银行与供应商之间的知识分割不是一个无足轻重的本地问题。13

可执行纠正权：一个建议概念，不是一张现成证书

本文提出可执行纠正权（Executable Correction Authority）这一说法，把它作为本文建议的分析工具，而非公认学术理论或现行监管标准。对于一项产生影响的决定，它要求三个条件同时存在：有明确人员承担复核任务；有足够信息理解决定依据与边界；在纠正仍有意义的时限内，拥有实际和法律上的权力去改变结果或暂停其生效。

无需发明新的神圣指数，就能把这个想法转成机构测试。机构选取一批已经实际发出的、有影响的决定，在安全环境中重演一项合理异议，并逐项询问：是否找到有权负责人？他能否不靠猜测重建主要理由？发现已证实错误时，能否在向受益人公布的时限内执行纠正？随后，计算同时通过三项条件的案例占全部测试案例的比例，并按决定类型和风险程度公布分布，而不是只给出会掩盖严重案例的平均数。

这项实验指标不能证明原决定正确，不会创造一项法律上撤回不可逆行为的权利，而且如果样本过于简单、时限过分宽松，还可能在表面上改善。它只揭示问责的一面：机构能否让纠正真正生效。本文的可检验假设是，与只有名义投诉通道的系统相比，尽早明确这项权力并定期测试的系统，错误被发现后仍持续生效的决定会更少。在把它当成科学结果以前，需要受控现场研究。

还有四个问题比“我们是否有人在回路中”更诚实：有权者看得到关键数据吗？他不仅能反对*结果*，也能反对*政策*吗？他能在难以撤回的影响执行前停止紧急决定吗？受益人无需理解模型团队内部语言，就能找到异议路径吗？英国信息专员办公室（ICO）与艾伦·图灵研究所（The Alan Turing Institute）的联合指南主张，机构应向受人工智能支持决定影响的人解释决定；这里的解释是机构关系的一部分，而不只是数学展示。14

最有力的反驳：人类不会因为是人就更公平

上述问题并不意味着每一份文件都必须回到员工桌上，由他凭经验和直觉决定。人会前后不一，也会疲劳；判断会受情绪、经验和工作压力影响。经过良好测试、使用一致标准的模型，可能减少某些差异，也可能发现人工审查漏掉的错误。若每一项事务都要人工复核，延迟也可能伤害等待服务的人。这是严肃论点，不是自动化的公关辩护。

2020 年成绩事件也不能证明所有统计决定都无效。统计监管办公室的报告明确承认，监管机构工作人员真诚追求准确结果，任务受到异常的时间与数据约束，而全面退回不用统计模型，可能错失真实公共利益。7 另一个问题是，如果人工决定没有记录理由，它本身也可能无法解释。

但这项反驳不是推翻论点，而是修正实施路径。问责不要求人重做每一步计算。监督可以按风险分配：对简单、可逆案例给予明确自动授权；审查敏感边界；超过授权范围时强制暂停；由独立程序抽查决定，包括员工未作修改就批准的案例。这不是适用于所有行业的一套配方，而是让权力与决定影响相称的方法。

英国信息专员办公室关于人工智能治理的指南，还要求明确谁决定数据处理的目的和方式，谁代表他人处理数据；不同阶段和关系可能带来不同法律身份。如果实际控制与法律权限不同，不能只凭合同把供应商称为“处理者”，就把所有责任压缩进去。这是一项在其范围内适用的数据保护规则，不是对全部人工智能输出的民事或银行责任作出的总括判决。15

因此，关键测试不在于谁把手指放在批准按钮上，而在于谁能说“这项规则不可接受”，并让异议产生可执行的机构结果。没有这种能力，再多签名也救不了决策体系。最坏情况下，签名只是权力已经被拆分以后，重新分配指责的一场仪式。

在下一个结果到来之前

英格兰以新的依据重新发放了那个夏天的成绩。8 月 13 日之后，已无法把学生重新送回考场；重新计算也无法返还一些人承受第一次结果影响的那些日子。因此，政策逆转既重要又有限。决定不只是数据库中的数字；它在特定时刻进入人的生活，即使后来修改，影响仍可能继续前行。

Ofqual 经历最深的意义，不是一句简单的“提防不公算法”。它显示，一项判断在纸面上可以极其清楚，在权限地图上却极其模糊。当模型能够撰写建议、识别风险并发送决定时，工作流停下来明确征求人类意见的时刻会越来越少。届时，董事长不该只问“哪个模型作出了决定”，而应问：“我的哪项决定让这个模型有能力造成这种影响？如果事实证明我们错了，谁有权停止它？”

2020 年 8 月印出的成绩只有几个字母，每个字母背后却有一条很长的决定链。如果现代机构从中学到什么，它就不会满足于在链条末端找到一个签名，而会去寻找一开始是谁握着笔。

参考文献

1. Awarding GCSE, AS & A levels in summer 2020: interim report，Ofqual，2020 年 8 月 13 日；原创初步监管报告。说明成绩机制、背景和估分边界；细节适用于英格兰。
2. Requirements for the calculation of results in summer 2020，Ofqual，2020 年 7 月 7 日发布，修订至 8 月 20 日；面向考试委员会的原始要求。页面包含政策改变后形成的版本，解读时需注意。
3. Grading gaps in summer 2020: who was affected by differences between centre assessment grades and calculated grades?，Ofqual；对 2020 年结果的后续官方分析。包括特定年龄和范围内 246,110 名学生的 457,420 项科目报考记录，不测量每项个别估分是否真实，也不覆盖英国全部学生。
4. Written statement from Chair of Ofqual to the Education Select Committee，Ofqual 主席 Roger Taylor，2020 年 9 月 2 日；一手声明，界定部长指示、监管机构责任，并陈述其辩护与自我批评。
5. Separation of Ownership and Control，Eugene F. Fama、Michael C. Jensen，《The Journal of Law and Economics》，1983 年 6 月 1 日；区分决策管理与决策控制的奠基性学术研究，不是人工智能系统的实证研究。
6. Formal and Real Authority in Organizations，Philippe Aghion、Jean Tirole，《Journal of Political Economy》，1997 年；有关知识分布如何让正式权力与实际影响不同的理论研究。
7. Ensuring statistical models command public confidence，Office for Statistics Regulation，2021 年 3 月 2 日；对英国各地区 2020 年成绩系统教训相对独立的官方复盘，并非法院判决。
8. 12 CFR § 1002.9: Notifications，美国消费者金融保护局；本文于 2026 年 10 月 31 日核验的美国现行规则版本。规定不利信贷决定理由通知要求，只适用于其管辖范围。
9. Consumer Financial Protection Circular 2022-03，2022 年 5 月至 6 月间发布的历史解释声明；监管机构的撤回指南记录记载，它与 2023-03 通函于 2025 年 5 月 12 日撤回。本文不把二者描述为现行指南。
10. AI Risk Management Framework 1.0: Core, GOVERN，美国国家标准与技术研究院（NIST），2023 年；有关领导角色与模型风险管理的自愿框架。
11. Article 14: Human Oversight，欧盟委员会，按 2026 年 7 月 27 日综合文本；实施时间表见 2026 年官方指南 AI Act: Regulatory framework。需区分规则通过时间与各阶段开始适用时间。
12. Key Principles of Governance in Financial Institutions Subject to Saudi Central Bank's Oversight and Supervision，沙特中央银行，2021 年 6 月 30 日发布，第二版为 2024 年 5 月 8 日；规则记录标为“有效”。这是金融治理一手监管来源，不是专门规定人工智能决定签署的文本。
13. Regulating AI in the financial sector: recent developments and main challenges，Financial Stability Institute / BIS，2024 年 12 月 12 日；有关监管和技术风险的机构比较研究，不测量特定沙特银行。
14. Explaining decisions made with AI，ICO 与 The Alan Turing Institute，经 GOV.UK 发布，2025 年 1 月 27 日；面向受影响者解释决定的实践指南，不是有关解释有效性的现场实验。
15. What are the accountability and governance implications of AI?，英国信息专员办公室（ICO）；截至本文审阅日可用的官方指南。解释数据保护中的控制者与处理者角色，不涵盖所有形式的法律责任。