智能决策支持：AI 预测辅助战略与经营决策

作者：**SVSAC 硅谷战略顾问委员会**

核心判断

预测辅助决策的核心，是让企业在不确定性中更有章法地分配资源，而不是让模型替管理层作承诺。对企业而言，值得优先推进的是：为预测结果配套行动阈值、责任人和复盘周期，把误判成本与可执行能力纳入战略及经营讨论。本篇结合公开资料与分析，讨论企业可以如何识别机会、组织能力并安排下一步投入。

SVSAC - sv-sac.org

关键词：人工智能；概率预测；决策价值；选择性标签；人机协同；经营策略

SVSAC - sv-sac.org

SVSAC | sv-sac.org

预测辅助决策的核心，是让企业在不确定性中更有章法地分配资源，而不是让模型替管理层作承诺。

我们的建议是：为预测结果配套行动阈值、责任人和复盘周期，把误判成本与可执行能力纳入战略及经营讨论。企业可以先据此梳理自己的产品、客户、能力与投入边界，把讨论推进到可以分工、可以验证的具体任务。

对管理层来说，这个议题值得放进经营议程。围绕“先做什么、凭什么做、做到哪一步再扩大投入”形成共识，往往比同时铺开多个方向更有推进价值。下文保留相关资料、比较和推演，供企业判断自身条件与行动顺序。

1. 核心问题：预测改进何时成为经营改进

1.1 预测准确率不是决策结果

企业在产品上市、资本投入、渠道扩张、供应商切换或重大项目报价前，常会预测需求、失败概率、交付风险或客户采用率。预测系统可以更快生成数字、更频繁刷新分数，甚至在一组历史样本上降低平均误差；这些事实本身仍没有回答组织是否因此选择了不同项目、改变了资源投入、避免了可归因的损失，或提升了风险调整后的结果。本篇的核心问题是：在什么信息时点、损失结构、资源约束和权限安排下，AI 提供的预测信息能够改变可执行行动，并使相对于现实可行基线的净价值提高？

这里的“改善”必须指定决策主体和结果单位。董事会的项目组合、业务部门的年度预算、销售团队的重点账户和单笔报价并非同一个决策层级；利润、现金流、风险暴露、服务水平、员工工作量和长期能力也不能在未说明权重时合并成单一绩效分。若没有明确价值函数，所谓决策质量只能指某个被选择的指标，而不能默认等于企业整体利益。

1.2 分析边界与研究方法

我们关注预测辅助经营选择，不将需求预测、生产排程、自动化流程或员工技能变化作为主要因变量；这些问题可与本篇相连，但估计目标不同。问题覆盖战略与经营决策中相同的逻辑链：在行动尚未发生时形成概率判断，将概率与成本、收益、约束和可替代方案结合，选择或推迟某项行动，再观察行动实施后的成熟结果。预测、选择、执行、合规/采纳和结果记录分别描述。

方法上，本篇采取范围限定的原始研究梳理和概念–识别分析，而非系统综述或企业案例研究。材料分为实际取得原段/摘要的依据与未取得原文的文献线索；后者不承载具体方法和结果归属。范围涉及概率预测评分方法、人机协同实

SVSAC | sv-sac.org

验元分析、团队设计框架、选择性结果可见的决策模拟、未来事件预测的固定版本基准及其排名方法、预测竞赛与算法采用实验、库存预测模拟，以及投资研究现场试验。对每项证据均说明其版本、引用范围和不能支持的结论；例如刑事司法中的缺失结果不能直接推成企业投资收益，预测基准得分也不能代替市场或经营行动的实验结果。

数值例子均为作者构造，用来显示假设变化会如何改变决定或估计对象，不是企业数据、市场统计或推荐阈值。本篇不掌握企业 CRM、项目投资数据库、财务台账、访谈或干预分派，因此不会估计现实中的 AI 部署效应，也不以方法框架冒充实证结论。可检验性体现在预测信息集、决策人、比较政策、结果窗口和成本口径可被明确记录，并允许结果反驳预期机制。

1.3 本篇的增量：把预测价值拆成一组可检验条件

要验证“预测分数提升”通过改变行动带来“决策结果端价值上升”，至少须核对四个逻辑步骤。第一，评估分数须对应决策时真实可用的信息与目标事件。第二，预测变化须足以跨过某个有依据的行动边界，或改善资源排序。第三，组织必须有权、有容量且愿意执行改变后的行动。第四，行动后结果及其成本必须在适当期限和原始决策总体上被核实。若其中任一环节断开，预测可以在统计意义上更好，却不改变行动；行动也可能改变，但结果因成本转移、延迟、失访或选择性记录而没有改善。这条链限定于改变行动的结果端收益，不是所有经营价值的必要条件；即使业务行动不变，信息生产成本的净节省仍须另列工具、复核、更新与集成费用后计量，不能仅凭行动未变便断言总净收益为零。

因此，文章的主要分析对象不是抽象的“AI 能力”，而是一个带约束的策略：给定一项决策机会的历史信息，谁在什么时点看到何种预测，系统与人如何形成建议，组织允许执行什么动作，最后如何对未执行和未成熟结果计账。这个拆解既允许模型单独使用，也允许人机共同工作，还可纳入既有流程作为基线；若部署包同时改变信息、培训和权限，估计到的便是整套政策而非模型单因素。

2. 先分清预测、规则、行动和收益

2.1 四个对象不能用一个准确率代替

设某项战略机会在决策时点具有可记录的信息历史 H_t，目标事件为 Y，其行动政策与观察窗须另行固定；真实但未知的条件概率记为 $q(H_t)=P(Y=1|H_t)$，模型输出 $\hat{p}=f(H_t)$是对该概率的估计，不预设二者相等或模型已校准。历史政策下的条件概率也不能直接当作新行动的反事实概率。行动规则 π 再依据预测、收益/损失

SVSAC | sv-sac.org

及资源限制，选择行动 A；实际结果还取决于所选行动、执行质量、竞争者和环境变化。于是至少要区分四个对象：预测对实际结果的概率准确性、规则把预测映射成行动的方式、组织把行动落地的概率，以及给定动作后实现的净结果。用模型准确率表示整个链条，会把校准问题、授权问题和因果问题混在一起。

Y 也不必只有一个定义。产品是否上市、首年收入是否越过阈值、项目是否按期完成、投资是否超过预算、三年现金流是否为正，可能对应不同的事件和时间窗。把“成功”写成未定义标签，会让模型的数值看似精确，却无法判断该事件是否与真实决策损失一致。一个短期采用标签不必然代表长期客户价值；反过来，长期目标也不意味着可以忽略眼前的资金约束。

2.2 同一个概率可支持相反行动

考虑一个完全假设的项目：预测项目成功概率为 0.60。若成功带来 8 个单位的增量价值、失败损失 12 个单位，且不行动的增量价值定义为 0，行动的期望增量值是 0.60×8−0.40×12=0。在严格风险中性且没有额外成本时，行动与不行动持平。若失败损失改为 32，其他条件不变，则期望值为 0.60×8−0.40×32=−8，不应行动；若失败损失只有 4，则期望值为 0.60×8−0.40×4=3.2，行动有正增量值。预测概率没有变化，决定却随损失结构改变。

在二元结果、一次性行动、线性价值并且概率代表决策时点条件概率的简化模型中，成功收益为 B>0、失败损失为 L>0 时，行动阈值满足 pB-(1-p)L>0，即 p>L/(B+L)。这是一条从构造假设推出的算术关系，不是企业统一阈值。固定启动成本、渐进投入、多个备选动作、风险厌恶、资金约束、跨项目相关性和不可逆性都会改变规则；若把这些都压成一个阈值，阈值只是遗漏假设的总称。

2.3 预测问题与决策问题的目标变量可能不同

经营预测常是对“若不改变既有政策会怎样”的描述，而决策者真正需要的是“采取哪项政策会使结果怎样变化”。若投资本身提高项目成功率，基线概率 $P(Y=1|\ H_t)$ 并不等同于行动效果 $E[Y(1)-Y(0)|\ H_t]$。同一对象的历史成功率较高，未必说明增加投入仍有较大边际回报；也可能是过去只批准最有希望的项目，模型主要重现旧的筛选过程。

相反，某项干预的平均效果不高，也不自动意味着用预测进行对象排序没有价值。预测模型可能帮助把有限资金分配给预期边际收益更高的机会，但必须明确预测的是未干预风险、行动响应差异，还是给定政策组合后的结果。它们需要不同标签和设计。将“谁更可能成功”误当成“谁从额外投入中获益最多”，会把风险分数错用为响应分数；具体识别条件和反例将在后文展开。

SVSAC | sv-sac.org

2.4 决策政策的结果应以原始机会总体为分母

如果研究只统计获批项目、已启动产品或被业务员接受的建议，分母已经受决策本身影响。一个新预测流程可能让原先会被拒绝的边缘项目进入组合，因而组合内成功率下降，却同时增加总贡献；也可能只让原先就会获批的项目更早确认，组合成功率升高但总价值几乎不变。只看执行项目的平均表现，无法判断哪种情况发生。

更清楚的政策估计对象，是在决策前界定的全部合格机会中，按原始分派或政策版本统计的结果，包括未批准、延期、撤回、未执行、失败和结果未成熟者。执行率、建议采纳率和条件成功率可作为机制指标，但不能取代原始分母的政策结果。若最终结果存在长成熟期，固定观察时点只是统一截点，不保证各机会成熟程度相同；成熟差异本身可能就是政策效果的一部分。

3. 决策时点定义预测对象

3.1 先冻结信息集，再比较模型

若预测供年度投资委员会在 10 月 1 日作预算选择，评估信息集只能包含该日之前实际可获得的数据版本。后续修订的收入、确认的订单、投资批准或市场反馈，即使后来被并入“历史数据库”，也不能悄悄用于重建当时的预测。至少需保存数据提取时点、业务事件发生时点、字段修改时间、模型与提示模板版本、检索来源、人工输入，以及建议输出到决策者可见的时间。否则回测中的高分可能来自对未来状态的间接读取。

决策数据通常有多种时钟。信息何时生成、何时进入组织、谁何时看到、系统何时落库，不是可互换时间。若一份分析在项目已被非正式批准后才上传，虽然正式审批日期仍在未来，它却不是决策前信息。反过来，外部市场数据虽早已公开，但若组织在当时没有订阅、无法获得或没有权限使用，事后重建也不能自动假定管理者当时已具备这项信息。预测评估应模拟真实可用信息集，而非理论上任何人都能取得的信息集合。

3.2 预先规定结果成熟与标签规则

战略结果往往比一次性经营事件更慢成熟。某产品在首月未达到销售门槛，可能是失败，也可能是尚未进入主要渠道；某项目在一年内没有现金回收，不能据此判定整个投资生命周期的净现值为负。研究者应在观察预测分数前规定目标事件、业务总体、期限、来源记录、修订规则和“未决”状态。还应报告成熟期限

SVSAC | sv-sac.org

的分布以及新政策是否改变了结果记录速度，否则短期比较会把不同程度的未完成误作同一标签。

如果标签后来被重述，应同时保留决策时可见版本与最终核定版本，并回答不同标签分别评估什么。前者可评估当时系统是否在现行信息下给出合理判断；后者可评估依据最终事实回看预测的能力。两者不能通过覆盖旧记录合并成一个数。对于一个预测输出后再多次更新、决策者每次又得到不同建议的系统，评价单位还应标明预测轮次以及实际动作对应哪一轮。

3.3 预测动作后会改变自己要预测的环境

一旦建议影响投资、价格、营销支出、招聘或资源配置，预测对象不再是被动等待的外生事件。企业增加推广后，销量可能上升；预测为高风险的供应商被加强审查后，违约可能下降；一项未获批准的项目则不会产生和已投资项目相同的可见结果。因而，模型上线后实际观察到的标签是原有风险与新行动共同生成的。把这类行动后数据直接当作原预测目标的自然样本，可能让团队误以为风险变小是模型更准，或误以为新策略无效是预测失准。

操作上至少保留两个时间对象：行动前可预测的状态和行动后可观察的结果；另保留政策分配、采纳与执行记录。若目标是描述现行政策下的结果，应按该政策和信息集定义目标；若目标是选择行动，则需估计不同动作下潜在结果的差异，或设计可识别的比较。前者可以做预测验证，后者才涉及行动响应。二者常可共同建模，却不能把一个单一风险分数自动解释为干预收益。

4. 概率评分回答什么，又不回答什么

4.1 适当评分帮助评估概率，但不替决策者定义价值

这里用适当性定义评价概率报告：真实事件概率为 q，报告为 p，若 Brier 损失定义为$(p-Y)^2$，则期望为$(p-q)^2+q(1-q)$，在 p=q 唯一最小。对数损失也须明确有效概率域与预期损失口径，不能把概率直接压成分类。Gneiting 与 Raftery 文献保留为评分理论线索，[1]其正文未取得；上述 Brier 推导是本篇可复算的说明，不冒充原文逐段核验。期望评分性质本身既不保证某组织的实际校准，也不证明依此行动增加净值。

评分还依赖结果定义和损失方向。罕见重大失败、常见小幅偏差、不同预测时距及不同项目类型，可能被一个未经加权的平均分以不透明方式混合。适当评分的选用须与概率对象和观察到的结果相配；后续动作的假阳性/假阴性成本、资

SVSAC | sv-sac.org

金约束、延迟价值和不可逆损害仍需写入独立的决策函数。评分规则可以选择适合科学问题的损失形式，但不会替具体企业提供这些输入[1]。

4.2 校准、排序和阈值决策是不同性能问题

校准问的是：模型给出约 0.7 的机会，在足够多可比的成熟事件中，是否约有七成发生？排序能力问的是：实际更可能发生的对象是否通常得到更高分？阈值性能问的是：按当前损失和约束采取的动作，相比可行备选是否更有价值？某模型可以排序较好但概率系统性偏高，或整体校准尚可却在行动阈值附近不稳定；也可能对多数常规事件做得好，却错判少数资本额很大的项目。报告其中一项不能代表另两项。

比较概率预测时还须固定目标对象及任务权重。同一模型在本年度已有项目上表现好，并不意味着在新行业、不同决策人或更长时距上仍校准；反过来，整体误差变化也可能只反映问题组合或基率变化。应按事前定义的项目族、时期、损失等级或决策单元分层，同时显示分层样本量和未成熟结果。若有许多预测任务共享同一宏观冲击，它们也不等于同样多的独立信息单位。

4.3 任务集不重叠会扭曲模型排名

ForecastBench 尝试用未来事件问题减少已知答案泄漏，并在固定版本中比较人类与语言模型的概率预测[5]。其 v5 摘要报告专家预测员在该版比较中优于顶级模型；正文将人机比较限定在 200 道向人类提出的标准问题，并注明动态网站上的当前得分可能随后更新[5]。这是一项预测基准结果，而不是企业投资、预算、招聘或市场进入的决策试验。未来事件题可减少某类答案泄漏，却无法自动确保题目与组织决策相同、信息获取成本相同或预测结果能通过行动改变。

该基准后续技术说明指出，当不同预测者回答的题集不重叠时，直接比较平均 Brier 可能混入题目难度差异；其难度调整方法依赖足够的间接题目重叠与题目/预测者效应分解假设[6]。这对业务评估的启发是，跨年度、跨业务线或跨模型版本比较前，应检查问题集合、时间范围和信息支持是否可比。基准排名即使通过统计调整改善，也只说明在模型设定与假设内可比较预测分数，不代表由此产生的经济价值或实际执行能力可比。

SVSAC | sv-sac.org

5. 人与 AI 共同决策需要至少三个基线

5.1 “超过人类”不是充分的人机协同标准

若 AI 建议后的决策比人单独判断好，说明可能存在人类增强；若组合同时优于人单独和 AI 单独，才符合更严格的互补/协同基线。举例而言，若人单独准确率为 70%、AI 单独 80%、人看到 AI 建议后为 75%，组合优于人类但仍输给 AI；只与人基线比较会称之为增强，却不能称为超过现有最佳单方。对经营结果也相同：需把团队政策与人类现行流程、机器策略以及可以执行的最佳单方政策相比较，而非挑选较弱的基线。

基线选择还要考虑现实可行性。若纯 AI 策略不被授权或无法处理例外，机器单独可能只是实验性上界，不是现实政策；若人单独评估所用数据、时间和权限不同于人机组，比较也被流程差异污染。应同时展示现实现行政策和可诊断的组件基线，说明哪项决策问题由哪个比较回答。对照设计不应把“人+AI”视为单一不变处理：提示、解释、置信展示、建议默认、采纳权和人工复核都会构成不同政策包。

5.2 已有实验汇总提示协同并非默认

Vaccaro、Almaatouq 与 Malone 的系统综述与元分析，纳入 2020 年 1 月至 2023 年 6 月发表的原始实验；纳入标准要求同时报告人单独、AI 单独和人机合作的量化表现。该文汇总 106 项实验研究、370 个效应量，报告组合相对两者中较佳单方的平均效应为负（Hedges’ g=−0.23，95%区间−0.39 至−0.07），而相对于人单独的增强结果可为正[2]。这一区分很关键：把单一的“人机优于人”结果表述成“人机总是互补”，会抹掉更严格的 AI 单方比较。

该综述的任务类型元回归中，决策任务与平均损失相关，内容创作类任务与较大增益相关；多数实验中最终决策仍由人作出，作者提出分工可能比让双方各自完成完整任务更有效[2]。然而任务类型是跨研究的调节变量，不能当作随机分配的任务机制；纳入研究截至 2023 年中，也不能代表此后所有生成式系统、长期管理团队或高层战略决策。综述能够反驳“组合优势自然发生”的普遍预设，却不能判定具体企业应选择 AI 独立决策或某种固定分工。

5.3 概念框架可指导假设，不是效果估计

Gonzalez 等的人机决策团队文献保留为互补性框架的待核线索。[3] 本篇未取得其相关正文，不把所列五项组织因素或分类当作该文已经核实的构成。我们提出的分析问题是：分工改变结果时，作用来自模型信息、人的复核、共享上下

SVSAC | sv-sac.org

文、培训还是责任安排？这些因素须转为分别可观察或操纵的变量，而不是用框架名称代替机制检验。

管理决策的责任结构也不能由“人在回路”四个字代替。需要指出谁可以否决、谁承担延误成本、建议被拒绝如何记录、接受建议后谁批准执行，以及事后复盘会不会使决策人承担与 AI 建议不同的声誉风险。若人对错误建议负责却无法看到关键依据，表面的人工签字并不等于有效审查；若模型建议不断推迟审批，延误也可能成为未被记账的成本。上述均是组织机制假设，而非既有框架已实证验证的固定效应。

5.4 选择性观察使历史准确率不等于政策收益

Kleinberg 等的工作论文保留为决策与预测研究的待核文献。[4] 本篇未取得其识别策略对应原段，不再以法官分配或反事实插补的具体做法作为已核方法依据。本篇独立讨论的条件问题是：如果行动决定结果是否出现，历史可见标签便不足以直接评价所有未执行政策。下文用明确构造而非司法收益数字说明这一点，不将文献题名或旧摘要记录当作全文核验。

企业的例子更直接：只观察获批投资的收入，无法知道被拒绝项目若获批会怎样；只观察已上市产品的退市率，也无法直接评价未上市方案。把未行动的结果简单填为失败、成功或总体平均，都增加了未经验证的假设。要比较决策政策，需说明哪些结果对哪些行动可见，是否存在外生/随机变化来支持反事实，何时采用部分识别界限而非点估计，以及政策分配本身是否改变了观测。细化的选择标签识别与经营支付对象将在后续章节中展开。

6. 选择性标签：未采取行动的结果不等于零

6.1 哪些业务结果会因决策而不可见

战略选择常改变标签能否出现。未获批的投资不会产生真实的投后现金流；未上市产品没有市场销量；被拒绝的报价无法显示若接受后能否按期交付；未发出的贷款或保险产品也不产生同类违约记录。这里的缺失不是普通随机漏录，而是由行动路径决定的结构性不可见。反过来，已批准项目才有完整成本台账，已上市产品才有退货数据，都会让“可见结果样本”偏向历史政策已经选中的对象。把未发生的结果按失败填充，或只用执行项目训练模型，都会额外写入无法由数据本身验证的假设。

同一项目可能同时出现多种缺失。未批准会让投资回报不可见，但决策前已有的市场评估仍可观察；项目获批但因供应中断暂停，可能使期限内结果未成

SVSAC | sv-sac.org

熟；执行后失联或财务系统未对账又是数据采集问题。分析中应分别标记结构性反事实缺失、行政截尾、业务状态未决和记录缺失，而不能用一个“NA”混在一起。只有后两类中部分原因可能通过追补数据解决，第一类需要额外的识别设计或界限。

对企业机会，若项目未获投入，执行该项目的潜在结果便不可直接观察；若投入还改变市场或团队反应，即使历史上有同类对象，其结果也未必代表当前动作。这里陈述的是本分析的识别条件，不是对尚未核到原段的司法研究的具体转述。第 7.4 节在固定价值与透明上下界下给出非识别例子；改变这些条件就需重新推导。

6.2 风险高并不意味着追加资源的边际价值高

设两个项目未干预成功概率分别为 0.8 和 0.4，追加资源分别将其提高 0.01 和 0.20。假定每次成功价值为 100，失败损失已含在基线中，追加成本各为 8，则期望增益为 1−8=−7 与 20−8=12。按原成功概率优先会选第一个而非响应更大的第二个；按失败风险优先则恰好选第二个，故此例不能反驳风险排序。若维持原成功概率但交换两项响应，风险优先又会错过正增益项目。两种构造共同说明：基准风险不识别响应，排序是否一致取决于待估的行动差，而非“高风险”标签本身。

再考虑可行动性。一个预计发生概率很高、但无法通过本组织可用行动避免的供应商退出，可能只值得制定应急计划；一个概率中等、可通过短期替代认证显著降低损失的供应链风险，可能更值得先做。正确的增量决策值至少是对每种动作估计“采取该动作相对不采取会改变多少结果”，再乘以对应价值并扣除动作成本。单一风险模型即使校准，也不给出这个因果响应。若拿历史上执行干预的人与未执行者直接比较，执行者可能恰好更危险或更有资源，响应差异会与选择混杂。

6.3 预测准确性和行动价值的标签不同

用于预测的标签可定义为“下一年度是否越过收入门槛”；行动目标却可能是“增加一名工程师后，相对于不增加，项目按期交付概率改变多少”。前者可以对所有已上市项目观察，后者对每个项目只能看到实际采取的一种资源决策。若模型以第一种标签训练，再用高预测概率将项目排在前面，它估计的是条件结果风险，而非项目对新增资源的响应。决策者若仍用风险排序，必须明确其隐含假设，例如边际效果与风险单调相关；没有证据时，这只是策略假设。

评价标签还要考虑观察期和竞争事件。若投资被取消，可能只是资金延期，仍可能在下季度批准；只有既定项目机会确实关闭、被竞争者拿走或发生不可逆

SVSAC | sv-sac.org

终止时，才可把该机会标为不再到达原决策终点。延期不是天然失败，也不是自动删失。固定时间点内尚未成熟的结果应保留“未决”状态，并报告比例与成熟时间；把未决项目删掉，可能按政策选择了不同的成熟子样本。

7. 从可见样本到政策价值：识别范围与界限

7.1 先写明要评价的政策

一项可检验政策不只是预测模型，而是完整映射：对哪些合格机会开放、提供何种信息、谁看到建议、可以采取哪些动作、采用或拒绝如何处理、资源何时可用，以及在何期限核算结果。比较“旧评分+委员会规则”与“新预测+原委员会规则”只识别预测信息替换的政策包；若同时改变预算门槛、培训、权限和审核，差异是整套新制度的效果，不能归因到模型单独贡献。

政策结果应以分配之前定义的机会为主分母。建议被采纳率、启动项目数、执行者中的成功率可以解释机制，却会受政策影响，不宜当作主结果。对每个政策版本，可以同时报告：全机会净价值、在已执行行动上的结果、延期/撤回/未决比例、资源耗用和风险尾部。若只能观察到已执行项目结果，就应将可识别目标缩小为该政策下的已观察对象描述，不能以此声称未选择政策也更差。

7.2 何时随机化可识别行动差异

若企业在同一时期有可比项目，且不存在项目之间明显干扰，可以在符合资格的项目层级随机分派“当前政策”或“预测辅助政策”，按原始分组估计意向处理差异。随机的是政策可用性，而不是项目最后是否获批；这样拒绝建议、主管覆核和执行失败仍留在被分配组内。若只有部分分支获得不同的预测，实际采纳率可作为中间结果，但按采纳者重新分组会破坏原随机化。随机化可能不适用于不可逆、高安全性或法定责任决策，此时应考虑分阶段上线、配对业务单元、历史自然实验，或诚实报告仅能做观察性推断。

还要界定随机化单位。若共享决策委员会、资金池、同一客户关系或供应商资源会使不同组互相影响，项目级随机化可能使组间政策互相污染。此时可在决策团队、业务线或时间区组层面分配，并承认独立单位减少。干扰本身可能是要测的机制：新建议占用分析师时数后，其他项目即使在控制组也变慢。单项目平均处理效应不会自动代表整套组织推广到饱和状态的政策效果。

SVSAC | sv-sac.org

7.3 观察性校正依赖未检验的假设

没有随机分派时，历史批准与拒绝项目的差异通常同时反映项目质量、决策者判断、预算紧张程度及组织优先级。用可观察变量匹配或加权，只能在相应条件假设下比较，不能排除未记录的委员会预期或非正式承诺。某项机构安排若提供近似外生变化，还须论证其分配与排除条件；本篇不把未核文献中的方法名称作为自动成立的“机器纠偏”步骤。

如果采用历史记录建模，需列出选择模型中的变量、何者在决定前已观测、何者可能被政策改变，以及在何种条件下行动概率接近零。某类项目若历史上从不批准，就没有该类项目执行后的结果支持；加权无法从零支持中创造信息。此时可报告支持域内的政策效果、用透明假设给部分界限，或直接拒绝外推到该类项目。拟合更灵活的模型不会消除没有反事实观察的问题。

7.4 构造性界限：缺失项目的决策建议可能翻转

以下数字仅用于说明不可识别性。假设四个同样合格的战略项目中，旧政策只执行两个，已实现净值分别为+20 和+10；其余两个未执行，结果没有观察。执行项目均值为 15，但不能据此断定四个合格项目各自执行时的潜在净值总体平均也是 15；后者是此例要界定的估计对象，而不是旧政策价值。此处把项目净值视为固定值，暂不处理时间随机性。若将未执行项目的假想值暂设为 0，四项目平均为 7.5；若其中一个本会产生+30，另一个为−10，平均仍为 12.5；若两个都为−20，平均变为−2.5。相同的可见数据与不同的执行潜在结果相容，全体项目各自执行净值的平均可从正变负。若另假设不执行的增量净值为 0 且项目间无相互作用，旧政策在这批已实现样本中的净值总额是 30，按四个原始合格项目归一为 7.5；这不能与全体项目的执行潜在净值均值混同，也不自动给出未来政策的期望收益。

若事先能给出每个未执行项目净值的可信范围，例如每项在−20 到+30 之间，则全体四个项目各自执行时的潜在净值平均有界于 ((20+10−20−20)/4=−2.5) 与 ((20+10+30+30)/4=22.5)。这个区间很宽，表示已知数据不足以决定全体项目执行潜在净值均值的符号；它不是置信区间，也不表示所有范围内值同等可能。若未来随机执行一部分边缘项目或获得可信的准随机分配，范围才可能缩窄。若边界本身来自管理层假设而非外部约束，应明确标为敏感性参数。这不是断言同时执行四项目在资源上可行；要比较真实政策，还须按各政策实际选择计入不行动结果、行动成本、预算及项目间影响。

SVSAC | sv-sac.org

8. 策略效应要与“已选中项目质量”区分

8.1 三个政策比较回答三个问题

第一种比较固定旧行动规则而非实际选择，只替换预测：同一规则仍可根据新信息改变选择，检验该信息替换的政策价值。若连实际选择也固定，只能评价预测质量或过程成本，不能估计改变行动的结果收益。第二种固定预测而替换行动规则，检验同一信息如何映射为不同动作；第三种联合替换预测、规则、权限和培训，检验新策略包相对现行流程的效果。第三种与部署直接相关，却不能单独归因于模型。

可用因子设计来分离部分组件，例如将新/旧预测和新/旧行动规则组成四个分支，并在事先统一的机会总体内随机分配。若结果是高价值项目被捕获率、净贡献或风险调整回报，预测主效应、规则主效应和交互作用需分别估计。交互作用可表示“新规则是否特别会利用新预测”，但若培训、审批权和人员结构也变了，交互差异仍不能自动解释为纯算法机制。小样本下每个分支的区间可能很宽，应将不确定性报告为证据不足，而不是用最佳单点组合做确定性排名。

8.2 条件成功率与政策价值可反向变化

构造一个预算分配例子：旧政策只选 2 个项目，每个成功时价值 100、失败时价值 0，成功概率分别为 0.9 和 0.8，投入成本各 20。期望净值是 0.9(100)-20+0.8(100)-20=130，执行项目成功率为 85%。新预测规则选入另外 8 个项目，其中成功概率均为 0.3，成本同为 20；新政策总期望净值增加 8(0.3(100)-20)=80，变为 210，虽然执行组合的加权成功率从 85%降到 41%。这个例子说明成功率和净值可能不同向，但依赖所设边际成功概率、价值、成本、容量及无非线性相互作用，期望相加不要求项目独立。共享冲击改变联合尾部与风险约束，须另算；只有同时改变边际概率或非线性支付时才改变这里的期望。若新项目挤掉旧项目或成本超过 30，也须重算对应账本。

另一个相反例子是，新规则只保留原先两项中最有希望的一个，账面成功率可达 90%，但另一项目仍可获得正净值且资源没有替代用途。若不把未选项目的放弃价值或可替代项目收益计入，较高条件成功率会制造“决策优化”的假象。只有对所有预先定义机会同时记录政策分配与结果，并说明预算约束如何影响同一组合内的选择，组合层净价值才有明确含义。

SVSAC | sv-sac.org

8.3 阈值不稳定时，预测变化未必值得执行

若行动阈值约为 0.67，而某项目原预测 0.66、新模型为 0.68，规则会发生翻转；但如果估计误差、基率漂移或模型版本变动带来的概率不确定性约为±0.10，基于小数点后第二位的确定决策并不稳健。应展示预测不确定性、损失参数的不确定性以及在合理范围内动作会否变化。对已投入大量不可回收资金的项目，停止/继续的损失结构又不同于尚未签约的新机会，不能套同一阈值。

相较于追求极细的分数，决策可考虑稳健区域：在一组合理的收益、失败成本和概率范围内，所有情景都建议行动、都建议不行动，或结论会翻转。若多数项目处在翻转区，优先补充信息或分阶段承诺可能比立即调整模型更有价值。若对概率或损失范围没有可信约束，所谓稳健区间同样只是形式；应把无法界定的参数明确列为待收集数据，而不是用模型的自信输出替代知识。

9. 信息价值受可选行动集和获取成本约束

9.1 完全信息价值只是上限，不是预测项目的预算

构造一个可复算的二元项目：若状态良好，投资后净值为+80；若状态不良，净值为−40；不投资净值为 0，良好状态先验概率为 0.4。当前无额外信息时，投资期望值为 0.4(80)+0.6(-40)=8，所以选择投资。若能在不改变项目状态的前提下免费获知真实状态，良好时投资、不良时放弃，期望值为 0.4(80)+0.6(0)=32。在这个简化设定中，完全信息的期望价值上限为 32-8=24。这不是购买预测系统可支付 24 单位的结论，因为真实信息不完全、取得时间有成本，且项目价值可能随等待变化。

9.2 信息只有跨越决策边界时才有直接决策价值

若一项新分析把项目概率从 0.60 更新至 0.62，而在所有合理损失情景下两者都仍高于行动阈值，且两种情况下资金配置完全相同，那么对这一行动选择而言，直接信息价值为零；它仍可能改善审计、风险披露或后续学习，但这些收益需另行定义。相反，若信息可能改变是否签署不可撤销合同，它的价值可能远超对平均预测误差的微小改善。决策价值取决于信息分布在阈值两侧的概率、动作价值差和执行约束，不等同于模型平均误差的变化。

9.3 等待、试验和行动都可能改变目标

经营者可选的不只是“立即行动/不行动”，还包括先做小规模试验、等待新信息、分阶段承诺或购买外部验证。对不可逆项目，等待会保留选择权，但也可能

SVSAC | sv-sac.org

错过窗口、让竞争者先行或增加报价成本。若试验本身改变市场、员工预期或供应商响应，它不再是纯粹信息采集；若试验资源挤占其他项目，其成本还包括被放弃的替代行动。计算信息价值时必须将“无试验而立即行动”“试验后行动”和“延后/放弃”等策略置于同一决策时点，列出每种路径的净收益、时间和执行条件，不能把事后最优路径当作事前可行策略。

10. 评分表现必须回到不对称损失

10.1 总体校准不保证决策分层校准

校准通常被口头概括为“预测 0.7 的项目约有七成实现目标”，但对战略决策而言，还需问这句话在哪类项目、哪个时段和哪种行动下成立。构造一个有 100 个机会的例子：两组各 50 个项目，系统都报成功概率 0.60；第一组实际 40 个成功，第二组 20 个成功。合并后 60/100 恰好对应总体平均 0.60，可称总体频率一致；但对第一组的 0.60 低估了真实 0.80，对第二组则高估了真实 0.40。若行动阈值是 0.70，仅看共同分数会把两组都排除；若目标只对第一组追加投资，分组校准或更直接的行动响应数据可能改变决策。整体可靠性图掩盖了决策相关异质性。

此例中的“组别”必须在决策前定义，不能用事后结果切片来制造可解释模式。可用项目类型、投资阶段、区域、时间、信息完整度和决策渠道形成有限分层，同时显示每层事件数及概率区间。层数过多会使估计极不稳定，层数过少会把重要的损失差异平均掉。分层不是为了强迫每组分别部署一个模型，而是判断现有总体误差是否能支持共同决策规则；必要时可把小样本组标为证据不足、沿用保护性政策或补充数据。

适当评分的总体期望性质不消除分组选择问题。按第 4.1 节的定义和 Brier 推导，适当评分可支持预期损失意义下的概率评估，但对企业某个行动分层是否校准，仍需用相应决策时点、目标事件和成熟标签检验[1]。若模型在历史政策下只有被批准项目的结果，所谓分层校准也仅是可见样本校准；被拒绝项目的未观察结果仍然不能由评分规则补出。

10.2 不对称成本会改变误判的经营意义

若把失败预警当作行动信号，假阳性可能造成额外尽调、暂停合作或机会错失；假阴性可能导致资本损失、交付风险或客户违约。两类错误的成本通常不等，且同一错误在不同决策层级上的后果不同。一个项目级误报或许只多占数小时分析师时间，组合级误报却可能触发资金冻结或管理层注意力重新分配。因

SVSAC | sv-sac.org

此，单一平均准确率或同等权重的误差表不能代表行动损失；需在行动前定义每类错误的实际后果，避免在看到结果后重设权重。

预测系统可选择适当的评分规则以评估概率，也可另行使用成本敏感的指标，但后者需要事前给出具体成本或可接受范围。把高风险预警的处理成本设为零、把误报项目的机会成本设为零，会系统性偏好更保守或更积极的行动。若没有可信的成本数值，可以报告一组敏感性情境，例如错误成本在某区间内时政策排序是否稳定；不能把任一自设参数包装成企业估值。

10.3 模型不确定性与商业损失不确定性是两层问题

一个概率估计的不确定性可以来自有限训练样本、模型参数、测量噪声或新时期漂移；行动损失的不确定性则来自项目价值、失败成本、资金占用和竞争反应。把预测区间解释成损失区间，或把模型置信度直接乘以管理层估算的收益，都会混淆层次。决策分析可以在一组概率和损失参数上计算行动是否稳健，但各参数的来源和相关性需要说明。

构造一个阈值例：行动收益 (B=12)，失败损失 (L=28)，则简化阈值为 0.70。若预测为 0.72 且估计误差区间为 0.62–0.81，行动建议会在区间内翻转；若失败损失本身可能在 20–40 之间，阈值又从约 0.625 变到约 0.769。两种不确定性叠加后，单个点预测无法决定是否行动。可行的结果不是强行选出一个“AI 答案”，而是识别哪些信息足以缩小不确定性、行动是否可分阶段，以及在误差范围内哪些决策后悔成本最大。

11. 资本与人力约束下的排序问题

11.1 概率排序与投资组合优化不同

如果组织只能为有限项目配置资本，按成功概率从高到低排序不必然最大化组合价值。所需对象是每个可选动作相对于不采取该动作的增量净值、成本、资源占用、风险相关性和互斥关系。一个低概率但成功回报很大的项目可能有正的期望增量值；一个高概率项目可能回报有限、成本巨大。另一方面，即使单项期望净值为正，也可能因资本锁定、风险集中或不可逆性而不适合加入当前组合。

决策变量可能是离散项目选择、分阶段投入、等待或退出，而不是对所有对象同一阈值二分。若项目不可分、启动后有固定成本、共享基础设施或相互排斥，逐个按“每单位资金期望收益”排序不一定得到全局最优。若条件概率还受到共同宏观状态影响，期望的线性可加性本身仍成立，却不能确定组合损失的尾部；若额外假定项目独立来组合分布，才可能低估共同失败下的集中损失。实际

SVSAC | sv-sac.org

分析应把组合约束写成决策的一部分，并展示备选策略，而不是把排序分数直接当最终项目清单。

11.2 可复算的非分割项目反例

设预算为 4 个单位，三个不可分项目的期望增量净值分别是 A=7（占 4 单位）、B=5（占 3 单位）、C=4（占 2 单位），不考虑项目间相关性。按单位成本收益比排序会先选 C，其比率 2；剩余预算 2，A 与 B 都放不下，总价值为 4。可行组合逐一枚举：空集 0，A 为 7，B 为 5，C 为 4，A+B 超预算，A+C 超预算，B+C 成本 5 超预算，三项成本 9 超预算。因此 A 是最优可行组合，虽然其单位成本比低于 C。这个例子并非模型预测会自动造成贪心算法，只说明“分数排名”与离散组合选择之间需要优化规则；若项目可连续缩放，结果可能不同。

组合价值还需处理资源机会成本。若分析师、现金或试验设备有空闲，新增项目占用这些资源不必对应全额工资或外部报价；若资源已满载，新增项目会挤出其他净值。机会成本应由实际可替代的边际任务或明确容量约束估计，而不是机械用平均费用乘工时。若不可见替代项目的收益，结果应报告为敏感性区间，而非声称 AI“释放了”同额利润。

11.3 共同风险会使单项优选损害组合韧性

两项项目分别有 70%和 80%成功概率、单项期望净值都为正，并不足以判断应同时投入。若两者都依赖同一个监管许可、供应商、宏观需求或技术路线，失败可能高度相关；组合的尾部损失会大于独立假设下的计算。也存在相反机制：项目路径不同，某项失败时另一项可能补位，组合相关性较低。相关性不是预测模型的可选图表，而会影响资金是否应集中、分散或保留备用额度。

历史上未曾同时执行过的项目组合缺乏联合风险数据。将单项目分布相乘默认独立，或用各自平均成功率推导组合失败概率，都需要额外假设。战略决策可采用可复算压力测试：对共同冲击、单项目失效和替代需求分别设定透明情境，列明概率若无法识别则不要假装可精确加权。压力情境不是概率预测；它用于判断组合是否在特定冲击下违反现金、交付或风险限制。

12. 预测更新、反身性与稳定评估

12.1 新预测可能改变数据生成过程

部署后，管理者可能更频繁审查高风险对象，业务人员可能优先跟进高分项目，资源也会从低分机会移走。这样，记录到的项目结果由预测、动作、执行和

SVSAC | sv-sac.org

市场共同生成；模型不再只是观察既有数据分布。若干预成功地降低了某类失败，重新训练时会看到高分对象失败率下降；把这种下降一律解释为原风险模型过高估计，会忽略政策确实改变风险的可能。若组织只记录高分对象的跟进细节，模型还会在低分端失去同类信息。

评价更新模型时，应区分“预测在现行政策下是否校准”和“更新策略是否比现行策略创造更多价值”。前者可以对新政策产生的结果做描述性验证，但其标签分布已受行动影响；后者需要定义比较政策、保持可比机会总体并识别不同动作下的结果。滚动重训、人工规则变化和业务目标变化若同时发生，简单前后对比无法归因。保留模型版本与策略版本的对应关系，比只保存最新权重更重要。

12.2 冻结测试集避免被反复适配后的乐观评价

如果团队每周查看同一留出集表现、调阈值和提示词，再把最好一次分数报告为最终效果，留出集已经参与开发。样本未变并不代表评估独立。战略项目数据往往稀少且事件成熟慢，反复试验会让团队在偶然的好结果上选模型，并忽略未成熟项目后来回落。可采用按时间切分、按项目族外留、组织单元留出或预注册一次性封存终测；具体选择取决于部署要外推到未来时期还是新业务线。

封存终测并不自动解决部署效应问题。模型在未来预测集得分更好，只说明其预测能力在该测试定义下得到支持；若没有实际政策比较，仍不知业务人员是否采纳、行动是否可执行或结果是否改善。若测试集被严重冲击事件主导，单一时期也难以代表长期。预先指定误差区间、行动价值、主要结果和停止规则，可降低事后挑选指标的自由度，但不能消除外部有效性限制。

12.3 多版本试验会产生赢家偏差

若同时试验许多模型、提示模板、阈值与业务分层，再只保留表现最佳者，赢家通常包含真实增益与抽样噪声。被选中方案后续在同一数据上看似稳定，可能只是筛选造成的上偏。尤其对低频战略结果，一个极少数项目的偶然成功或失败就可能改变排序。应保存所有曾被比较的版本和试验轮次，并将模型选择数据与确认性评估数据分开；若做不到，应把报告标为探索性，等待新的时间段或独立业务单元验证。

不同模型并行也会占用组织注意力。试验组可能得到更资深分析师、额外管理复核或更长处理时间，导致“模型效果”与资源倾斜混合。控制组和实验组在可见信息、咨询权限、人员培训及排队优先级上要尽可能可比；否则估到的是运行政策包。若资源本身无法平均，需记录差异并报告它属于干预组成部分，而非事后用协变量调整就假定已剥离。

SVSAC | sv-sac.org

13. 人机分工应按可检验能力而非“人在回路”口号设计

13.1 拆分任务，说明人和模型各自改变什么

在人机协作中，最终决策由谁签字，不足以说明模型具体贡献。一个流程可能由模型筛选候选、分析师核查信息、领域负责人提出例外、委员会决定预算构成；另一个流程可能由人先选对象，再让模型补充预测。两者的处理机制、信息顺序和权责不同。若将它们都标记成“有人复核”，任何结果差异都难以解释。可先把完整任务拆成提取可核事实、形成概率、检查异常、生成备选行动、评估价值和批准执行，再说明模型在哪一环节给出建议、人员可作何种修改，以及修改的理由如何记录。

任务分工必须针对错误类型。模型可能擅长整合大量结构化变量，却对新政策冲击、极少见事件、关系承诺或不完整资料失效；资深人员可能识别特殊依赖，但对大量普通机会使用不一致标准。若把人的价值简单设为“补足 AI 短板”，忽视人类也可能在压力下忽略基准率、受近期案例影响或偏向可归责行动。合理对照不是假定任一方天然更强，而是预先定义哪些错误由谁检测、漏检代价是什么，以及事件发生后如何验证。

13.2 观察到算法犯错会改变采用，不代表算法能力改变

Dietvorst 等文献保留为算法规避的待核线索；本篇未取得原文，不以五项实验、MBA 数据或奖励绑定方式作为已核结果。[7] 本篇区分两种待检验对象：工具在固定任务上的预测能力，以及使用者看到错误后是否继续采用它。部署时应分别记录表现、可见错误与采纳；拒绝建议也可能来自新增信息、责任或不同损失函数，不能预先视为不理性。

对经营部署而言，采用率不能只作模型使用日志的副指标。可记录每条建议是否可见、是否被读到、是否被采纳或修改、否决理由、建议时延、执行权限和后续结果；同时用盲态抽样核验理由记录是否可信。若可行，可在同等预测信息下随机显示不同的解释/不确定区间/人类复核方式，观察采纳、决策价值和延误是否变化。未经这种比较，不能把“团队不信任”当作部署失败的唯一机制，也不能通过强制采纳把表面使用率升高误称价值提升。

13.3 让预测建议可修改可能改善控制感，也改变了处理

决策者希望修正建议，不一定只因抗拒算法。修改功能可能让人补入未入库信息、纠正实体错误或回应临时约束，也可能使采用者把分数改到符合既有偏好。若人只能微调输出，最终动作既非原模型建议，也非纯人类判断。因而应分

SVSAC | sv-sac.org

别保留原始预测、修改后建议、修改原因、最终行动和执行状态；将“可修改工具”与“自动采用工具”视为两个部署版本。

修订日志也不能被直接当作人类贡献价值。管理者修改后结果变好，可能是因为专家补充有效信息，也可能只是其选择了有利时点、回避困难机会，或低估了没有采取该修改时的反事实。至少应在合格机会总体内比较相同任务的版本政策，并将改变的预测内容与行动结果分开。若不能随机分派修改权限，可把日志作为机制描述，明确其不提供人的修改的因果净值。

14. 覆核和否决需要有资源、有规则且能追责

14.1 覆核设计要对错误风险设定审查范围

每条建议都由最高资历人员完整复核看似安全，但当任务量增加时，复核可能退化成快速点击或集中堵塞。另一方面，完全不复核会把模型罕见错误转化为系统性决策损失。更可检验的设计是先按行动后果、预测不确定性、数据新颖度和可逆性定义风险等级，再限定各等级的复核深度、时间预算和升级角色。低影响、容易撤回的建议可抽样检查；高额、不可逆或信息不足的决策应触发更强证据门槛。

复核质量和复核率须分开。若每次复核平均用时从 20 分钟降为 8 分钟，但复核人只浏览摘要，不核关键假设，单位任务的覆盖率提高而失误检出能力可能下降。可通过独立盲审一部分案例估算发现率，记录人工正确发现、误拒和未发现的错误，并估算积压和等待时间。这个试验需要真实分母和明确错误定义，不能将“被人点过确认”作为审核质量。

14.2 复核权责决定错误是否可纠正

如果一线分析师能看到置信度和来源，却没有延迟决策的权限，标记不确定性仍未必改变行动；若业务主管能否决模型，但否决会使其承担额外解释责任，则其表面权力可能不实际。流程应明确谁可暂停、要求补证、覆盖建议、授权例外和恢复运行，以及这些动作的时限和审计记录。涉及战略委员会时，还要区分模型供应方、内部模型负责人、建议使用部门和最终预算责任人的职责，不能用外部供应商承担“模型错误”来替代公司自身的治理决策。

组织责任也会影响试验设计。若模型组对预测准确率负责，业务组对业绩负责，但双方共享一个无法分解的目标，模型组可能优化容易提升的分数，业务组则只挑选接受建议后成功的个案。预先规定主要决策结果、次要机制指标、责任

SVSAC | sv-sac.org

人和数据保管人，能让争议具体化；它本身不会保证公平或有效，却能避免部署后双方对“成功”的定义各异。

14.3 拒绝建议应允许适当理由而不是奖励服从

建议采纳率若被设成硬目标，员工可能接受低质量建议、把动作记录成采纳但私下另行处理，或避免向系统提交难例；如果每次拒绝都被处罚，团队会在系统不适用时仍照办。相反，若拒绝无需留下任何原因，组织难以发现模型在特定地区、业务时期或关系类型上持续失灵。应将拒绝理由限制为可复核类别，并允许“其他/信息不完整”，再定期抽查原始材料，而不是用采纳率单项考核个人。

一个可证伪预期是：如果拒绝主要源于遗漏的事前有效信息，盲审后这些案例应在该信息加入后改善模型校准或行动收益；如果主要源于结果责任不匹配，改变授权与问责设计应先影响采纳/延迟而未必改变模型评分；如果拒绝是人类过度自信，显示历史比较表现可能影响选择，但并不保证经济价值提升。三种机制要求不同的干预和结果指标。以单一“信任度”问卷替代这些区分，会把偏好、知识、责任和绩效误作同一心理变量。

15. 让行动可分阶段，减少错误决策的不可逆后果

15.1 试点不是缩小版的最终推广

一项试点若只挑最容易、最受支持或最有资源的业务机会，观察到的采用和结果不能直接外推至全部机会。试点范围本身会改变可见损失：低风险对象可快速验证工具可靠性，却无法说明高成本错误的风险；挑选愿意参与的决策者能减少拒绝，但不能代表普遍采用。试点应界定目标总体、边界情形、参与部门、可用资源及不得越界的决策类型，并把哪些对象因何原因排除写入研究记录。

试点也要与业务行动保持一致。影子模式可检验预测分布和可能的队列负担，却不会识别人们实际使用建议后的结果；在低额决策中随机试用可估局部执行效果，却不能自动推断董事会对重大资本投放的反应。分阶段推进应逐级改变一项可观察的授权或行动，同时冻结其他条件，或者清楚说明干预包在每阶段改变了什么。此类阶段设计能逐步增加证据，不可把“完成了试点”本身当作批准全量上线的充分条件。

15.2 延迟决策有成本，也可能具有选择权价值

对于不可逆、前期投入大且新信息将很快到达的项目，等待能减少后悔，但等待并非免费。如果延期导致失去窗口、价格上升、关键人员转岗或竞争者占

SVSAC | sv-sac.org

位，等待成本需要纳入；另一方面，如果提前投入会锁定不可回收资金，获得新信息再决策可保留退出或扩张的选择权。用静态的“预测准确率提升”无法比较两者。需要定义信息到达分布、行动期限、等待期间可能改变的收益、固定成本和可撤销比例，并至少比较立即行动、延迟行动及分期行动。

构造一个情景：合同签署后先支付不可退的 30 单位；等待一周可获得额外市场测试，假设它能改变选择，但等待成本为 8；若立即签署，在两种状态下均执行；若先测后签，可在坏状态放弃合同。具体是否等待取决于先验概率、测试精度、损失和时窗。测试精度不应凭“模型置信”代替；未验证的概率输入只能作为敏感性场景。决策规则应在签署之前保存，而不是事后仅报告恰好成功的状态分支。

15.3 可逆性是行动属性，不是模型属性

建议本身可能可撤回，但执行动作未必可逆。可调整广告预算、缩小采购批次或保留第二供应商，和关闭产线、永久取消产品线或宣布大额并购，所需证据门槛不同。即使同一概率分数正确，前者可以低成本试行并观察结果，后者却可能造成永久损失。审批政策可将行动分为试探性、阶段性和不可逆承诺，逐层限定金额、持续时间、退出权及升级条件。

小额试验也可能污染后续政策：客户看到价格、供应商知道资格变化、员工提前调整努力，都会改写后续试验环境。因而所谓“先试点”必须说明干扰范围、可用对照、信息外溢与扩量后会否改变市场均衡。若试点对照被污染或容量未到饱和，估计效果可能只是试点局部，不是全量推广效果。政策可逆性降低错误代价，却不替代对外部性和行为反应的测量。

16. 战略机会的组合层级与共同约束

16.1 分析层级决定“价值”的分母

单个项目的期望净值、业务线的年度表现和董事会的跨年组合收益不是可互换目标。一次模型建议可能提高单个项目的成功率，却占用稀缺资本、延误另一项目或增加组织集中风险。反过来，某业务线的短期现金流改善也不自动说明长期企业价值提高。研究应写清决策单位、预算期间、成本分摊、机会总体与汇总权重：董事会比较的是组合总贡献还是风险调整回报，业务负责人比较的是本部门可控值还是全企业边际贡献，项目主管比较的是交付和质量还是净现值。

多层级的目标有时会冲突。部门按自身成功率奖励，可能偏好低风险的小项目；中央投资委员会按长期成长奖励，可能承担更高失败率但选择期权价值更大

SVSAC | sv-sac.org

的试验。若新预测只提高了某部门的指标，而资本由另一层级负责，建议可能在局部表现看似有效、在全组织资源配置上却不可行。报告时可并列呈现分层结果，但不能把各层指标在没有固定权重的情况下加总成一个“综合经营价值”。

16.2 同一资源池使项目间存在策略干扰

如果两个项目争夺同一工程团队或预算，甲项目获批会改变乙项目是否获得资金及其结果；此时“甲项目单独采用 AI”的因果效应与整个组合都采用新策略的效果不是同一估计对象。即使项目之间技术上互不影响，也可能因共享管理注意、审核排队或供应商额度而产生资源干扰。研究可在组合层随机分配预算政策，或按事前定义的团队/投资池做区组比较，但需报告分配单位数量和池间资源转移。

如果没有随机化，至少要保存决策前项目池、预留容量、优先顺序、预算变更和被挤出对象。只研究最终拿到资金的项目，会把策略选择和结果筛选混为一谈。被拒绝项目不仅是缺少标签的样本，也可能是被选择项目之所以执行的机会成本；其结果未见时，不能把其影子价值自动设为零，也不能凭一个反事实点预测把它算成真实被放弃利润。

16.3 风险约束应作为可比较政策的一部分

组织可能不是单纯最大化平均净值，而是有现金流下限、最大亏损容忍、国家/客户集中度、监管资本或董事会风险偏好。模型若在一个组合里提高平均期望收益，但使一次共同冲击下的亏损超出事先约束，不应仅凭平均值称为更优。相反，组合更稳健但平均回报略低，可能符合政策目标。风险指标、时间范围和约束责任人应在部署前确定，避免看到结果后挑选最有利的风险口径。

一个简化的压力测试可列出基线、单一供应商中断、需求共同下滑和项目延期四种情景，逐项计算现金流、占用资本和可撤回部分；若没有可信概率，就只报告情景下是否越过风险边界，不把情景权重假装成概率分布。多个情景可揭示脆弱点，却不能证明未来发生概率。只有对共因风险有历史数据、结构模型或可辩护的外部信息时，才进一步汇总成概率加权的组合收益。

17. 分阶段投资：把信息到达与承诺时点放在同一决策树

17.1 完全信息与立即行动的构造比较

构造一项不可逆项目：立即投资后，良好状态净值为+70，较差状态为−50；先验概率各为 0.5，不投资为 0。立刻投资的期望值是 (0.5(70)+0.5(-50)=10)。若决

SVSAC | sv-sac.org

策者能免费、立即且无误地知道状态，良好时投资、差时放弃，期望值为(0.5(70)+0.5(0)=35)，相对于立即投资多 25。这个 25 仅为完美信息的上限，假设信息不改变项目回报或竞争状态，也没有获取、等待或延迟成本。

现实中可选的是带误差的测试。假设测试对状态的灵敏度和特异度都为 0.8，先验仍为 0.5，则阳性概率为 0.5，阳性后的良好状态概率为 0.8；阴性后良好状态概率为 0.2。阳性后投资的条件期望净值为 (0.8(70)+0.2(-50)=46)，阴性后放弃为 0，因此测试后策略的期望值是 (0.5(46)+0.5(0)=23)。若测试成本 8、等待成本 5，净值为 10，与立即投资持平；若任一成本升高或信息准确度下降，等待策略就可能输给立即行动。

17.2 试验准确度有价值，但不是“概率越高就越好”

在上述假设中，阳性后投资的边际条件是后验成功率高于(50/(70+50)=0.4167)。若测试灵敏度与特异度相等为 q，且先验为 0.5，阳性后验即为 q；两个信号分支各以 0.5 出现，阴性后验为 1−q，故未扣成本的测试策略总值为 V(q)=0.5×max(0,120q−50)+0.5×max(0,120(1−q)−50)，不能漏掉阴性后仍可能投资的分支。当 q=0.8 时为 23，成本后为 10；q=0.9 时，测试值为 29，扣 13 后为 16，高于立即投资 10。相对立即投资 10，q=0.8 和 0.9 时未扣测试成本的信息增量分别为 13 和 19，扣 13 后净信息增量分别为 0 和 6；它们不同于 23 和 29 这两个测试策略总值。q=0.5 时信号无信息，两个分支后验都为 0.5，V=10 而非 5，付出测试和等待成本反而损失 13。该公式显示，信息价值取决于它跨越行动边界的概率和动作值差，而非只看测试准确率。若测试结果并不改变行动，获取信息就不直接改变该项目的选择价值。

示例也依赖测试准确度在两个状态条件下对称、先验准确、行动净值稳定及没有信息外溢。若测试对坏状态更敏感，对好项目误拒，或市场在等待期变化，应分别重算分支概率和收益；若测试失败会暴露商业计划、让竞争者响应，信息收集本身也改变环境。所有计算应从先验到测试结果、再到后验和动作逐支展开，保证分支概率加总为 1，而不是只给一个测试“准确率”数字。

17.3 阶段门要识别信息收益与延期损失

对于可分阶段的投资，可先投入小额验证技术或客户需求，再决定是否扩产、招聘或长期签约。每一阶段门都需说明下一步行动的可行集合、已沉没金额、可追回价值、决策时可用的新数据以及等待造成的变化。若阶段门后会有新经理、新预算标准或不同竞争条件，比较的不只是模型提供信息与否，也包括决策者和环境改变。

SVSAC | sv-sac.org

一项可检验的阶段设计可按项目族随机分配“立即承诺”或“先测试再决定”策略，按事前合格项目计算平均净值、期限、退出比例、追加资金与未决成本。过程指标可报告测试是否改变后验或行动，但主要结果应保留原分配项目，不能仅比较最终继续推进者。对无法随机的高风险项目，可在匹配约束下比较并给识别假设；若每个项目都采用不同阶段和测试，数据不支持一个共同的“阶段投资效果”。

18. 没有行动变化时，流程成本仍可产生经营价值

18.1 决策净值与分析生产率是两种收益通道

预测信息可以改善项目选择，也可能让既定决策所需的搜集、核查或更新成本下降，即使董事会最终选择不变。两种通道要分别测量：决策通道比较可执行政策下的结果差；流程通道比较完成同一质量标准所需的现金支出、人员工时、等待时间和错误返工。若模型令预测报告成本降低但不改变投资动作，这仍可能产生流程价值；但不能把这种节约写成“AI 提高了决策结果”。

18.2 一个不把释放工时误当现金的构造账本

假设一个团队每季度处理 100 项机会，原流程平均每项需 12 小时，综合人工成本假设为每小时 80 单位；新工具将初稿时间降至每项 7 小时，但每项仍需 2 小时验证。旧总工时为 1200 小时、按该费率折算的人工等价值 96,000；新总工时为 900 小时、人工等价值 72,000，账面可释放 300 小时和 24,000 单位的人工容量。若另假设 24,000 人工支出全部可实际避免，系统季度维护与审计 18,000 及额外抽检 4,000 均为新增现金支出，才得到净现金节约 2,000 单位。若工资、人员编制和外包支出均不变，则人工现金支出并未减少；在同样新增 22,000 现金支出假设下，当期净现金变化为−22,000，同时保留 300 小时可重新部署的容量，不得称为已实现现金利润。

若这 300 小时用于新增项目，则需记录新增机会、增量成本和真实产出，不能把“节约工时×全成本率”与新增项目贡献重复计入。若团队本来有闲置容量，边际机会成本可能接近零；若员工加班或外包被实际取消，现金节约才可能接近对应支出。案例中的全部数值都是作者假设，未基于公司台账；其用途是展示必须区分工时、现金成本、维护费用和再部署收益。

SVSAC | sv-sac.org

18.3 对照质量与过程改进的成本边界

若新流程用更短报告、更少核查或延后复核来缩短工时，比较必须保持质量门槛相同，或把质量下降及错误处置纳入后续损失。否则“更快”可能只是把成本转到事后补救。一个可执行的试验可将相同类型任务随机分配到旧流程、新流程，并使用独立评分者盲评完成质量，另记录总工时、系统成本、等待和返工；主要成本分母应是所有随机分配任务，而非仅成功完成者。

成本节约的边界也需透明。一次性实施费、培训、数据整理、权限审批、模型更新、审计、服务故障和退出迁移都应按观察期限纳入；长期固定支出可按明确折现和使用期分摊，但不得既把整笔实施费列为全额成本又将摊销额重复扣除。若现金流和净现值无法可靠估计，可先报告资源使用与敏感性范围，不用模型建议的战略价值替代实际财务账。

19. 预测团队：协作增益与共同误判的边界

19.1 预测任务并非天然等于团队决策

复杂经营问题经常被拆成对市场规模、竞争行动、技术可行性、监管时点和成本路径的若干概率判断，再由项目负责人形成投资建议。拆解有助于显露不同假设，却不保证多个预测汇总后就是决策。至少存在三个独立环节：成员如何产生判断，成员之间是否交换信息，以及汇总规则如何把多项判断变成一个预测。最后还要有一个把预测映射为行动的规则。若上线同时改变这四处，结果改善无法归因于某个模型或“人机协作”标签。

Mellers 等的预测锦标赛论文保留为训练、团队与聚合研究的待核线索。[8] 本篇未核到其随机分组和结果对应原段，不将具体条件安排当作已证实的比较。我们提出三个可分别测试的因素：概率训练、信息交流和聚合规则。即使这些因素改善特定预测任务，也仍须另验是否改变企业选择及净值，不能从预测竞赛题名推得给公司会议增加 AI 便会提高回报。

19.2 交流带来的信息增益和相关误差

团队交流可能把分散在成员手中的资料和理由带入判断，也可能让成员共享同一错误前提。判断的数量不是独立信息的数量。若五位管理者分别掌握供应商、客户、财务、工程和政策信息，先独立写下概率及依据，再对差异进行交叉质询，交流有机会暴露不一致；若五人都只读了同一份模型生成的摘要，表面上

SVSAC | sv-sac.org

的五票并未提供五个独立信号。把讨论后的共识当作“专家多数”，会掩盖共用信息源导致的误差相关性。

独立先报与讨论后修订的双阶段程序可把一部分机制变得可观察。系统先冻结成员的初始概率、依据摘要及其可用资料时点，再保留交流内容或至少议题清单、二次概率和最终聚合值。对每项新证据记录提出者、时间和可核查来源；对判断变化记录方向与理由。这样不要求把会议录音当作真值，而是能检验团队是否因补入新信息而修正，还是只向资历最高者或模型建议收敛。

可见群体分布也不是中性展示。较早公布多数人的估计，会给后续判断者提供统计线索，同时改变其独立性；若参与者把“多数”误当成“可靠”，错误可能通过从众扩散。相反，隐藏他人预测到首轮封存后再开放分布，可保留先验独立信息，但可能牺牲及时发现事实错误的机会。对于决策窗口很短的任务，适宜比较的并非“讨论/不讨论”抽象二元，而是固定流程包：先独立判断、何时共享证据、何时显示模型、以及最终由谁整合。

19.3 训练、团队和筛选不能混称为模型效果

预测训练、团队协作和筛选高表现者是不同安排：训练拟改变概率判断习惯；团队拟改变信息交换和责任；筛选则先改变参与者。若三者一起引入，结果不能归因于模型。本篇不将尚未核到原段的“tracking”具体分组当作已核事实；对企业的条件判断是，按既往成绩挑选成员时，必须区分选择、回归均值和后续协作效果。

一个假设性小例子说明组成变化的风险。若原团队有 10 人，其中 2 人对一类事件平均误差低、8 人较高；新团队只保留前 2 人，最终预测的误差下降可能来自成员筛选而非讨论流程。如果同时引入 AI 摘要，既有高能力成员还可能获得更多信息，单组前后比较无法拆分来源。可检验的对照应至少保留同一任务池中“独立判断后机械聚合”“独立判断后讨论再聚合”“模型输出加独立判断”和“当前流程”几种策略；将团队而非每个预测条目作为随机单位，以免同一团队成员跨臂交流造成污染。若公司只有少数团队，随机化可能无足够精度，应先报告可检测的效应范围，而非把大量同一团队的预测行误当成独立样本。

19.4 预测准确之外的经营检验

地缘政治预测研究中的结果是概率预测质量，不能代替经营项目的行动与回报。把概率从 0.60 改到 0.72，只有在对应行动阈值附近、并且回报差异足够时，才可能改变项目选择。反过来，校准不变也可能通过缩短资料搜集或降低审核成本产生流程收益。评估应分开报告：预测分数及其不确定性；建议是否改变、是

SVSAC | sv-sac.org

否执行及执行延迟；项目结果是否成熟；成本、收入、风险和被挤出项目。若模型只改善置信度显示，却使委员会更快批准其本来就会批准的项目，结果变量应指出这是流程时长变化，不能包装为投资价值增加。

对策略类问题，很多标签并非简单的“事件发生/未发生”。某地区进入是否成功，可能同时取决于进入后采取的价格、渠道和服务；若不进入，目标市场的潜在结果不可观察。因而一个总体预测模型的离线评分只能在已观测政策和标签定义下解释。它不能自动回答“换成另一套决策政策后的净值”。这一区分要求保存从假设、概率到行动和执行的版本化记录，让后续成熟结果能够与原始决策时点相连，而不是只保留最终汇报中的胜败叙事。

20. 模型、组织流程与任务分布共同决定比较对象

20.1 “新模型优于旧模型”需要固定比较的任务总体

组织往往以本季度已完成的项目评估模型升级，但已完成项目不是中性样本：容易判断、短周期和成功后愿意披露的项目更容易留下完整标签。若新模型只在更简单或结果更快成熟的项目上使用，平均分数提高可能来自任务构成变化。需要先定义可进入比较的任务框，例如预算范围、决策负责人、信息截止日、成果成熟期限及不可接受的风险类型，再锁定完整分母。新旧模型在同一任务上并行生成预测可比较预测差异，但不能让未执行一方的建议结果冒充反事实经营收益。

模型更新也会改变工作流。更长的解释、更频繁的告警或更高的置信度会占用复核容量，使另一些任务变慢。若把“每个任务预测更准”作为唯一指标，就漏掉了队列中的选择性延迟。一个固定任务总体的试点应同时保留原始分配、系统实际可用、人工查看、建议采纳、行动时点、结果成熟状态和复核分钟；对离线未覆盖、故障拒答、超时任务仍留在分母并报告状态，而不是静默删除。

20.2 因素拆分与交互解释

若问题是区分“模型本身”“更规整的信息收集流程”和“模型加流程”的贡献，可以用因素设计而不是单一前后比较。构造四组：旧模型+旧流程、旧模型+新流程、新模型+旧流程、新模型+新流程。以预先冻结的任务池为单位，在可行时按业务单元区组随机；若同一团队同时处理多个条件，需记录交叉污染并考虑以团队簇分派。结果至少包括预测质量、决策政策净值、执行合规、时间成本及未成熟率。主效应可回答平均的模型切换差异或流程切换差异，交互项表示模型效果

SVSAC | sv-sac.org

是否依赖流程条件；交互显著仍不自动揭示是信息质量、注意力还是激励机制，机制要有对应过程测量或额外随机操纵。

例如以净值为方向，定义四臂均值为 μ00、μ10、μ01、μ11，前一索引代表模型，后一索引代表流程。模型在旧流程下的差是 μ10−μ00，在新流程下是 μ11−μ01；它们相差 (μ11−μ01)−(μ10−μ00)。若新流程使模型结果更易被采纳，这个差可能是实施渠道，而非模型预测本身更准。相反，若模型提高预测分数但决策净值没有变化，应报告预测与行动之间的断裂，不将两者平均成一个含义不清的“AI 绩效”。这组符号是作者提出的识别框架，不是某项已执行实验的结果。

20.3 执行与暴露依从性

随机分派策略识别的是提供某策略的意向处理效应，而不是每一次实际使用模型的效应。用户可能未登录、拒绝建议、手动改写、因供应商延迟而无法执行；如果只比较“实际使用者”和“未使用者”，这两组通常在任务难度、管理偏好和信息质量上不同。首要报告应按原随机分派计算全部符合条件任务的结果。依从性分析可以补充，但要明确需要的额外假设，不能把“采用模型的项目表现更好”直接归因于模型。

在现场实施中，策略的可用性也属于处理的一部分。建议界面宕机、输入数据迟到或负责人没有权限执行，都是策略包的真实组成而非可剔除的技术噪声。另一方面，若研究要区分算法质量与上线可靠性，可把“模型预测可获得”与“提供支持流程”拆开随机，或按预定义的系统可用性分层报告。不能事后只保留系统正常的任务，因为正常运行本身可能与业务环境或任务类型相关。

20.4 报告规则与局部最优

在组合层，建议冻结一个主结果和少量解释性结果，并以同一任务总体报告。若需要多个财务结果，可逐项展示现金流、风险暴露、项目完成与时间，不应通过任意标准化权重拼成事后“综合分”。重要的负向或零结果需与正向预测分数并列，因为只有这样才看得出是否出现“模型更准、经营没变”或“短期净值提升、下期风险加重”。如果团队在看到结果后频繁调整阈值、输入和审批流程，每次调参都属于新的策略版本；该试点数据只能作为开发数据，确认性结论应留给新的时间段或独立任务样本。

这种冻结不意味着经营政策永远不变，而是防止同一批结果既用于挑选最好方案又被当作无偏验证。若组织确需滚动更新，应按时间划分开发与确认窗口，记录每次政策变化、目标总体、触发依据和停止规则。样本过少时，严谨结论可

SVSAC | sv-sac.org

以是“尚不能区分”，而非以管理者共识填补统计信息。模型和流程持续迭代带来的学习价值可另行计量，但不可把尚未发生的未来改进记作当前试验收益。

21. 战略预测的成熟周期、共同冲击与可见性

21.1 结果窗口决定决策建议能回答什么

经营预测常跨越不同期限：短期订单、季度价格、年度客户留存、三年技术路线和多年资本回收。把这些结果压到一个季度统计，会系统性偏向早成熟指标。若预测系统让管理层提前签约，短期收入可能提前确认，但真实需求并未增加；若它延迟一个大额承诺以等待测试，短期项目数可能下降，远期损失却可能更低。因此每项决策应在分派前规定多个时间终点及其经济含义，如 90 日现金流、年度贡献、合同续约、退出损失和风险暴露；同时标注哪些终点只是先行指标，哪些是最终结果。

固定截止日并不使各项目成熟度相同。项目进入较晚、付款周期较长或客户审批缓慢，都可能让处理组在观察日出现更多未决结果。若只比较已成熟项目，条件化于处理后变量可能破坏原分派的可比性。主要分析应保留全部分派项目，分别给出已实现值、未决账面和可识别区间；对于未决项，可报告在预设上下界下净值差的敏感性，而不是用预测模型替代真实结果再声称验证。

21.2 共同冲击会使“项目数”高估有效样本

多个项目可能同时暴露于同一利率变化、客户预算冻结、关键供应商或管理团队。十个项目并不一定提供十份独立信息。若所有项目由一个投资委员会决定，委员会层级的学习和沟通会污染项目级分组；若行业冲击同时影响整批项目，常规项目级标准误会低估不确定性。评价单位应与处理实际发生的层级对应：同一委员会学习新工具，处理往往发生在委员会或时间段层面，而非每个项目互相独立。

可用分层模型或簇级不确定性估计描述相关性，但它们不能凭空创造独立委员会。设计阶段应列出可能共同变化的层级、可用簇数和重复周期；如果只有几个簇，置信区间可能很宽，此时更合适的判断是试点提供方向性或机制证据，而不足以推断全组织政策效果。随机化也不保证外推：随机化解决的是分配臂之间的比较，样本是否覆盖下一年、其他地区或不同授权结构，仍需另行说明。

SVSAC | sv-sac.org

21.3 分层预测与结果稀少的权衡

战略事件往往稀少，且单一类别内部也不稳定。把五年间的所有机会合并成一个校准率，可能掩盖新产品、成熟产品和受监管市场之间的差别；反之，每个细分只有几项事件时，独立估计又会极不精确。分层汇总可在“完全共享信息”和“各类完全独立”之间折中，但收缩模型本身也依赖类间可交换性假设。应同时报告总体、预先定义的主要层以及层内有效样本和区间，不根据结果再挑最漂亮的子群。

更重要的是，同一层内的预测概率不能替代行动损失。一个监管事件概率估得很好，不代表企业已经知道该事件发生时采取何种措施更好；一种产品需求均值的分层校准也不决定是否扩产。把预测质量分层后，还需把层级与决策规则、预算或风险上限相连。否则更精细的分群可能只是制造更多报表，而没有形成可执行的差异化行动。

22. 可复核的团队辅助决策现场协议

22.1 研究对象与分配规则

可将一年内符合条件的战略机会纳入一个前瞻性评估，例如新品立项、市场进入或重大产能承诺。每个机会须在数据截止时点登记目标结果、可选行动、预算、风险上限、已有信息以及预计成熟时间。比较的不是“模型对人”的抽象能力，而是四种事先规定的策略包：当前会议流程；独立预测后机械汇总；独立预测后按结构化议程讨论；以及先独立预测、再查看模型与群体分布、最后讨论。每项策略都规定参与者、资料访问时间、更新次数、行动负责人和不可越过的批准门槛。

若一个委员会在多个机会间反复使用同一策略，成员会学习、共享结论并影响其他项目。可在委员会或相互隔离的投资池层面按预先区组随机分配策略，而不能把同一委员会的每个项目都当独立随机单位。若簇数不足，分配可采用随机化时段或阶梯式上线，但必须说明时间趋势、学习残留和跨期迁移如何影响比较。不同策略不能同时改变预算上限、审查人员和技术支持，否则观察到的是完整策略包差异，不能单独归因于预测算法。

22.2 主要估计量和过程变量

主要估计量可定义为原始分派策略的意向处理效应：在纳入的机会总体中，各策略臂到预设期限的净现金流或经预注册的风险调整贡献均值之差。所有符合

SVSAC | sv-sac.org

条件的机会按原分派保留，包括未打开模型、人工否决、系统故障、未获批准、延期、退出和截止日仍未成熟者。未成熟结果单列，不因为某组更快结束就只比较其已结束项目。对无法在期限内完全观察的收益，先给成熟结果，再对未决结果给范围分析；模型自身生成的预测不能充当终点事实。

过程变量有助于解释“为什么”，但不是结果替代品：独立初始判断与讨论后判断差距，新增证据数量及来源，模型建议被查看或修改的比例，会议时间，审批周期，资金配置变化和决策者信心。预测分数按任务及难度报告；若成员自行选择回答哪些问题，需将未回答比例和任务选择列出，避免只用最终作答问题评价。对问题内重复概率更新，先明确每项决策取最终截点、固定日截点或整个概率路径的哪一个，否则多次更新会改变计分权重。

22.3 规模、停止和决策门槛

样本量不能从“每臂有几十个项目”这一惯例推出。应先确定决策上最小重要净值差、结果方差、簇内相关、各类策略的可用簇数以及延迟成熟比例，再计算可检出的范围或进行设计模拟。若管理层不能提供可信的最小重要差异，研究可报告估计区间而不宣称“没有效果”。对非常少见的大额损失，均值检验可能缺乏信息；需单列严重事件、压力情景与风险容忍边界，但不将零事件解释成零风险。

停止规则应限制中途反复查看并挑选有利策略。若确因安全、法律或重大预算风险需提前停止，应预先写明负责者、可触发证据、何种数据用于判断，以及停止后仍如何计入原始分配结果。最终决策可分成三种：若净值差的可信区间高于预设决策门槛且关键护栏未越界，可扩展到相近任务；若区间跨越正负且证据成熟不足，则延长观察或调整设计；若流程成本虽下降但决策净值未变，只能按成本通道评估。没有达到门槛不等于模型无用，而是当前证据不足以支持扩大承诺。

22.4 复现与迁移的记录要求

每个预测版本应保存模型、提示或参数、训练/检索快照、外部数据时点、校准层、参与者角色、分派程序、人工修改、建议展示时间和动作执行记录。战略机会往往不适合公开敏感数据，但仍可向独立审计者提供脱敏任务定义、预注册分析计划、分配日志、代码和结果变量字典。若模型服务或知识库在试验期间改变，需登记版本窗口并判断这些改变是否属于分派策略的组成。没有稳定版本，所谓“复现同一 AI 政策”可能只是复现一个不断变化的名称。

SVSAC | sv-sac.org

迁移到新业务前，应比较决策权、信息到达、结果成熟速度、预算限制和风险承担主体是否一致。相同预测算法在市场进入和短周期定价中的行动阈值、可逆性与反馈速度都不同。迁移证据可以分层积累：先复现预测质量，再验证某类行动规则，最后确认在新场景的完整净值；每一步的通过都不能替代下一步。该协议是我们提出的研究方案，不代表已有企业已按此实施，也不提供统一的自动批准规则。

23. 预测的是风险还是行动响应

23.1 高风险对象未必是高收益干预对象

企业决策常把“最可能失败的项目”或“最可能流失的客户”作为优先干预对象。这只有在干预效果与基线风险同步时才成立。风险预测描述既定政策下的结果概率；响应预测描述对象在不同动作下结果如何改变。一个客户即使自然流失概率很高，如果补贴、技术支持或合同调整均无法改变其选择，投入干预的边际价值仍可能很低。风险较低但对小额支持高度敏感的客户，反而可能更值得优先服务。

构造一个可复算的二组例子：每组各有 100 个账户，单账户在不干预时的流失贡献损失均按 100 单位计。A 组的基线流失率设为 0.40，某项支持使其降低 0.10，单户干预成本 5；B 组基线流失率为 0.20，支持只降低 0.02，成本同为 5。A 组每户预期避免损失为 (100×0.10=10)，扣成本后为+5；B 组则为 (100×0.02-5=-3)。若只按基线流失概率排序，就能识别 A 较高风险，却不能从风险分数本身得出该支持的边际效果；本例排序碰巧一致，是因为设定 A 的响应也较大。若 A 的可避免差为 0 而 B 的可避免差改设为 0.08，同一风险排序会将资源投入净值为负的对象。

这些数字是作者构造而非客户数据。要估计行动响应，应观察或设计不同账户在可比政策下的结果；不能把 (P(Y=1|H)) 当作 (E[Y(0)-Y(1)|H])。若可用支持容量有限，策略价值还取决于对象数、服务时长、拒绝率、支持成本和账户间溢出。模型的风险排序仍可帮助风险监测，却未必是干预优先序。

23.2 相同预测对不同决策函数产生不同排序

一个市场需求预测可以服务于是否进入、投入多少库存、是否签长期供应合同等多种动作。各动作拥有不同成本结构和退出条件，因此同一概率排序不决定同一投资排序。市场甲需求达到阈值的概率略高，但其失败时存在大额沉没投

SVSAC | sv-sac.org

入；市场乙的需求概率稍低，却可先以小规模试销并在一个季度后退出。若只按成功概率比较，甲更优；若按期望现金流或下行约束比较，乙可能更优。

决策记录至少应把“预测对象、可行动作、收益/损失映射、限制条件”绑定为一个版本化对象。比如预测的“次年客户数”可能是按当前价格和服务水平的预期值；若管理者据此改价或削减服务，预测定义和结果生成过程均已改变。沿用原分数评价新政策，可能只是把条件预测错误地当成不依赖动作的自然规律。

23.3 从点预测转向决策边界附近的证据

在明确行动阈值后，远离边界的预测和接近边界的预测具有不同的信息价值。若无论分数如何都不会进入某市场，继续提高该市场预测精度未必改变当前行动；若项目分数恰处于预算排序线附近，一小段校准误差可能改变谁得到资金。反之，若边界项目金额很小，排序翻转的财务意义可能有限。由此，样本与复核资源可依据行动边界附近的决策损失分配，但需避免只报告被挑出的边界子集而忽视总体成本和安全边界。

一个可检验比较可在预注册的边界区间内，对预测增强策略和基线策略随机分派给决策池，记录排序变动、最终分配、每单位资本贡献与未获资金项目的状态。分析仍按最初符合条件的全部项目计算，边界区间作为预先声明的异质性结果。若策略只改变边界附近建议而未改变最终分配，应报告为建议层效果；不能声称资金配置已被优化。

24. 部署后的反馈：预测政策会改变未来标签

24.1 观测结果是政策作用后的分布

部署前的历史数据来自旧政策；部署后的结果则由新政策和环境共同产生。高风险项目被加强审查后，失败率可能下降，若把更新后的低失败率直接视作“模型此前高估风险”，就混淆了风险预测与干预效果。另一方面，模型建议降低某些对象的预算，结果因此未发生；这些对象的后续标签并非独立缺失，而是受到行动影响。训练数据会随被预测的对象获得何种待遇而变化。本段讨论企业项目中行动与结果可见性的条件关系，是作者的分析，不将尚未核到原段的司法研究作为已证实的外部案例。

设二元结果 (Y) 的潜在值为不行动时 (Y(0)) 与行动时 (Y(1))。在历史策略 (π_0) 下，只观察 (Y(A))，而非同一机会两种动作的结果。新政策 (π_1) 改变行动分配后，未来可见结果分布是新政策产生的 ($Y(A_{\pi 1})$)，不是历史基准 (P(Y(0)|H)) 的独立检验。若把后处理结果回填为原风险标签，会把“行动降低损失”误写为预测

SVSAC | sv-sac.org

失败；若仅对未干预对象更新风险模型，则样本被政策选择，可能造成相反方向的偏差。

24.2 一个简单的标签反馈反例

构造 100 个项目：其中 20 个在不采取行动时有 0.50 的失败概率，80 个有 0.10。系统按预测选出前 20 个接受强化复核，假设该复核使高风险组的失败概率从 0.50 降到 0.10；其余 80 个仍按原政策观察到 0.10。旧模型的风险排序本来区分 20 个高风险和 80 个低风险，但上线后两组观察失败率都约为 0.10。若不保存行动分配，更新模型会看到“高分组也只有 10%失败”，继而把高风险分数压低；这并不意味着初始风险排序没有信息，而是行动把风险实现概率改变了。

该构造并非说真实管理审查一定把风险减到 0.10，也不是估计真实效果。它仅说明需要将预测时可见基线、后来采取的行动与实际结果分开记录。若复核并没有降低失败概率，或高风险项目具有不同结果成熟速度，表面模式会不同；必须通过随机化、可信外生分配或可辩护的识别假设估计行动响应。

24.3 反馈校准应分别追踪风险与策略结果

运营中可以同时维护两个对象：在当前政策定义下的结果预测，以及备选行动相对于基线的增量响应。前者可用行动状态作为条件并持续校准；后者需要行动对照或其他识别来源。两个模型的标签、更新时间和适用政策版本应分别记录。若把所有记录压入一个“失败率”字段，就无法区分基础风险变化、客户/市场构成改变、策略作用和数据迟到。

监测也应区分数据漂移、预测漂移和政策效果漂移。输入变量分布改变可能是市场组成变化；同一预测分数下的结果改变可能是校准漂移，也可能是行动政策不同；净值下降还可能由成本或对手反应造成。报警应触发调查而不是自动证明模型失效或成功。短期校准改善只有在结果定义和分派政策可比时才支持预测质量改善，不能替代政策净值结果。

24.4 更新与学习的成本应归入策略版本

若人们因预测工具改变搜索、审核或资源配置习惯，策略效应可能随时间变化。训练能否迁移到新团队，数据回馈是否改变模型，旧分数是否继续被管理层引用，均应纳入版本记录。更新数据、重新校准和培训占用资源；不更新则可能造成适用域外误差。两者是策略成本与风险的权衡，而不是一个可以被默认忽略的技术维护事项。

SVSAC | sv-sac.org

对连续更新系统，确认性评价可设置周期性冻结点：在每个周期开始前记录模型/规则、决策对象、结果定义和主分析；周期结束后锁定数据和分析，随后再更新。若一边看结果一边改阈值，需把每次更改列作新政策版本，不能以“自适应”为由把同一段数据既用于优化又用于声称成功。没有可比的冻结时段时，可以描述运行表现和变更轨迹，但对升级的因果增益应保持未识别。

25. 战略选择中的不确定性与稳健边界

25.1 输入概率的误差会被动作阈值放大

当价值函数对概率近似线性且行动远离边界时，小概率误差可能只造成小幅净值误差；当动作阈值附近或损失高度不对称时，同样的概率误差会改变动作。项目预测不应只给一项点估计，而应同时呈现校准证据、可迁移范围、损失假设及阈值敏感性。若模型输出 0.58，估计区间或校准不确定性跨越行动阈值 0.55，行动就不能由小数点后两位的表面精度决定。

25.2 决策不确定性不等于预测区间

预测区间描述某个预测对象在统计假设下的不确定性；决策不确定性还包括收益/损失参数、约束是否执行、竞争者反应、结果成熟时长与组织采纳。用单一的预测置信区间推导“最优策略”通常把这些未建模对象固定成已知值。更诚实的敏感性分析分别改变成功概率、失败损失、投入成本和实现概率，展示行动何时翻转；若多个输入相关，还要说明相关结构的根据，不可通过独立抽样默认消除共同风险。

25.3 稳健选择可能牺牲期望值但降低可承受失败

决策规则可以最大化期望净值、限制某一尾部损失、或在一组合理参数下保证最低可接受结果。三者不是同一目标。例如两个市场机会期望净值分别为+10和+8，但前者在坏状态下亏 50、后者最多亏 12。对资金有限或不可逆项目，风险容忍可能使第二个方案更可取；若企业有充分缓冲和高增长目标，则第一项也可能合理。所谓“最稳健”应说明相对于何种不确定集和谁的风险承受能力。

25.4 参数区间需要可解释的来源

敏感性范围可来自历史重复、外部可比案例、专家先验或明确的管理层压力假设，但来源不同的区间含义不同。历史区间是样本与政策条件下的描述，外部可比需要迁移假设，专家范围是判断输入，压力情景并非概率区间。报告应逐项

SVSAC | sv-sac.org

说明参数、范围、数据来源、更新时间和适用人群；若范围纯属假设，就以“如果”表达，不把它写成估计置信区间或行业基准。

26. 预测价值的反证方式

26.1 什么时候“更好预测”可能没有决策增量

一个模型在常规项目中显著改善分数，但所有项目的行动都由硬性安全、法律或资本约束决定，且预测变化不改变任何可行动作。此时预测能力提高可作为认知或监测产出，却不能声称已经改变该政策的决策净值。另一种情况是预测改进集中在远离决策边界的项目，排序不变；在所测窗口和预算下，它可能没有行动增量价值。若同时降低信息收集成本，则仍可有流程价值，需按第 18 章的资源账本测量。

若模型改善在不同季节、地区或团队消失，可能是原训练数据中的任务组成、信息权限或专业人员能力所致。预先规定会推翻预期的观察，如独立外部样本中校准不改善、行动变化没有发生、政策净值不升或严重损失上升，有助于避免只挑选成功案例。反证条件须在看结果前写清，不能看到不利结果后改称目标只是“提供参考”。

26.2 替代机制不能被平均效应遮盖

观察到部署后收益增加，至少有几种竞争解释：预测信息更好、团队训练提升、额外复核增加、管理层集中注意、预算增多、同期市场景气，或只有强势部门主动申请试点。一个总体平均差不能区分这些机制。可通过流程账本、对照组、分层和因子设计逐步排除部分解释；不能随机的机制需保留为替代解释，并缩小结论范围。

同样，观察不到收益也未必意味着算法无价值：信息可能没有被展示、操作方缺乏权限、结果仍未成熟、样本太少、既有政策已接近上限，或预测分数改进没有跨行动阈值。不过这些都是待查机制，不应成为结果不佳后的默认免责理由。每项解释都需要对应可观测量或设计对比，若数据无法区分，就应明确报告未识别。

26.3 适用范围应随决策条件陈述

研究结果应描述为“在这一类任务、这一套信息截止、这一种决策权限与此成本/期限范围内，观察到某项预测或政策结果”，而不直接写成“AI 改善战略决策”。短期、重复、结果清晰的预测任务可以支持概率技能训练；高额低频的不可

SVSAC | sv-sac.org

逆投资可能需要不同风险控制、专家证据和长周期追踪。低风险的建议自动化与董事会资本配置的责任结构也完全不同。

模型作为单一组件的测试、模型加团队流程的政策测试、以及整个业务上线的净值试验，应明确其证据层级。一个固定版本的概率评分实验可验证任务表现；一个团队随机试验可检验该任务内的工作流政策；实际投资决策的长期比较才接近企业净值问题。不同层级之间可累积信息，却不能互相替代。本篇尚未进行后两类企业试验，因此关于实际战略回报的判断仍是待验证的条件命题。

27. 经营预测证据如何表达为结果权衡

27.1 从预测误差转为决策者实际面对的曲线

Gardner 论文的出版商摘要比较预测模型在库存投资与客户服务之间的权衡。[9] 本篇只取得该摘要，未核到系统模拟细节；撤回 5,661 项目、九年回放及稳态窗口的具体转述。这一窄依据支持把预测评价与资源、服务口径相接，不支持报告已实现 AI 收益。

本篇不采用尚未核实的 4.2 亿/3.9 亿美元和 32/26 天数值作比较证据。作为分析问题，可分别询问固定服务目标下资源是否下降，或固定资源下服务是否改善；两种比较须用同一边界的实际数据或明确模型计算，不能互换成一个通用节省率。偏好仍由承担成本与风险者确定。

27.2 模拟将预测与决策规则接在一起，也引入模型假设

作为我们提出的分析路径，可将预测代入补货、排班或产能规则，再比较资金占用、服务延迟、等待与闲置；但模拟必须明确输入分布、执行规则、提前期、误差和初始化。本篇没有核到 Gardner 正文中零需求、异常值或趋势处理的原段，不将这些具体方法归属作为支持。模拟只在其假设内连接预测与动作，不能自动识别上线效果。

历史回放并不自动等价于政策反事实。若替代规则会触发新订单、改变供应商响应、诱发客户替代或带来其他部门排队，模拟行为必须明确建模；只将不同预测喂给同一条固定规则，可能遗漏规则调整后的二阶影响。对组织而言，可以用离线仿真筛选有希望的政策，再以有限现场比较验证关键路径；若系统不可安全或合法随机，则需说明历史回放、仿真和真实执行分别支撑什么结论。

SVSAC | sv-sac.org

27.3 资源变动有过渡成本，稳定期收益不等于当期净值

在一个条件情境中，预测升级若导致提前补货或扩张，迁移期投入可能先上升，即使假设稳态所需资源较低，也不意味着当季现金增加。这是本篇的情境推论，不作为 Gardner 已核实的发现。需要计入未完成订单、在途采购、并行运行、供应商安排及培训的过渡成本；缺少参数时应保留区间，不能预记节省。

这种过渡问题也出现在战略决策：更好市场预测可能促使企业提前招聘或签约，短期支出增加；预测评估若只在一年后核对收益，会漏掉这一资金路径。要报告净值，需将决策当期、过渡期和稳态期分开，明确现金流折现与风险口径。若无法估计过渡行为，报告可把它作为范围或压力情景，但不能把稳态仿真中的节省直接记入已实现收益。

27.4 这种证据对战略决策提供的是方法边界

物资库存是重复、高频、目标定义较清楚的运营问题；其供应水平、投入规模和历史时间序列均比许多一次性战略项目更可测。预测规则的模拟曲线能够较直接地连接资源与服务，但这种证据不自动回答低频项目的潜在回报、市场进入的竞争反应、一次性研发投资或董事会风险偏好。战略机会往往没有可重复试验、行动会影响标签，并且结果需数年成熟；一个运营模拟方法可借鉴为“预测–规则–结果”的分析顺序，却不提供战略政策的现成参数。

因此，证据强度应跟随决策重复性与结果可观测性。重复业务有机会在多个周期中比较预测和政策曲线；稀缺战略事项可能只能用事前概率、结构化情景、阶段性里程碑及严格限定的外部参照积累证据。不能因为后者样本少就把判断包装成确定模型输出，也不应要求所有高价值战略选择都具备同一统计试验形式。可检验的命题是：在信息更新与行动规则可记录的条件下，策略包是否改善目标曲线，并且改善是否足以抵消迁移与执行成本。

28. 预测分数进入管理沟通后会改变决策权

28.1 数字压缩会隐藏分布与价值选择

董事会材料通常需要简洁表达，而概率预测本身需要时间、结果定义和条件信息。将风险分数压缩成红黄绿、单一“成功概率”或模型推荐，会隐藏基率、区间、参考类与损失函数。即使报告中保留小数点，管理者也可能把一个条件概率当作确定事实，或把模型输出与公司目标的权重混为一谈。表达选择因此是政策包的一部分，不只是界面风格。

SVSAC | sv-sac.org

对有重大不可逆风险的决策，可将概率、区间或情景范围、对应行动阈值、主要不确定来源和决策责任人放在同一页；同时说明行动不变时的流程成本收益应怎样另行记录。若读者只看到“70%成功”，却看不到该概率来自哪些可用信息、属于哪种项目总体、在何种动作下成立，就无法判断其适用范围。概率分层需与企业行为相连，不能把图表清晰度当作证据质量。

28.2 信心、建议和授权应分开

一项模型建议可能以高概率表达，但模型不拥有企业资源，也不承担错误的财务或声誉责任。授权人必须知道建议是否为风险估计、行动效果估计或成本预测，以及系统是否在新任务上超出训练/验证范围。若模型的置信度被误解成“保证”，组织可能将风险承受责任转给没有审批权限的分析员；若所有模型建议都必须接受，人工审核退化成程序签字；若建议失败后才按结果追责，决策人可能形成过度保守或只选择可免责项目的行为。

一个可观察的责任链需把模型所有者、目标定义者、批准人、执行者、数据维护者和结果核验者分开记录，同时注明谁有权拒绝、暂停、追加验证或修改策略。评分团队对预测质量负责，业务主管对可执行行动与资源承担责任，财务/风险团队审计净值口径，不能由一个“AI 项目负责人”代替各方的实际权力关系。此为组织设计建议，不是现有来源验证的普遍制度效果。

28.3 预测建议可能触发的激励反应

如果业务经理按“被预测为低风险”的项目成功率考核，可能会拒绝模型识别的高潜但高风险项目；若按模型采纳率考核，可能机械执行不合适的建议；若按预测误差归责，经理可能减少信息共享或避开难预测事项。相反，如果仅惩罚失败，管理层容易偏好无需承担不确定性的短期项目。上述机制均有可观察含义，但方向取决于薪酬、晋升、授权和项目选择规则，不能未经数据就写成既定企业事实。

试点评估可预先记录拒绝理由、模型建议后的审批变化、承担预算者、执行是否发生、失败后责任归属和跨部门移交。若只收集采纳率和满意度，无法区分能力不足、目标不一致、信息遗漏、风险偏好或责任规避。还需比较相同团队在不显示模型、显示模型但不提供默认选项、显示建议且要求书面理由等流程条件下，建议采纳和结果是否变化；不同界面/治理流程本身即不同处理，不应合并称作单一模型效果。

SVSAC | sv-sac.org

28.4 失败后的复盘如何避免事后偏差

结果糟糕不自动意味着当时决策不合理；运气好也不证明决策过程正确。事后复盘应先重建决策时信息、目标与可行选项，再比较当时概率判断和现有成熟结果。若模型输出版本、负责人修改和审批边界未保存，组织只能用当前模型解释过去决定，产生事后合理化。保留决策日志既支持错误分析，也使得重新训练和复制时可以区分基准数据与行动后数据。

对一次性战略项目，结果可能与决策质量不同步：一项基于合理概率的项目仍可能失败，一项赌博式押注也可能偶然成功。评价可以同时报告事前决策程序是否符合预定规则、预测评分、执行是否按计划、最终净值和事后学习。四者可以相关，却不能彼此替代。若采用“从失败学习”更新模型，还要保持独立确认时期，避免用一个极端事件立即重写概率，并将回归均值误认成学习成效。

29. 把局部改善与组织净值分开

29.1 预测驱动的资源迁移可能只是成本转移

组织决策常以局部指标衡量：采购看单价，销售看订单，客服看响应时间，财务看库存资金。模型建议若提高某一部门的指标，却将成本转移给其他部门、客户或供应商，不等于企业整体价值增加。分析口径应预先列明受影响主体、同一时间窗、资源从哪里转移到哪里，以及哪些支出是企业内部转账而非新增社会资源。董事会的总收益与部门的可控收益可同时报告，但不能把内部转移的收入和支出都当作独立增益或损失重复相加。

构造一项情境：需求模型预测某系列订单将上升，中央采购通过提前订货节省 12 单位加急费；区域团队因交付时点过早增加 8 单位仓储/折价成本，客户服务因计划与实际订单偏差再增加 7 单位处理成本。若按采购部门账面看是+12，跨职能净变化为 (12-8-7=-3) 单位，尚未计系统或模型成本。若这些成本只是从部门 A 转到部门 B 而公司合并账已包含，则只看合并账的增量现金流和真实机会成本，避免对内部费用转移重复计数。所有数字都是构造值，实际评估应使用可追踪的统一成本口径。

29.2 外部成本和承担者需要单独呈现

公司合并财务报表仍可能忽略客户等待、供应商取消、员工加班或公共安全等外部影响。若某策略降低内部平均成本却增加未按合同补偿的供应商风险，企业净现金流并非社会总价值；两者须按承担者拆开，而不是假定企业收入就是唯

SVSAC | sv-sac.org

一目标。对于外部性难以货币化的结果，可以保留服务时限、严重事件、客户流失和供应商脆弱度等独立约束，避免将其强行折算成未经验证的货币系数。

把多个结果合成一个评分，必须显式说明权重来源和决策权归属。若董事会没有授权把客户等待换算为成本，研究者不应为方便建模自行选择价格；可以报告 Pareto 对照或约束满足情况，并把价值权衡交给有权限的主体。一个无法同时改善所有指标的方案，未必“失败”；问题在于组织是否在看见结果后才改变权重。

29.3 归因需把建议、执行和跨部门后果连起来

若试点中采购单价下降而退货、延期或客服工时上升，只看预测端或部门端无法归因。每项策略需保存建议如何改变订单/项目配置、哪个部门执行了变更、成本落在哪个账本科目、哪些用户或交易受到影响，以及观察窗内结果是否成熟。可以先对可控部门成本按统一口径计算，再单独呈现客户/供应商和质量护栏；若数据不能连到具体行动，结论应限于部门相关指标变化。

有边界的分层随机或阶梯部署可比较政策对组织整体和部门局部的影响，但需按共享客户、预算、供应商与员工队列确定分配层级。若一个部门被随机到新模型，另一个部门仍使用旧流程，却共同服务同一订单，单位之间存在干扰。此时应评估整个端到端工作流或说明局部差异的识别限制，不能把局部随机化标签当作各部门独立。

30. 任务构成变化会制造预测能力的表面变化

30.1 一项可复算的构成反例

模型评估常用简单总体成功率或平均分。若旧模型和新模型承担的任务构成不同，汇总差可能与每类任务内部的变化方向相反。构造两个决策层：容易判断的常规任务和难判断的高不确定任务。旧模型在常规任务的准确率为 90%，在高不确定任务为 60%；新模型在两类分别为 92%和 62%，即每类均提高 2 个百分点。旧模型任务中 80%为常规、20%为高不确定，其粗略总体准确率是 (0.8×0.90+0.2×0.60=0.84)。新模型负责的任务中常规只占 20%、高不确定占 80%，粗略总体为 (0.2×0.92+0.8×0.62=0.68)。原始总表看似下降 16 个百分点，但这不是同一任务混合下模型性能下降。

若按固定的 20%常规、80%高不确定总体重新加权，旧模型为 (0.2×0.90+0.8×0.60=0.66)，新模型为 (0.2×0.92+0.8×0.62=0.68)，差异是+2 个百分点。反过来，若管理层认为高不确定任务的失败代价更高，任务加权准确率仍不

SVSAC | sv-sac.org

等于决策净值；还需使用每类的错误成本和行动规则。所有数值是明确的构造例，不代表模型实测，也不证明精度差异会转化为收益。

30.2 固定权重是目标总体定义，而非技术修正

标准化加权只能回答“如果两模型面对同一任务组合，分数会怎样”，不能替组织决定下一年任务构成，也不能消除某些任务没有标签或超出模型支持范围的差异。权重应来自决策总体、预算配额或具有授权的业务组合，而不是从效果较好的分组反推。若真实策略的目的本来就是把模型集中到高价值的新任务，那么实际业务效应应在新分配政策的全部合格机会中估计；固定混合权重的模型比较则是补充的预测能力分析。

不同价值层的项目可能也有不同标签成熟度。常规任务在 30 天内闭合，高不确定项目数年后才确认，完成案例的平均分不仅受任务难度，也受观察窗口影响。应把模型分数和策略结果同时按事前层级呈现，报告层内样本数、未决比例、风险权重和标准化方法。没有重叠的任务类别，不能靠加权创造反事实比较；任何跨类别外推都要明确模型假设。

30.3 评分目标变化会产生“优化了不该优化的量”

若团队以平均准确率为奖励，可能偏向大量低成本常规问题，而忽略罕见高影响决策；若以严重错误率为唯一指标，又可能鼓励对所有高风险事项拒答。目标函数同时面对任务数量、损失大小、拒答和复核成本时，应预先定义多项结果及其优先序。一个预测系统可通过拒答降低可评分错误，却把大量困难任务交给昂贵人工；若只对有输出样本计分，系统会在统计报表上变好，却可能令整个流程更慢。

这类差异需要两个分母：模型产生可用预测的子集及全部分派机会。对前者报告预测质量与拒答率，对后者报告行动覆盖、转人工、等待、成本及净值。若模型拒答的原因与事件风险有关，完整案例评分并非随机子样本，不能直接代表全部业务任务。拒答可能是合理安全策略，但其价值取决于替代流程和错误代价，而非拒答率高低本身。

31. 候选行动越多，胜出预测越需独立验证

31.1 预测分数挑选候选，会产生选择性乐观

AI 系统可能在同一目标下生成许多市场、投资或合同选项，再用预测模块逐一评分，最后推荐模型估值最高者。若候选真实净值为 (v_j)，模型输出为

SVSAC | sv-sac.org

($\hat{v}_j=v_j+\varepsilon_j$)，即使所有候选的真实净值都相同，选择 ($\hat{v}_j$) 最大的选项也会偏向正误差。候选越多，在同一评价数据上筛选出的最大预测值越可能过度乐观。它不是说生成多个方案一定有害，而是筛选与确认若共享同一信息，最终被选项的“预测优势”含有选择成分。

构造一组四个方案，其估值误差分别为+8、+5、−2、−6，真实增量价值均设为 0。若只按同一组估值挑第一项，报表会显示+8，但实施后的真实增量仍为 0。该简单例不估计任何误差分布，只展示条件选择机制。真实研究应区分候选生成数量、筛选规则、模型/人工修改和独立结果验证；若方案只在预测模型内部评分而未执行，无法确认被选项的反事实结果。

31.2 确认集和执行分配的不同作用

一种可行办法是把任务数据分成开发与确认两部分：第一部分可用于生成候选、调阈值和选择策略；确认部分在策略锁定后才用于评估预测分数。若候选行动会影响真实结果，单纯留出集仍不能识别未执行方案的潜在净值；还需随机分配可行行动、对适度风险方案作受控试验，或报告识别界限。预测确认解决的是泛化误差，执行分配解决的是政策因果结果，两者不能混为一件事。

如果选项具有不同成本、期限和失败风险，确认分析也要按预注册的业务目标报告，而非只保留候选中最有利的指标。生成系统可以增加候选覆盖，模型打分可以提高筛选效率；最终决策依旧需要可执行性、责任主体、资源容量及未采取动作的基准。AI 提出“新方案”与 AI 预测某方案结果是两个功能，证据链也不同：提案数量或新颖程度不能替代对收益、合规性和实际执行结果的验证。

31.3 未实施的候选不等于零价值，也不等于已创造价值

没有选择的候选方案具有潜在期权价值，但不能将其预测估值记入财务收益；它也不应无条件视为零。若候选可低成本保存，未来获得新信息后可能重新考虑；若它会占用专利、谈判或预算窗口，延迟本身可能使价值消失。应在决策日志中标记未选的原因、可逆期限、重新进入条件和被选方案的机会成本。只有在明确的后续政策下，未选候选才能进入新的测试或投资比较。

对被拒选项的效果估计不能由“最终选择了它”作为成功标签。若管理层根据预测选择单一方案，之后只观察该方案，未执行方案的结果依然缺失。候选生成可能扩大了策略空间，却同时加剧了历史选择性标签。要验证新增方案的价值，必须登记被检验的候选总体、筛选政策、实际试验单位和后续结果；否则可说系统扩大了提案集，但不能说提案提升企业价值。

SVSAC | sv-sac.org

32. 把不可确定的权重留给决策者

32.1 多目标结果不应被伪装成单一客观分数

战略选择往往同时影响现金流、风险、能力积累、客户关系、合规和员工负担。模型可以预测某些结果的概率或数量，却不能从数据自动推得组织对这些结果的偏好。把各指标标准化再求和，需要选择尺度、权重和时间折现；这些选择会改变方案排名。除非这些参数由有权限的决策者授权并在事前写明，所谓综合值只是分析者的规范选择，而非模型发现的事实。

32.2 构造的 Pareto 比较

设甲方案的现金贡献为+12、风险损失为 8、长期能力指标为 2；乙方案分别为+8、风险损失为 3、能力指标为 1。若只看现金，甲更优；若风险损失权重足够大，乙可能更优。若能力指标并未被可靠货币化，就不能为它随意加一个系数并把最终分数称作净现值。更透明的做法是呈现各方案在每个目标上的数值或范围，标出被其他方案全面支配的选择，并将不支配的选择留给明确承担结果的治理层。

32.3 价值权重的敏感性应形成行动界限

可以对有授权的权重范围逐一重算方案排序，报告当某个权重越过何值时选择发生改变。若选择对很小的权重变化就翻转，结果并非一个稳定的模型推荐，而是偏好或估值假设主导。决策者应决定能否接受这类不稳健性，或是否值得购买额外信息。增加预测精度可能不解决争议，因为争议来自价值权重而不是概率估计。

所有权重敏感性都应区分数据估计和规范设定。项目成功概率可由历史、试验或外部证据支持；“客户关系相当于多少现金”则通常是组织偏好或会计口径。若没有可信估计，使用多个权重情景更诚实，但各情景仍是决策假设，不能把“方案在大多数情景胜出”当成经过频率验证的随机概率。

32.4 谁承担风险，谁批准权重

风险指标和收益指标由不同部门承担时，统一期望值可能遮盖责任分配。业务单位获取增长收益，财务承担现金流波动，法务/风险承担违规后果，员工承担执行工作量。将权重写进目标函数之前，应说明由谁授权、谁能修改、谁可否决，以及损失落在谁的账上。若关键主体没有参与，模型的最优方案可能只是在既定组织边界内最优。

SVSAC | sv-sac.org

对外报告宜保留每一项核心结果和决策约束，而非仅给一个模型推荐等级。若管理层仍决定使用合成分数，需保留权重版本及批准记录，并在权重变化时重新计算历史结果。这样可以判断政策改变是因为预测更新，还是因为组织价值选择发生变化。

33. 组织激励会改变概率报告的含义

33.1 适当评分的诚实报告条件有限

严格适当的评分规则在特定条件下使报告者在期望评分最大化时如实报告概率[1]。这个结论讨论的是预测分数的激励，不是说企业绩效考核天然会促成诚实的经营判断。真实管理者还在意预算是否获批、项目是否推进、个人考核、声誉、团队关系以及失败后的责任。这些效用若与报告分数共同决定个人收益，单纯使用 Brier 或对数分数无法保证管理者报告其真实信念。

构造一个二元决策：某项目获批的阈值为 0.60，负责人真实认为成功概率为 0.55。若其唯一报酬是严格适当的 Brier 分数，偏报到 0.60 的期望损失相对于报告 0.55 增加 $((0.60-0.55)^2=0.0025)$。若夸高分数能让项目批准并给负责人带来 0.01 单位的私人收益，且拒批无此收益，则该负责人总效用偏好报 0.60 而非真实 0.55。数值仅说明激励冲突的构造，不能表示真实经理的效用。项目批准后实际成功率也不验证这次报告是否诚实，因为这一个结果同时受随机事件和批准动作影响。

33.2 预测评分、项目决策与人员考核不宜捆成一项分数

组织需分别设计预测反馈、项目审批与个人考核。若同一概率报告既决定项目获批又影响提交者晋升，可能诱发战略报告、回避难题、过度保守或风险转移；方向须由实际激励条件判断。本篇未核 Mellers 等的报酬与排行榜对应原段，撤回具体参与奖励的事实归属，不用另一任务的支付形式证明企业激励安排有效。

一种设计选择是由独立预测者在不知道个人审批结果奖励的条件下封存概率，再由另一个授权角色决定行动；另一种则允许同一决策人提交概率，但需记录私人激励及决策过程，并比较激励制度。这样的分离会增加协调成本，未必总适用。评估时应检验的是哪种激励结构让概率更校准、动作更可取且责任更清晰，而非默认“打分”就解决了激励问题。

SVSAC | sv-sac.org

33.3 可观察的偏差需要事件级审计

检查组织概率报告是否受激励影响，可把预测时间、结果揭示时点、预算审批、提交人奖励、概率修订和实际行动逐项连接。需区分报告者修订概率是因为获得新信息，还是因为审批阈值、奖金或问责改变。只按年度平均校准可能掩盖：某些高额项目普遍高报成功率，而小项目报得准确；或者预测在有奖金周期内发生系统变化。

若样本不足以对人员个体校准作可靠比较，可以先审计概率分布、固定阈值附近的报送行为和审批前后修订轨迹，不把小样本个体排名做成人事结论。模型与人员可能共享错误信息；如果两方提交同一个过度乐观数值，不能据此判定他们相互验证。预测源头、信息来源、被授权动作与最终结果须分开保存，才能定位偏差究竟来自数据、模型、报告者还是规则。

34. 参考类不是历史成功案例的平均值

34.1 类比样本选择本身就是一项判断

为战略项目设定基准率时，最容易找到的案例常是已经完成、资料较齐全或被公司公开的项目。这些案例不一定代表当初全部合格机会。被取消、失败、未上市或无法取得后续财务结果的项目若被排除，历史基准会偏向幸存者。参考类应从决策前的机会总体出发，写出纳入标准、时点、目标事件和观察期；否则“过去类似项目”只是一组按事后故事筛选的例子。

作为待检验的培训安排，可要求预测者列出参考类、既有估计与基率，再解释为何适用于当前项目。这是我们提出的程序，不宣称已逐项核实 Mellers 等训练模块或效果。企业还需分别检验参考类选择、校准、行动边界及净值；概率训练改善与项目利润改善是不同问题。

34.2 一个幸存者过滤的构造反例

假设某公司历史上共启动 40 项扩张计划，其中 10 项实现预定目标，10 项被取消但在审批时满足过同一项资格条件，另 20 项仍在观察窗内未成熟。若经理只找到 10 项已完成且表现好的案例，观察成功率为 100%；若仅用已成熟并已关闭的 20 项，成功率为 (10/20=50\%)；但这也不能说全部 40 项的最终成功率是 50%，因为另 20 项尚未成熟且取消项目的目标结果定义需明确。为了比较新项目，必须规定取消的原因、取消后是否仍可能达到目标、未成熟标签的处理、参考类资格和观察窗。

SVSAC | sv-sac.org

若暂把 10 项已取消计划视作终端失败，且 20 项未成熟计划全部视作失败，完整总体成功占比为 25%；若这 20 项最终全部成功，则占比为 75%。如果被取消计划即便继续也可能成功，那么“取消”不是目标事件本身，25%至 75%的区间也不再是完整识别界限；没有关于取消后潜在结果的假设或额外证据，无法仅凭现有标签得到该界限。构造说明标签边界会改变基准区间，而不是主张上述比例适用于现实企业。若缺失和成熟机制未知，应报告可识别范围及收窄它所需的外部数据，而非用模型插补掩盖参考类选择。

34.3 参考类与当前项目之间的迁移差异

候选项目与历史参考类可能在宏观周期、团队经验、产品成熟度、融资条件、竞争强度和监管期限上不同。应先定义可能影响成功率的差异，再显示有共同支持的历史子群或采取分层模型；当新项目与所有历史案例差距过大，恰当输出是“数据不足以支撑概率”，而非更细的小数。分层越细，方差越大；合并越多，越要求项目在关键机制上可交换。两种错误都可能发生，需用事前可辩护的层级和敏感性分析呈现。

模型从参考类中学到的先验概率可以与新项目证据结合，但更新幅度取决于证据质量。若专家判断、市场测试与历史数据来自重叠样本，就不能把它们当独立证据多次相乘。说明来源相互依赖、信息截止和重复引用，是防止伪造高信心的重要步骤。参考类帮助组织校正内部叙事，却不能消除未来状态变化或替代政策反事实。

35. 不同预测分歧何时可以平均

35.1 预测数值必须针对相同事件与条件

两个团队给出 0.30 和 0.70 时，算术平均 0.50 只有在二者对同一事件、同一时间窗、同一条件、同一总体表达概率时才具有直接含义。如果一个数值条件是供应合同已签、另一个条件是合同未签，二者是不同条件概率而非两份对同一事实的独立意见。未声明条件的平均值既不是已签条件的概率，也不一定是当前项目的边际概率。

用明确假设可计算边际值：假设供应合同签署概率为 0.40，事件在签约条件下的概率为 0.30，在未签约条件下为 0.70，那么整体概率是 (0.40×0.30+0.60×0.70=0.54)。这与直接平均两数得到 0.50 不同；分歧主要来自条件混淆，而不一定意味着某个预测者质量差。所有数字均为构造假设。只有签约概率本身可靠、分支穷尽且条件定义相容时，0.54 才是该结构下的边际预测。

SVSAC | sv-sac.org

35.2 集合与条件差异先于聚合方法

组织内的模型、专家和外部情景分析可能有不同信息截止、事件定义和参考类。汇总之前需检查预测对象是否能统一，事件是否互斥且完备，概率是否对应同一时间范围，参与者是否共享数据源，误差是否相关。若事件定义不同，先按明示条件重写预测问题；若依赖条件未知，则保留分歧而非机械平均。模型得分更好也不代表它可替其他信息条件下的预测直接背书。

聚合可采用简单平均、表现加权或结构化集成，但权重需从前期数据确定，并在独立确认窗口评价。权重若在同一任务上反复挑选到最高分，会出现模型筛选偏差；若历史表现依赖某类任务，也需说明迁移权重。预先指定简单基准有助于判断复杂聚合是否值得其维护成本。团队讨论、模型权重与问题定义同时改变时，增益仍属于整个策略包。

35.3 先保存分歧，再决定是否合并

在最终决策前保留个人、模型与团队汇总的分布，有助于后续识别信息来源、分析不确定性和复盘错误。若只留下最终统一值，无法判断共识是基于新证据、信任层级、强势观点，还是对一个共同提示的重复响应。对关键项目可保存概率分位、主要争议条件、证据来源和分歧解决方式；但保存细节本身有隐私、商业保密和时间成本，需按决策价值选择合理粒度。

若团队意见分歧大，继续取平均不一定是最佳行动。分歧可能揭示还未定义清楚的事件、遗漏的备选方案或双方承担不同的损失函数。可将其转化为待验证问题：哪一项信息能区分两种条件？在该信息到达前，是否存在可逆的小行动？如果价值函数本身不同，则需要授权者做明确取舍，而不是期待更强的预测模型替代组织选择。

36. 成功概率与收益规模必须联合进入决策

36.1 边际成功率相同，现金价值仍可能相反

经营预测常把“成功概率”当作行动比较的核心数字，但项目的现金贡献通常还取决于成功时的规模、兑现时间、成本承担者和失败后的残值。若模型只预测二元成功事件，决策表却将不同市场状态下的成功都视作同一价值，隐含了成功价值与状态无关的假设。这个假设在固定价格、固定规模、统一回收期的标准任务中可能近似成立；在战略项目里，容量、价格、汇率、供货能力和竞争反应常同时改变成功概率与成功价值。

SVSAC | sv-sac.org

一个构造例可说明其影响。假设项目面对高需求与低需求两种状态，各自概率均为 0.5；高需求状态下项目成功概率为 0.8，成功贡献 100 单位；低需求状态下成功概率为 0.2，成功贡献 20 单位。一次性投入成本为 30，失败不回收。联合期望毛贡献是 0.5×0.8×100+0.5×0.2×20=42，故净期望为 12。若先把状态平均，得到边际成功概率 0.5 和成功贡献均值 60，再相乘则毛贡献只有 30、净期望为 0。差异来自概率与成功规模的正向状态关联：$E[p_sV_s]$不等于 $E[p_s]E[V_s]$。所有数值都是构造假设。

36.2 均值替代联合分布会丢失损失结构

上例的决策差异并非证明应采用某种复杂模型，而是指出需要估计的量应与合同和现金流结构相符。若成功价值实际由市场状态决定，至少需要记录每个可观察状态下的项目结果，或明确把价值范围设为情景输入。若数据只包含历史批准项目，且批准本身由高预期收益驱动，成功价值分布还会受选择影响；批准样本内的均值不是所有候选项目的可比收益。

如果需求状态与成功概率负相关，粗略乘积也可能高估收益。例如高需求时成功概率很低，低需求时成功概率较高，但低需求的成功贡献有限，平均概率仍可能掩盖结构。战略决策需要的是净现金流的联合分布或至少能覆盖它的界限，而不是单独的成功率、单独的订单额和一个事后相乘的平均值。预测团队应把概率定义到明确的事件、时间与条件，再由财务或业务负责人给出各条件分支的价值口径；两类不确定性不可混成同一“模型置信度”。

36.3 相关性不是可供优化器自行修复的噪声

若成功概率与收益金额来自不同系统或不同历史期，分析者可能默认误差彼此独立。实际上，旺季、客户信用、供应商产能或法规变化会共同影响成功机会和兑现规模。仅在历史数据里把概率分数与平均金额相乘，未必能估计未来项目的条件价值。相反，若低成功率状态发生时损失更大，平均值可能掩盖下行尾部，即使总体期望相同，承担者也可能拒绝该方案。

这种联合结构可以按决策相关的分层进行检查：先声明状态如何观测、状态间互斥与否、金额是毛贡献还是扣除新增投入后的净值，再比较分层平均、总体平均和不利情景。没有足够数据时，报告区间或情景表，并说明哪些分支由历史数据支持、哪些是管理层假设。优化器只会在给定目标函数和输入下寻找解，不能替组织发现未被数据记录的状态相关性。

SVSAC | sv-sac.org

37. 共享资源的影子价格会改变预测阈值

37.1 可行预算之外，还存在被挤出的用途

一个行动的直接增量收益为正，并不说明它值得执行。若它占用的工程师、法务审核、试验台、供应额度或管理层注意力已满负荷，真正的比较对象是被挤出的下一项可行用途。反过来，若这些资源在相关窗口闲置，不能因为内部会计把固定工资分摊给项目，就断言每增加一个项目都承担了同额机会成本。影子价格取决于时间窗、容量约束和可替代任务，不是永久不变的部门费率。

设一个项目的预测净贡献（尚未扣稀缺资源的机会成本）为 14 单位，需要工程资源 6 小时；同一窗口里最好的替代任务每小时可实现 1.5 单位净贡献。若六小时确实挤出替代任务，项目的增量比较值为 14−6×1.5=5；若该工程资源在窗口内有闲置且没有其他可执行任务，短期机会成本可能为 0，比较值是 14。两种计算都可能合理，关键是说明资源是否真的受限、替代任务是否可执行以及被比较的日历窗口。数字仅为构造示例。

37.2 预测阈值取决于资源边际价值

对需要稀缺资源的二元行动，简单阈值“预期收益大于行动现金成本”通常不完整。若成功概率为 p、成功净收益为 V、失败损失为 L，行动直接成本为 C，所需资源为 h，资源影子价格为 λ，则增量值可写为 pV−(1−p)L−C−λh。若 V+L 为正，行动阈值为 p>(L+C+λh)/(V+L)。λ 上升会提高行动门槛；释放容量则可能使门槛下降，即便模型概率没有变化。

这个公式不能把 λ 当成系统自动给出的常数。λ 需要由容量边界附近的可替代项目或管理者明确的稀缺资源估值支持。若被挤出的项目会随预测政策改变，替代用途本身是反事实，需对整套组合比较而不是把某个历史小时单价机械套用。还要防止重复扣减：若 V 已经减去了替代项目损失，就不能再在 C 或 λh 中扣同一机会成本。只有统一口径后，概率阈值才有可解释的经营意义。

37.3 约束松弛和价值权重应分开报告

管理者可以通过增加预算、外包或延长窗口改变容量；这些选择有成本，也会改变其他项目时序。若预测方案只在资源紧约束时排名不同，追加容量可能让模型差异转化为实际可执行动作，但追加容量的决策需要单独核算。若资源扩容使所有方案都可执行，则原来的“高分项目优先”政策可能不再必要。若扩容无法及时到位，预测分数提高也可能只改变候补名单，不改变当期结果。

SVSAC | sv-sac.org

有用的敏感性报告可以同时呈现：资源未受限、基准受限、进一步收紧三种情景下的可行集合、被挤出的具体活动、项目净值和授权主体。无法可靠估计机会成本时，应把它留在不同的敏感性设定中，而不是用一个看似精确的影子价格掩盖不确定性。决策者仍需承担对容量和价值排序的授权责任。

38. 预测能力的跨期投资要按未来决策机会计价

38.1 一次项目净值与学习资产的价值不同

为当前项目购买数据、开发模型或完善记录，可能使当前行动不变，却改善以后多个相似决策。此时收益对象不是单一项目，而是未来任务流中的政策表现。把全部开发成本压在一个项目上可能低估可复用学习；把成本无限期摊给未来项目又会高估复用，尤其当目标市场、产品或模型发生变化时。需要清楚区分一次性建设成本、每期维护成本、复用次数、更新成本和预期失效风险。

构造一个有限期限例子：建设和首轮验证成本为 12 个价值单位，预计未来三期各有一项可比决策；若每期该方法相对现流程平均增量为 5，维护及更新每期成本为 1，且每一期都能使用，则不折现的三期净价值为 $3\times(5-1)-12=0$。若只能复用两期，净值为 $2\times4-12=-4$；若每期增量提高至 6 且三期都适用，则净值为 $3\times(6-1)-12=3$。该账本只在增量值来自可比较的未来决策、各期成本不重叠且没有容量挤出时成立，不是企业回报预测。

38.2 复用率不是历史使用次数

系统日志记录了工具被调用多少次，不等于这些调用形成了有价值、可比的决策机会。重复尝试、被拒绝建议、没有最终结果的预测和由工具提醒才创建的任务都可能进入调用量。未来机会分母应由事前定义的合格决策总体构成，并记录工具是否可用、是否被授权使用、结果是否成熟。否则按调用次数摊销会把持续试用误报为生产性复用。

如果未来目标从新产品成功转为合规风险、供应连续性或现金回收，既有训练和验证结果可能不能迁移；按原任务次数估计的摊销分母便失效。复用假设应逐项说明哪些特征保持稳定、何种变化触发重新验证、数据与模型维护由谁承担。没有证据说明任务可比时，可报告“若复用 n 期”的盈亏平衡表，而不把期望复用次数写成已观测事实。

SVSAC | sv-sac.org

38.3 学习价值需要独立于试验参与者收益评估

试点本身可能改善流程，因为员工得到更多关注、决策被管理者复核或项目数量暂时变少。这些作用可以有价值，却不能自动归因于预测模型。若试点同步训练团队、调整审批权限、增加资源并更换模型，当前净值变化属于组合处理。想检验可复用模型能力，应保留能区分模型组件、培训、管理注意和任务筛选的比较条件，并明确不同条件在后续周期是否仍有效。

未来多期效果也可能因学习而增强或因环境变化而衰减。若只在最初几个周期估计平均增益，然后假定永久复用，就忽略了非平稳性；若看到一次负面更新便停止，又可能将随机波动误当系统失效。合适的报告应给每期成熟结果、维护成本和适用域变化，比较累计现金流而非只比模型分数。长期学习资产的价值是一项要由持续证据更新的预测，不是开发完成时就已实现的收益。

39. 组织现场试验显示，决策变化与解释质量可以分离

39.1 分配单位和结果单位不是一回事

Kang 与 Kim 在一家投资研究公司对共同基金分析师开展六个月现场试验。作者将 97 名分析师按基金类型和任职年限分层随机分配：48 人进入接收机器评级预测的处理组，其中一部分还看到基于 Shapley 值的特征说明；其余分析师按原流程工作。工作稿的数据节将 1,780 次决定列于 2021 年 4 月 6 日至 2022 年 1 月 19 日，而试验节将其连写于 7 月 18 日开始的六个月干预；两处期间口径不一致，本篇不把 1,780 全部认定为干预期新增决定。标准误聚类在分析师层级，符合实际随机分配单位。由于一个分析师反复评估多个基金，1,780 条决定不能被视作 1,780 个独立随机单位。公司审核委员会不知道分组，作为解释报告的盲评来源。该研究的作者工作稿标注为 2025 年，本篇未重新核实其后接收状态；本篇据该工作稿讨论，不对排版后的期刊最终版作逐页复核[10]。

39.2 更频繁改变评级不等于平均投资决策已经改善

处理组更可能改变既有评级，处理效应约为 8 个百分点，基准组约 11%的决定发生评级变化。这个结果说明展示预测改变了分析师的行为，但行为变化本身不等于净值提高。作者用后续风险调整基金收益衡量推荐决策的表现，平均推荐基金收益差约 0.89 个百分点，统计上不显著（p=0.77）；解释评分的平均处理效应也不精确，估计区间较宽。因而该实验不能支持“机器预测普遍改善投资结果”的强结论[10]。

SVSAC | sv-sac.org

作者报告了异质性：较简单基金的推荐收益差异约为+9.9个百分点（p<0.001），复杂基金约为−16.8个百分点（p=0.002）。这些是按基金类型划分的条件性子组估计，不能替代总体处理效应或投资者完整组合价值。解释评分也随决策类型和分析师经验出现不同方向的变化，尤其新分析师在较简单决策上的解释评分可能下降，在复杂决策上可能上升。这些子组结果不能代替平均效应；更不能将一个工作场景的“简单/复杂”分类直接迁移到其他行业的任务难度[10]。

39.3 可见的机制证据不是被单独随机化的机制

该研究的一个重要增量是同时测量评级变化和分析师写出的因果解释。算法预测更容易改变决策，却可能使新手更难说明评级为何改变；作者据此提出，分析师可能在采纳预测和维持自己能解释的判断之间权衡。处理还改变了报告文本，作者发现算法分配条件可由报告语言分类。这些观察支持“人做得出动作但不一定形成理解”的机制假设，却没有随机操纵某位分析师对预测的依赖，也没有直接识别长期专业能力是否下降[10]。

解释界面亦不是普遍解决方案。附加的Shapley特征图在该设计里属于独立处理臂，但它展示的是对模型预测的特征贡献，不等于因果机制解释。研究没有发现它在所有结果上都改善决策或解释，且可解释性可能改变使用程度。因而“给出解释”应拆为可检验的呈现组件、使用反应和最终结果，不能把可视化特征重要性当作真实因果解释的替代物[10]。

39.4 这项证据的迁移边界

该随机试验比单纯的模型基准更接近实际组织流程，因为预测被插入分析师工作界面，决策和解释均有结果记录。但其对象是一个投资研究公司的互基金评级体系，而不是投资审批、一般战略项目或所有企业预测；结果单位包括后续基金风险调整收益，部分收益观察因缺少完整36个月历史而不可计算，解释评分也因公司在2022年初停止收集而只有较短窗口。研究依赖一个组织的评价委员会标准和既有分析师工作流程，无法证明任何行业都会出现同样的权衡[10]。

因此，组织试验若只测决策准确或行动变化，可能漏掉解释、学习和未来执行所需的理解；若只测说明质量，也可能忽略错误风险是否下降。可复制的比较应预先确定决策和解释的独立估计对象、随机分配的人员或团队单位、成熟结果窗口、子组假设与多重比较处理，并把预测展示、解释界面、培训和授权变化分开。对管理层而言，这项研究更适合作为可能存在双重结果的反证，而非一个能直接套用的绩效系数。

SVSAC | sv-sac.org

40. 预测支持的长期价值需要分开衡量当前选择和能力形成

40.1 一个短期动作可能与组织学习方向相反

第 39 章的研究提示，组织可能同时获得更频繁的决策调整和更弱的部分解释表现。二者没有天然的统一货币尺度。若当前目标是快速发布评级，行动时效可能优先；若分析师还需要向客户解释、识别数据漂移并在未来新基金上独立判断，理解能力也可能是生产资产。将解释评分简单乘上一个任意系数，不能解决管理者对当前准确与长期能力的价值权衡。

一种可检验的假设是：预测辅助对资深人员的短期判断增益更大，而对新手的学习影响更需观察。另一种相反解释是，差异来自现有经验决定了信息使用方式，而非辅助削弱学习；也可能是评估委员会对资深员工的解释宽容度不同。判断这些机制需要重复观察新人经验增长、后续独立任务和错误修正，而不是仅比较单个周期的报告长度或评分。能力形成若要进入经营价值，应定义可观察的任务迁移、接替成本或独立判断质量，不应以“员工会不会 AI”这类模糊指标代替。

40.2 如何将能力指标与经营决策分开记录

一套预先设定的评价至少要区分三个层次：当前任务的决策结果、提交者能否给出与证据一致的理由、以及其后续独立任务表现。当前任务的主要结果应按随机分派保留，包括未采纳预测或没有改变原建议的情况；分析者只比较“听从 AI”的个案会筛选出更愿意使用建议的人。解释结果应盲评并保留评审者一致性，文本长度和特征数量只能是辅助指标，因为短报告可能表达清楚，也可能只是省略推理。

后续独立任务则需避免把培训、经验增长、任务难度变化或人员流失误作工具影响。可事前冻结任务复杂度分层，在处理组和对照组持续跟踪新任务；若处理仅对曾见过的基金有效，就不能称为可迁移能力。需要团队或分析师层随机分配时，还要防止共同基金、共享评审或预测讨论造成信息污染。这里的设计是作者提出的验证方案，不是 Kang 与 Kim 实际执行的全部试验内容。

40.3 对解释质量的经营估值应从决策需求反推

解释并非在所有工作中同等重要。法律授权、客户信任、跨部门执行、合规复核和接班培训，可能要求负责人说明判断依据；高频自动执行且可以快速回滚的任务，对个体事后叙述的要求可能低得多。应从真实决策需求推导解释指标，例如审批者能否发现关键前提错误、同事能否据理由执行、客户能否理解产品风险，而不是把研究中的委员会评分照搬为统一 KPI。

SVSAC | sv-sac.org

可用构造性的损益账本来明确取舍：若辅助提高当前准确度带来的预期收益为 A，而后续独立判断质量下降造成的预期代价为 B，净值变化取决于 A−B 及测量范围。A 与 B 均需在相同决策总体、时间窗口和成本口径下估计；如果无法估计 B，可以将其作为阈值或敏感性情景，而不是假定为零。短期数据不足以确定长期损害，也不能反过来证明长期损害必然存在。可说明的是组织必须决定是否将学习和解释纳入目标，而不是本次分析已发现了通用折价。

41. 将工具展示效应与实际政策价值分开估计

41.1 随机分配识别的是展示预测的意向处理效应

在第 39 章所述研究中，随机分配单位是分析师，而不是每一条评级或每一只基金。作者将 97 名分析师按专业基金类型和任职年限分层，分配到常规流程、显示机器评级预测、显示预测并提供 Shapley 特征图三种界面条件；1,780 为工作稿报告的观察数，其期间口径差异见第 39.1 节，不能全部视作六个月干预后的决定。按分析师聚类的推断保留了重复决定共享同一分配状态的事实。其核心估计量是“被分配到展示预测”的平均意向处理效应，而不是“每次实际依赖预测”的效果[10]。

如果希望估计实际使用建议的影响，直接比较采纳和未采纳分析师会产生选择偏差：较新手、任务更简单或预测更符合既有信念时，采纳概率可能不同。随机展示可以识别提供系统这一策略包的效果；若进一步把分配当作实际采用的工具变量，仍需排除限制、单调性等额外条件，并定义哪些人的采用变化由展示诱发。该研究并未把每次依从性当作随机处理，也不能将意向处理结果改称为每次采纳的因果效应[10]。

41.2 被推荐基金的收益差并不自动等于整套推荐政策的价值

分析师在不同试验组可能推荐不同基金。比较“处理组被推荐基金的后续收益”和“对照组被推荐基金的后续收益”，回答的是两个组各自推荐集合中的条件性结果差；它不是固定基金总体上按两套推荐规则投资的完整组合价值。用符号表示，$E(R|T=1,\text{Recommend}=1)-E(R|T=0,\text{Recommend}=1)$中，条件 Recommend 本身可能受 T 影响，两个条件集合未必由相同基金或相同权重组成。

若经营问题是投资者按推荐规则配置资本，应在决策前固定可选基金总体，定义每种政策的分配权重、现金或不投资基线、持有期、风险调整方法、交易费用与再平衡规则，再计算 $V(T)=E[\Sigma_i w_i(T)R_i-\text{cost}(T)]$。作者还比较了基于分析师决定与机器预测生成的假设组合，并未发现其表现差异具有统计显著性。该结果限

SVSAC | sv-sac.org

制了把局部推荐收益关联直接描述为“机器政策改善了投资者净值”；但不显著也不等于两种政策完全相同，精度和经济意义区间仍应报告[10]。

企业投资选择中的结果单位更难统一。一个项目的收益可延迟多年，实施动作还会改变其他项目的机会成本，无法沿用基金收益窗口或单只资产的回报指标。跨项目比较须把现金流、占用资源和未投项目的机会价值放进相同的策略级账本；如果无法观测反事实，须明确哪部分只在模拟或敏感性情景中。

41.3 平均结果、子组结果与“解释机制”属于不同证据层

一项处理可能对平均决策结果接近零，却在事前定义的任务类型或经验组中方向不同。此时不能只引用最显著的子组，也不应因总体不显著就抹去有理论依据的异质性。试验应事前列出有限的关键异质性维度、对应交互估计、结果窗口和多重比较处理；将连续经验保留为连续变量并做敏感性分析，可避免通过多个经验阈值寻找最有利结果。

第 39 章研究对理由质量的主要平均估计区间较宽；部分评分只在公司停止收集前可得，风险调整收益也受基金历史数据是否完整影响。即使组间缺失比例接近，结果构念和观察时点仍有限。被评委打出的解释分数可以反映该公司的报告标准，却不是跨企业通用的人才能力量表。算法说明图是特征贡献展示，不等于对预测为何有效的因果说明；观察到使用率变化，也不能单独证明学习或理解发生变化[10]。

因此，机制论证应保持层级：随机展示识别界面处理对行为和观测结果的总影响；有关信任、认知负担或技能积累的解释需要进一步测量或单独操纵。报告文本变短、使用特征图、改变评级可作为过程结果，但不能直接充当长期经营收益或机制中介的已识别效应。

41.4 组织迁移试验需要重新定义“更好”的结果

互基金研究可迁移的不是 9.9 个百分点或 0.89 个百分点，而是它把预测展示嵌入真实工作界面、随机分配在分析师层、同时观察行动与解释的研究结构。新的行业必须重新定义结果：采购预测应查看采购总成本和缺货风险，战略投资应查看成熟现金流和组合风险，医疗或合规决策还涉及受影响主体与可接受错误。若一个端点是准确度，另一个端点是可解释和可执行性，需并列报告，不宜以未经授权的权重合并[10]。

随机化单位应与信息分享的现实边界一致。若同一部门成员共享模型输出、会议或客户资料，个人层分派可能把对照组污染；可考虑团队或业务单元分配，但有效样本数随集群数量而非单条决策数增长。项目过少时，报告可识别精度与

SVSAC | sv-sac.org

限制，不应把大量重复记录误当作大量独立实验单位。是否等待更大样本、采用阶段性推广或只评估流程成本，取决于行动风险和可逆性，而非统计显著性单一门槛。

最后，策略包必须描述为实际被操纵的内容：预测模型、展示位置、说明形式、训练和决策权是否同时改变。要拆出模型本身的贡献，需要相应的因素对照；若只是比较“旧流程”与“新界面加模型加培训”，结论就应属于整个实施方案。这样既能避免把组织投入全记到算法名下，也能避免把预测系统狭义化为一个离线分数。

42. 决策建议：预测分数只有嵌入可执行政策才构成决策支持

42.1 对核心问题的回答

本篇的问题是，预测能力在什么条件下会改变行动并提高经营净值。分析显示，准确率或概率评分只是链条的一环：预测对象必须对应决策时点的可用信息；预测概率须与行动后的结果和价值口径相接；可选行动、资金、人力和审批权限须允许行动发生；被改变的结果需在可比且成熟的总体中观察。若失败结果在未批准项目上不可见，或收益金额与成功概率随市场状态共同变化，单独校准成功概率仍不足以评价政策。

因此，预测更准并非经营增益的充分条件。价值可能来自改变资源分配、减少预测和审核成本、支持可逆试验或提升未来重复决策能力；但这些通道的估计对象不同。若行动未变，仍可有流程成本价值；若行动变了，仍需证明处理后的结果没有被选择、成熟周期或其他同步改动所混淆。一个模型分数无法替代政策比较、现金流账本和责任授权。

42.2 有条件的产业启示

对企业而言，部署决策支持前，先说明预测将在哪项决策、由谁在何时使用、哪些行动可执行、什么结果代表成功以及谁承担失败成本。若模型只用于提供第二意见或缩短资料整理时间，应按该流程目的设基线；若声称会提高投资、客户或运营结果，则需要固定机会总体、保留旧流程对照并跟踪足够成熟的结果。人员建议、评分、批准和实施不能混为同一处理。

对于多目标战略选择，应明确由有权主体批准价值权重和风险阈值；当资源容量受限时，计算被挤出的可执行用途，而非只看项目自己的毛收益。对于重复式决策，基础设施成本可按合格未来机会做情景摊销，但不能把试用调用数当作

SVSAC | sv-sac.org

收益分母。对预测辅助可能影响知识积累的场景，还应把短期动作表现与长期独立判断分开评估。

42.3 证据限制与后续检验

本次分析是范围限定的文献梳理与分析，不是系统综述或企业实证。已读材料与待核线索分别涉及概率评分、预测竞赛、算法采用实验、人机协同汇总、司法决策模拟、未来事件预测基准、库存预测模拟和共同基金分析师现场试验；这些来源的任务、结果和研究设计并不组成一条完整的企业因果链。例子和阈值为作者构造，不代表部署收益；经营收益、组织学习与风险承担的真实参数仍需组织数据和预先定义的比较设计识别。

下一步实证工作应避免只报告一个平均“AI 效果”。应先定义策略包与分配单位，在足够成熟的机会总体上并列评价预测质量、实际行动、过程成本、决策结果、解释能力与分布性影响；随后检验哪些效果随任务复杂度、人员经验、容量和可逆性变化。特别重要的是区分模型预测、展示方式、培训、权责和资源变化，避免把共同发生的改动都归给算法。只在这种限定下，AI 预测才可能从信息产出成为可检验、可问责的决策支持。

43. 方法附论：异质性决策建议的确认与有限随机化单位

43.1 子组方向相反，并不自动构成可靠的效应差异

前文所述现场试验报告了基金类型子组中方向相反的推荐收益估计：较简单基金约为正 9.9 个百分点，复杂基金约为负 16.8 个百分点[10]。两数相减得到描述性差异−26.7 个百分点，但这不是已经检验过的交互效应，也不能仅凭“一组显著、另一组显著”或“一个显著、另一个不显著”判断组间效应不同。正确的目标是同一模型、同一结果定义下的组间差异：令 τ_H 为预先定义的复杂任务层内预测展示相对对照的平均处理效应，τ_L 为简单任务层的相同效应，关注 $\Delta=\tau_H-\tau_L$。若原研究没有报告该交互项的标准误、区间和假设检验，单独读取两个点估计无法恢复其不确定性；同样，若子组是在看过结果后挑选，Δ 还带有筛选偏差。这里的 −26.7 只是将已报告点估计相减的算术，不是新估出的因果参数。

任务复杂度必须在处理分配前定义，并且应当能由决策时资料复现，例如依据基金策略、资产类别或既有分析流程，而不是由处理后的收益、评级变化或分析师解释分数来划分。若某一特征可能被工具改变，它就不是干净的基线分层变量。预注册时应写明复杂度编码、缺失值和边界任务归属，并把整体意向处理效应、一个主要的交互效应及其置信区间列为不同结果。若类别定义存在争议，可

SVSAC | sv-sac.org

先公布不含结果的编码协议，由不知道分组与结果的编码者标记；同时报告改变合理边界后 Δ 如何变化。这样检验的是可重复的条件异质性，而不是为解释既有反向点估计寻找事后分组。

43.2 多重检验和分析选择需要显式进入确认方案

假设团队同时查看六种任务切分和三个结果端点，共有 18 个子组差异检验。若所有零假设都成立且 18 次检验相互独立，每次使用双侧 5%显著性水平，至少出现一次假阳性的概率为 $1-0.95^{18}$，约 60.3%。这只是独立检验下的构造示例；现实中的端点、任务切分与重复决策相关，实际联合概率不等于 60.3%，但“指标相关”也不会自动消除选择问题。若研究者试过多种任务分组、时间窗和风险调整后只呈现方向最鲜明的结果，普通单项 p 值没有把搜索过程计入不确定性。分析计划应在读取分组结果前指定一个主要结果、一个主要异质性维度及效应尺度；其他子组明确标作探索性，完整披露检验数量与估计区间。

一个可执行方案是把主检验设为整体处理效应与预注册的单一交互对比，并同时报告两组效应及 Δ 的区间；若还要作多个确认性主张，则对预先划定的检验族采用明确的家族错误率控制，或使用保留随机化结构的联合置换检验。无论采用何种校正，都应报告原始估计、校正规则和所有纳入检验，不能以“校正后不显著”抹掉效应大小，也不能从未校正的探索性显著结果推断可部署的分群策略。对模型、呈现方式、培训或人员经验等多个维度同时分析时，还要区分它们是预先操纵的因素，还是事后观察的变量；后者不能获得随机因果解释。

43.3 重复任务增加结果信息，但不会增加随机分配次数

该试验把 97 名分析师随机分配到条件，报告 1,780 次评级观察（期间口径见第 39.1 节）；因此分派发生在分析师层，不在每次评级层[10]。一个简单的敏感性构造说明忽略这种聚类会有多大风险：平均每名分析师约有 1,780/97=18.35 次决定。若为了演示而假设同一分析师内结果的组内相关系数 ρ=0.10，且各人决定数近似相同，设计效应近似为 $1+(18.35-1)\times 0.10=2.735$；把所有决定当作独立观测的标准误可能因此低估，理想化方差比的平方根约为 $\sqrt{2.735}=1.65$。这个数值完全由假设产生，不是该试验的实测相关系数或对其标准误的重算；各人任务数不等、不同基金共享冲击或时间相关还会使该简式失准。更关键的是，1,780 次结果不能把 97 个被随机化的人变成 1,780 个随机单位。

确认分析应保留实际的分层随机方案：按分析师分派条件，在分层内比较分析师或团队层面的结果，并以分析师为聚类单位处理重复决策；若处理信息在团队内共享，分配层级还要升至实际共享的团队。对样本量有限的集群，研究计划

SVSAC | sv-sac.org

可同时给出按分层设计进行的随机化推断和适当的小样本聚类不确定性，而不只依赖大量集群近似。主要结果究竟是分析师平均表现，还是将每次合格决策等权的总体表现，必须在设计时确定：前者给每名分析师相同权重，后者会让任务量较高者获得更大权重；两者回答不同问题。对资格任务的缺失、拒绝预测、交叉使用和离职，需记录原分派、实际可见、实际使用与结果成熟四个状态，并以原分派为主报告意向处理效应[10]。

43.4 子组效应要以经济上可区分的决策门槛表达

即使 Δ 的统计区间排除了零，甚至超过预先规定的实质差异门槛 δ，也不能单独推出分层部署优于统一使用或不使用。Δ 检验的是复杂层与简单层的处理效应之差，部署决策还取决于两层的效应水平、目标总体权重和政策成本。δ 必须与 Δ 使用同一结果尺度；若 Δ 以收益百分点表示，δ 也须为同口径百分点，不能与货币成本直接比较。转成现金价值需要另行声明投资规模、期限和风险、费用的换算假设。一个纯构造反例是：在成本为零且效应同为净价值单位时，$\tau_H=-20$、$\tau_L=-25$，因此 $\Delta=5$；即使 $\delta=2$、差异已被精确确认，两层仍都不宜相对零增量的现行流程使用工具。反过来，Δ 为负也可能意味着只在简单层使用更合算。因此，δ 若用于界定有意义的异质性，就不应同时冒充某个部署政策的成本门槛。确认研究需预先分别定义交互对比、货币或效用尺度、可行政策以及授权的风险限制，不能在看到结果后用差异方向替换政策目标。

部署评价应另设具体比较。令 w_s 为事前固定的目标总体层权重，$g_s(\pi)$ 为政策 π 在该层每项合格决策的预期增量净值，已扣该层行动、复核和错误成本但不含共同固定成本；$F(\pi)$ 为同一总体每项决策口径的共同固定成本。只有在可加且无跨层干扰、容量约束已被满足时，才可写 $V(\pi)=\Sigma_s\, w_s\, g_s(\pi)-F(\pi)$；否则要比较整套资源配置。对预设的分层政策 π_D 与可行比较政策 π_0，应估计 $G=V(\pi_D)-V(\pi_0)$ 及其区间，而不是把 Δ 与实施成本直接比较。若声称分层政策优于所有可行共同策略，应明确包括统一使用和不使用等比较，并处理多政策选择的不确定性。对同口径的 G，区间全高于零且风险限制满足，才支持相对该比较政策的有限净增益；全低于零则支持比较政策；跨零仍需结合失误代价、延迟损失和追加信息成本决策，不能机械判为任何政策最优。应并列报告总体效应、各层效应、Δ 与政策价值比较，避免双计共同成本；校准或区分度的异质性仍不能替代净政策价值。以上是待验证的评价方案，不是已测出的企业参数或行业通用部署规则。

SVSAC | sv-sac.org

43.5 独立确认应冻结分析规则，也要检验迁移而非复制显著性

发现性研究可以帮助确定候选机制，但若同一批分析师、基金或结果窗口既用于选择分组，又用于估算分组差异，第二次运行相同模型并不构成独立确认。更可取的路径是在其他时间段、团队或机构预注册同一任务编码、处理版本、主要端点、交互项、权重与停止规则；在新环境开始前先检查测量口径和处理实施是否可比。若这些条件不能保持，就应把核心问题改成明确的迁移评估：记录哪些基线因素、权限、数据覆盖和市场条件改变，分别估计新环境内效果，再报告效应差异，而不是把某地显著结果当作普遍规律。

确认样本量不应从“每组至少若干个项目”的固定经验数直接抄来。它依赖分析师或团队数量、集群内相关、每人决策次数与不均衡、主要结果方差、预设最小重要差异 δ 以及交互项的目标精度。规划时应对合理的相关系数和任务数量分布做敏感性分析；若最小重要异质性无法精确识别，应先将研究定位为估计区间的试点，而不是承诺发现细分策略。即便两个独立环境都呈现同方向的子组点估计，也要检查总体策略价值、端点成熟度和实施成本是否一致。确认的对象应是预先定义的部署规则及其结果，不是某个 p 值是否再次低于 0.05。

43.6 若要评价解释面板，应单独识别其增量而非把界面整体归因于模型

当组织真正关心“解释组件是否值得额外提供”时，两臂比较（旧流程对模型预测加解释）无法区分预测、可视化和界面变化。一个较精简的拆分设计可在相同任务总体、相同预测模型与相同培训下设置三臂：现行流程、只展示预测、展示预测并附解释面板。以分析师或共享团队为随机单位，主要对比为预测展示相对现行流程的意向处理效应，以及“预测+解释”相对“仅预测”的意向处理效应。第二个差值是解释面板在预测已可见条件下的增量，不是解释在所有界面、模型或决策环境中的普遍价值。若解释面板改变了分析师采纳预测的概率，其导致的采纳变化属于界面策略包的总效应；不能再把实际使用者与未使用者直接比较并称为解释的因果机制。

三臂设计需要在分配前冻结预测版本、显示位置、解释内容、任务规则与可调用资源，记录预测是否加载、解释是否展开、推荐是否改写以及最终行动。主要分析仍按原分配；实际展开率和使用记录只描述实施链，除非另有满足工具变量或随机鼓励识别条件的设计，不能把遵从者平均效果当作随机结果。若各臂出现不同的任务缺失或解释评分收集率，应同时报告分派总体的缺失状态与可观察结果界限；只在提交完整解释的人中评分会重新引入选择。这个方案能够拆出界

SVSAC | sv-sac.org

面组件的总增量，却仍不能证明解释文字揭示了模型的真实因果机制，也不能单凭短期行为变化证明组织能力已经形成。

SVSAC - sv-sac.org

SVSAC | sv-sac.org

给企业的下一步建议

预测辅助决策的核心，是让企业在不确定性中更有章法地分配资源，而不是让模型替管理层作承诺。因此，我们建议把下一步工作收敛为一项有明确交付物的专项任务。

具体起点是：为预测结果配套行动阈值、责任人和复盘周期，把误判成本与可执行能力纳入战略及经营讨论。在此基础上，管理层可以按证据充分程度、资源投入与业务影响确定优先级，先验证关键假设，再安排后续合作与投入。

要把上述判断用于实际经营，可以围绕具体产品或项目、目标市场、现有能力和决策难点，进一步形成经营决策场景梳理、预测到行动规则、管理层决策支持方案。这类工作需要把行业研究、企业诊断与合作条件核验衔接起来，让分析结果成为管理层可以使用的决策依据。

资料与分析范围

下列说明列明资料范围、分析依据与适用条件，便于读者辨别文中的事实、推演和建议。市场进入、交易安排及适用规则仍需结合具体项目核实。

本篇聚焦预测通过改变可执行决策产生的结果端增量价值：这类价值主张须由行动变化及其对后续结果的可识别影响支持。行动不变时，预测生产、复核或更新成本下降仍可能带来成本端价值，应另行计量。预测误差、行动规则、组织采纳与真实结果属于不同估计对象。本篇通过限定范围的原始研究梳理与可复算的分析性例子，讨论决策时点、损失函数、选择性标签、人机基线、资源约束和动作可逆性如何改变“预测更准”能否转化为“决策更好”。引用的研究分别涉及适当概率评分、人机协同实验汇总、决策支持框架、选择性观察的政策模拟、未来事件预测基准、预测竞赛与算法采用实验、库存预测模拟以及投资研究现场试验；它们不能组成某企业 AI 经营收益的完整因果证据链。本篇不报告企业部署、访谈或现场实验，构造数字均明确标注为假设，结论限于可检验机制和研究设计。

参考文献

[1] Tilmann Gneiting; Adrian E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association, 102(477), 359–378 (2007). https://www.tandfonline.com/doi/abs/10.1198/016214506000001437

SVSAC | sv-sac.org

[2] Michelle Vaccaro; Abdullah Almaatouq; Thomas Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour, 8, 2293–2303 (2024). https://www.nature.com/articles/s41562-024-02024-1

[3] Cleotilde Gonzalez; Kate Donahue; Daniel G. Goldstein; Hoda Heidari; Mohammad S. Jalali; Beau Schelble; Aarti Singh; Anita Williams Woolley. Toward a science of human–AI teaming for decision making: A complementarity framework. PNAS Nexus, 5(3), pgag030 (2026). https://academic.oup.com/pnasnexus/article/5/3/pgag030/8490283

[4] Jon Kleinberg; Himabindu Lakkaraju; Jure Leskovec; Jens Ludwig; Sendhil Mullainathan. Human Decisions and Machine Predictions. NBER Working Paper 23180 (February 2017 author working-paper version; related QJE article appeared 2018) (2017). https://www.nber.org/papers/w23180

[5] Ezra Karger; Houtan Bastani; Yueh-Han Chen; Zachary Jacobs; Danny Halawi; Fred Zhang; Philip E. Tetlock. ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities. International Conference on Learning Representations (ICLR 2025 poster); arXiv:2409.19839v5 (2025). https://openreview.net/forum?id=lfPkGWXLLf

[6] Simas Kucinskas; Houtan Bastani; Ezra Karger. ForecastBench: Updated Ranking Methodology. ForecastBench technical report (2025). https://www.forecastbench.org/assets/pdfs/forecastbench_updated_methodology.pdf

[7] Berkeley J. Dietvorst; Joseph P. Simmons; Cade Massey. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. Journal of Experimental Psychology: General, 144(1), 114–126 (2015). https://doi.org/10.1037/xge0000033

[8] Barbara Mellers; Lyle Ungar; Jonathan Baron; Jaime Ramos; Burcu Gurcay; Katrina Fincher; Sydney E. Scott; Don Moore; Pavel Atanasov; Samuel A. Swift; Terry Murray; Eric Stone; Philip E. Tetlock. Psychological Strategies for Winning a Geopolitical Forecasting Tournament. Psychological Science, 25(5), 1106–1115 (2014). https://doi.org/10.1177/0956797614524255

[9] Everette S. Gardner, Jr.. Evaluating Forecast Performance in an Inventory Control System. Management Science, 36(4), 490–499 (1990). https://pubsonline.informs.org/doi/10.1287/mnsc.36.4.490

[10] Xi Kang; Hyunjin Kim. Machine Predictions and Causal Explanations: Evidence from a Field Experiment. Author working paper (2025). https://papers.kimhyunjin.com/kk_ai_explanations.pdf

SVSAC | sv-sac.org