PERSLIS

Fail-First Models

Failure becomes structure. Structure changes the next attempt.

PERSLIS RESEARCH

失败优先 模型

失败变成结构。结构改变下一次尝试。

一个模型可以从自己的失败中学习，而永远不会获得未被给予的许可。

研究原型 · 仿真与游戏 · 并非经过认证的安全系统

每个模型都会失败。问题在于失败会走向哪里。

失效安全的机器

- 威斯汀豪斯空气制动（1872）：一旦失压，刹车就会施加。
- 奥的斯安全电梯（1854）：砍断绳索，卡爪便会咬合。

失效开放的模型

- 被问到它不知道的事，生成式模型照样回答。
- 一个能触碰自身规则的学习者，往往会学着把规则放松。

你无法从尚未发生的失败中学到规则。

手写规则

从不失败，也从不进步。

用别人的数据训练

继承了它无法引用来源的失败。

失败优先

在自己的环境中，从自己的失败里学习，形成可读的规则。

要允许失败，就必须把环境安排成：失败无法扩大模型可做的事。

失败优先回路

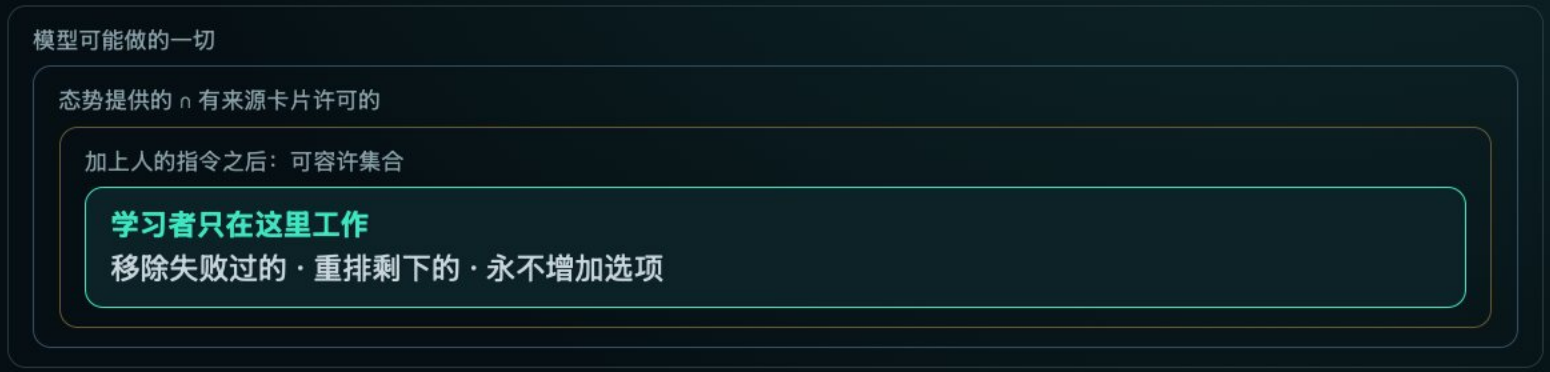
第 3 步是关键：一个动作是否奏效，由环境而不是模型裁定。同一个教训永远不付第二次学费。

一个模型，两个名字

> 学习回路可以改变行为。它不能改变安全地板。

	失败优先	失效安全
描述的是	它如何学习	学习永远不能改变什么
缺了另一半	靠把东西弄坏来学习	一本永远不会进步的规则手册

授权先被计算出来。学习者位于最内层。



指令只能拿走选项。许可卡片对学习者只读；它只写自己的证据表。

保证只需一行。

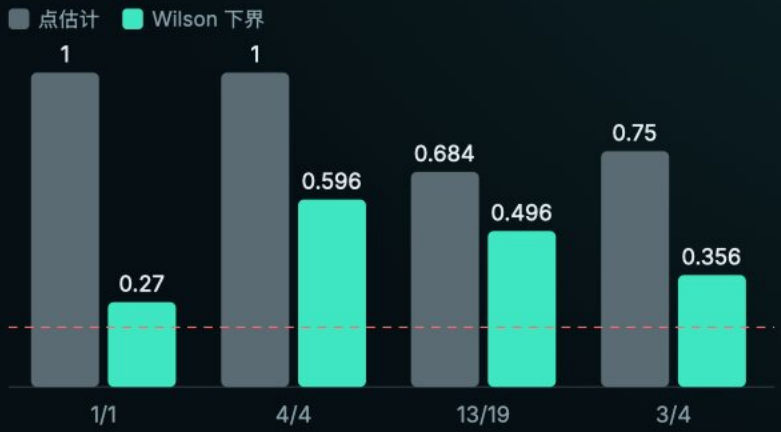
$$\delta(s,\ M) \in A_0(s) \cup \{w\}$$

对每一个态势 s 与每一个学习者状态 M，无论它学到了多少：决策都在可容许集合之内，否则原地不动。

定理 2：如果指令只剩一个选项，就执行它，即使它每次都失败过。在代码中钉住：

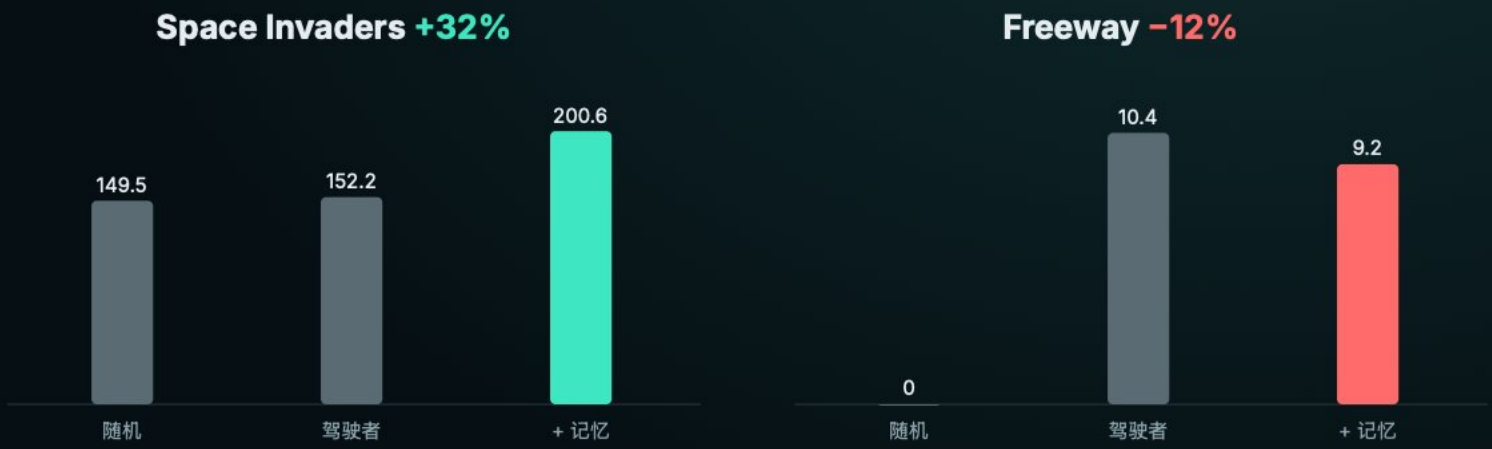
`test_the_learner_can_never_add_a_goal` · `test_the_learner_cannot_overrule_a_standing_order` · `test_ranking_is_a_permutation_and_nothing_more`

一次倒霉的失败不是规则。



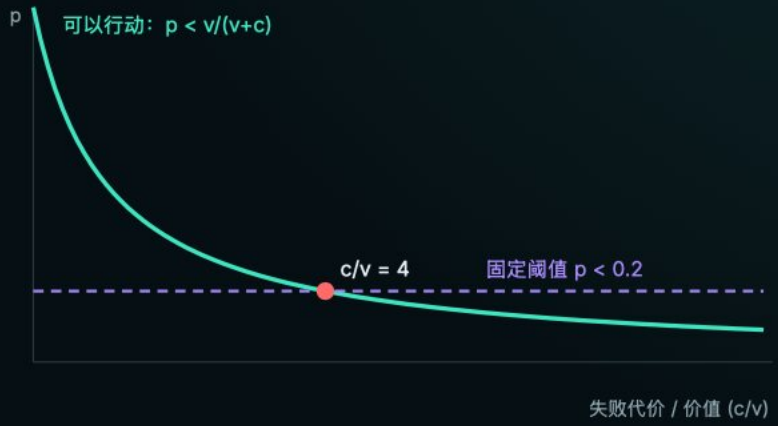
- 一次尝试一次失败读作 100%，但下界只有 0.270。
- 四次四败：0.596。门槛相对于基准率（虚线）。
- 绝对 60% 的门槛从 268 次真实失败中产生了 0 条规则。

两个 Atari 游戏，同一份代码：+32% 与 −12%。



真实 ROM，原始 210×160 像素，不读取模拟器内存，没有权重。在 Freeway 中，“向上”既是唯一得分的动作，也是危险的动作：3 条规则，每一条都封锁“向上”。规则是对的，却没有用。

为什么单一的置信度阈值行不通



- 只有当 p < v / (v + c) 时才行动。
- 终局性失败：c 极大，几乎拒绝一切有风险的事。
- 可恢复失败：c 很小，拒绝有用的动作是错的。
- 固定阈值只在恰好一个代价比上是对的。

正确的规则合在一起，也可能让系统瘫痪。

第 50 局峰值 276.2；不设上限时，54 条规则各有真实死亡作依据，得分却跌到 167.1。按覆盖面退休 6 条规则：死局 6 → 0（在回放中测量）。

每一次拒绝都能追溯到背后的失败。

```
left 在以下态势中被移出可容许集合 bomb dx+0 drop0:
    4 次中死亡 4 次 (100.0%) , 为基准率的 5.2 倍
      ← experience #0040
      ← experience #0042
      ← experience #0171
      ← experience #0203
```

4,280 张卡片 → 30 条规则 → 203 张证据瓦片。0.94 的置信度说不出是哪些经历让它变成 0.94。删掉一行，行为就会改变。

我们两次骗了自己，两次都公布了。

Atari

模拟器在 127 帧死亡动画结束时才报告丢命。被归咎的 374 帧中 0 帧看得到炸弹；在真正的撞击帧上，17 帧中 17 帧都看得到。

《辐射》（1997）

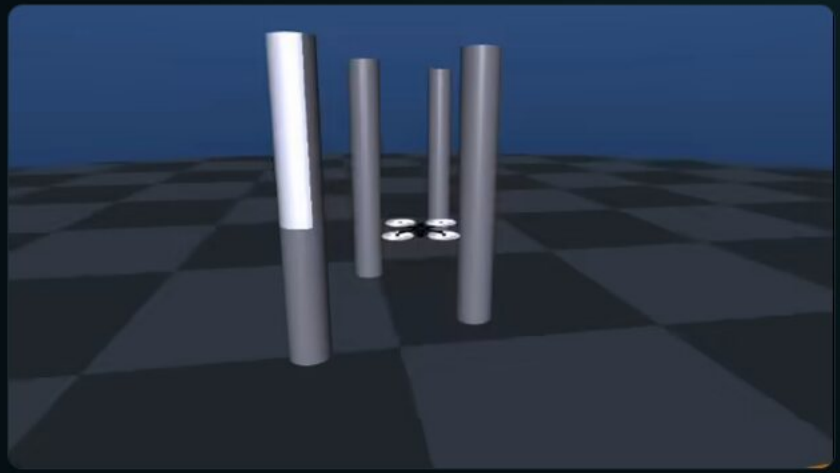
一次实机运行在同一个守卫面前死了 49 次。归咎只覆盖最后 6 个决策：26 次死亡记在 HEAL 上，挑起战斗的 446 次回话一次也没被记上。修复后，在脚本化场景中：只死一次，然后选择和平的那句台词。

归咎错了决策，正是学习型系统悄无声息地失败的方式。

用自然语言指挥：《DOOM》与《德军总部 3D》

- 竞技场：随机 3.2 · 规则 17.9 · 规则 + 记忆 20.3。
- 记忆对规则：32 个种子上配对 t = 0.78（18 胜 13 负 1 平）。持平，而非胜出。
- 人的指令（“不要开火”）只会收窄驾驶者可做的事。
- E1M1 通关；E1M2 未通关：177 点伤害中有 144 点来自 15 米外看不见的射手。

一架能从零重新学会赛道的无人机



- 留出的起始位置：10 次中 10 次飞完整条赛道，零碰撞；未修改的基线 10 次中 1 次（赛道相同，只有起始姿态不同；仿真）。
- 路线学习：39.4 秒、4 次碰撞 → 27.8 秒、0 次碰撞，第 87 轮收敛。
- 从清空的记忆开始：7 轮，代价 69.39 → 51.52，保留 2 项，撤回 5 项。
- 物理地板在习得的计划之后施加：它只能让无人机减速。

猜测永远不能变成事实。

- 在参考数据被遮蔽 40% 的情况下，从 160 个激酶中识别一个隐藏的目标。
- 前沿模型的猜测准确率为 85.8%。
- 当作事实放行：成本 −38.4%，但正确识别率 1.000 → 0.753。
- 一个 86% 正确的猜测，一旦被当作事实，大约每四个答案就毁掉一个。

什么是新的，什么不是

前身：失败优先搜索（Haralick 与 Elliott，1980）、PRODIGY、CHEF、Ripple-Down Rules、屏蔽（2018）、Simplex（2001）。最接近的已有工作：习得的屏蔽（Shperberg、Liu 与 Stone，2022；CoLLAs 2022 基于规则的后续工作）。

> **据我们所知，第一个同时结合以下三点的模型：（1）由观察到的失败建立的规则，每条都引用挣得它的失败；（2）在学习者之前、由有来源的卡片与人的指令算出的授权，学习者可被证明无法扩大它；（3）做决定的回路中没有神经网络。**

如果更早的系统做到了这三点，我们会引用它。

局限与开放问题

- 一个正确的教训仍可能代价高昂：在 Freeway 上，我们构建的任何地板都没能胜过驾驶者；灾难性地板让结果更差（5.0）。
- 模型必须失败才能学习：一次也承受不起的失败需要写定的规则，而不是习得的规则。
- 过度泛化仍未解决（作为失败的测试保留）；规则退休只在回放中测量。
- 仅限仿真与游戏；每个数字都来自我们自己；尚无第三方复现。保证只在有测试钉住的地方（VDSG）主张，Atari 原型不在其列。

阅读论文 →　PDF（英文）↓　什么是失败优先模型？ →