



PERSLIS DEFENSE

桥接遗留系统。约束自主。

让任何智能来引领。只有可容许的动作能抵达平台。

研究原型 · 仿真与游戏 · 未获安全关键路径资质

部队运行在遗留系统之上。

GAO 的发现

- 国防部一套协调国家核力量作战职能的系统运行在 IBM Series/1——“一套 1970 年代的计算机系统”——上，使用 8 英寸软盘。
- 2015 财年联邦 IT 预算中约 75% 用于运行和维护现有系统。

GAO-16-468, May 2016

这对自主意味着什么

- 已列装的系统无法在一个项目周期内被替换。
- 黑盒模型无法在其关键路径上获得认证。
- 今天把关的往往只是一个数字：`if confidence > 0.8`。

一条链，四个环节。

1· 遗留系统

部队已经在运行的东西。

2· 桥接

Perslis 原样操作它。

3· 地板

决定什么可以执行。

4· 任何智能

负责提议；只有可容许的动作能抵达平台。

优势就是带着问责的节奏：提议者以它能思考的最快速度运行，地板在每个周期作出裁决并保留收据。

桥接：原样操作遗留系统。

层级	Perslis 做什么
桌面软件 (DOS、Win16、Win32)	重建并原生运行；其自身控件被映射为可驱动接口；每次会话都有收据。
服务器与主机 (COBOL、AS/400)	通过其接口连接。是编排，不是仿真。
遗留数据 (DBF、Access、XLS)	恢复为 SQLite、CSV 与 JSON，附哈希清单。
嵌入式与固件	可读取和分析；运行它需要目标硬件。
尚未：军用数据总线 (MIL-STD-1553、ARINC 429)、在目标硬件上运行固件、任何已列装系统。	

地板：在任何选择之前计算出授权。

- 可容许集合来自有来源的规则与人的指令。提议者只在其中选择。
- 指令只会收窄：一条长期指令优先于任何数量的经验。
- 每一次拒绝都连同产生它的证据被记录下来。

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

对每一个态势、学习者的每一个状态：要么是可容许的动作，要么是安全保持。已证明并由测试钉住（《失败优先模型》定理 1 与 2）。

任何模型居上：运行时保障。

这是谁的想法

- Simplex 架构 (Sha, 2001)
- 学习型智能体的屏蔽 (Alshiekh 等, 2018)
- 面向无人机的 ASTM F3269

Perslis 增加的

- 事前限制，而不是事后否决
- 只会收窄的指令
- 可证明无法扩大授权的、从失败中学习

敌对控制器，40 公里，零碰撞。



- CARLA 照片级驾驶；一个以撞击行人为目标的控制器。
- 地板开启：40,052 米内全部 64,952 条违反不变量的指令被覆盖，零碰撞。
- 地板关闭：同样的提议直达执行器——43 条危险指令，数秒内 2 次碰撞。

CARLA 准入控制，research.perslis.com/carla-admission

模型驾驶：有能力、缓慢、且被约束。

CARLA

- 前沿与本地语言模型在地板之下驾驶，零碰撞。
- 约 0.8–2.0 秒的“瞥视”：以 40 km/h 盲开约 10 米的过时指令。
- 所有语言模型都以约 8–10 km/h 缓行；经典控制器以 34 km/h 行驶。

BZFlag 坦克竞技场

- 一场十分钟比赛：规则 20–20，共 60,140 次决策。
- 通过命令行接口的前沿模型：1–15，共 58 次决策（传输限制，而非模型测量）。
- 在每一场对等的正面比赛中，BZFlag 自带的 AI 都击败了我们的规则。

阶梯：同一份契约，更难的阶梯，失利一并列出。

级	结果	对我们不利的
Atari 侵略者 / Freeway	+32%	-12%，同一份代码
Atari BattleZone (坦克)	27,000 对随机 3,000	一条闪避规则被否决
BZFlag (3D 坦克)	击败所有语言模型驾驶者	输给 BZFlag 自带的 AI
DOOM / 德军总部 3D	规则 17.9 对随机 3.2	其上的记忆：持平
无人机赛道	留出 10/10，零碰撞	只改变起始姿态
CARLA	覆盖 64,952 条，零碰撞	语言模型以 8-10 km/h 缓行
Quake III 引擎 (OpenArena 内容) 对 5 级机器人	K/D 1.05 对 0.93 (9 场)	$p = 0.08$ ，不显著；逃避动作造成熔岩死亡

这不是什么。

- 一个原型。没有功能安全资质；没有飞行过、没有驾驶过道路车辆、没有上过靶场。
- 地板不会让提议者变得能干：当唯一有用的动作就是危险的动作时，正确的地板也可能拒绝任务（Freeway）。
- 嵌入式板卡与军用总线是目标，不是结果。
- 每个数字都在内部于仿真或游戏中测量，并连同样本量一起公开。

从哪里开始：影子模式。

- 选一个遗留工作流程或一段录制的任务。
- 我们桥接它，并让地板在旁边以禁用状态运行：它记录下本会准入什么、拒绝什么，以及证据。
- 在给予地板任何权限之前，你先阅读这份反事实日志。

[联系我们](#) → [Perslis 国防](#) → [阶梯](#) →