

AI 技术每日分析：Cashfree 推出 Relay，AI Agent 开始按业务结果收费

中国高技术产业发展促进会新质生产力工作委员会

博雅云创 & 中科创新驱动

2026 年 8 月 31 日星期一

摘要

今天 AI 技术方向的新动态并不集中在新一轮超大模型参数竞赛，而更多落到“Agent 真正接管业务以后，系统怎么运行”这一层。印度支付基础设施公司 Cashfree Payments 正式向商户推出 Relay，把购物车召回、失败支付重试、货到付款确认、订阅催缴和争议处理交给可持续运行的 AI Agent，并计划从免费试用逐步转向按业务结果收费；美国得州则暂停州级资金继续采购 Flock Safety 的 AI 车牌识别摄像头，说明 AI 视觉系统在进入公共治理后，数据共享、审计和使用边界会直接决定其能否继续扩张；开源本地推理工具 KoboldCpp 发布 1.120 版，新增 DirectIO 模型加载并快速适配 Qwen3.8-Flash-Next、Ling-3.0-flash 等新模型，显示开放模型生态的竞争正在从“有没有权重”转向“普通硬件能不能高效跑起来”。在生产级 Agent 安全方面，最新行业分析进一步提出：身份、委托关系、短期凭证、可归因遥测和全局“kill path”应先于复杂网关策略建设。AI 产业正在从模型能力竞争进入运行时、合规性和业务闭环竞争。

目录

一、Cashfree 把支付运营交给 Relay：Agent 开始按“业务结果”而不是 Token 定价	2
--	---

二、得州暂停州级资金购买 Flock 摄像头：AI 视觉开始被要求证明 “谁能查、查什么、查多久”	2
三、KoboldCpp 1.120 发布：开放模型竞争正在下沉到“本地推理工 程”	3
四、Agent 安全开始从“网关拦截”转向身份与委托关系	4
趋势观察	4
参考资料	5

一、Cashfree 把支付运营交给 Relay: Agent 开始按“业务结果”而不是 Token 定价

8 月 30 日, 印度支付公司 Cashfree Payments 宣布面向商户推出 Relay AI Agent。Relay 可以处理失败支付重试、放弃购物车召回、货到付款订单确认、订阅扣款失败处理以及支付争议等任务, 并可通过电话、消息等方式持续跟进客户。Cashfree 表示, 产品上线初期免费, 后续计划采用基于结果的计费模式, 例如按照成功挽回的购物车、成功处理的争议等结果收费。

这件事值得关注的地方, 不只是支付行业又多了一个 Agent, 而是商业模式开始改变。大模型 API 通常按 Token 收费, SaaS 通常按席位收费, 而 Relay 尝试把 Agent 价格直接绑定到“完成了多少业务结果”。如果这种模式跑通, 企业采购 AI 时会更容易把成本和收入、回款率或运营效率关联起来。

支付运营又是非常适合 Agent 的场景: 任务频率高、流程相对标准、数据结构化程度较高, 同时又需要根据客户行为做持续判断。Cashfree 官方产品页显示, Relay 的多个 Agent 共享业务记忆, 并允许商户设置明确规则, 让 Agent 按照批准的流程执行, 而不是自由发挥。这种“规则约束 + 共享记忆 + 闭环执行”可能比泛化聊天能力更接近企业 Agent 真正的产品形态。

二、得州暂停州级资金购买 Flock 摄像头: AI 视觉开始被要求证明“谁能查、查什么、查多久”

8 月 30 日, 《The Verge》和《Texas Tribune》相关报道显示, 得州州长 Greg Abbott 已要求州政府机构暂停使用州级资金采购 Flock Safety 的 AI 车牌识别摄像头。此前调查显示, 得州相关机构自 2023 年以来至少投

入 3000 万美元扩展 Flock 网络，其中部分资金来自为打击催化转换器盗窃而向汽车保险保单加收的 1 美元费用。

Flock 系统不仅识别车牌，还会记录车辆品牌、车型、颜色以及部分外观特征，并允许不同执法机构在一定条件下共享查询。随着网络规模扩大，得州多地陆续出现执法人员被调查或被指控滥用系统的案例，一些城市也开始取消合同。

这件事对 AI 行业的启示很直接：视觉模型的识别准确率并不是公共场景规模化的唯一门槛。系统必须同时回答访问权限、跨机构共享、查询留痕、保存期限、误用追责和公众知情等问题。未来面向政府、园区、交通和公共安全的 AI 视觉产品，治理能力会越来越像产品本身的一部分，而不是部署之后再补的一份制度文件。

三、KoboldCpp 1.120 发布：开放模型竞争正在下沉到“本地推理工程”

KoboldCpp 1.120 在 8 月 29 日发布，并于 8 月 30 日进入开源社区集中讨论。新版加入 ‘-usedirectio’ 模型加载方式，同时允许 mmap 与 mlock 组合使用；还增加了对 Qwen3.8-Flash-Next 和 Ling-3.0-flash 的完整支持。Kobold Lite 则加入用户可配置的 JavaScript 工具，可与标准 Tool Calling 协同。

这些更新看起来不像“大模型发布”那么轰动，却代表开放模型生态最重要的一层：推理运行时。模型参数越来越大、MoE 结构越来越复杂、上下文越来越长以后，用户真正面对的是加载速度、内存映射、量化质量、显存占用、工具调用兼容和不同硬件的适配问题。

KoboldCpp 这类项目的价值，就是把快速变化的模型能力翻译成普通开发者可以运行的本地软件。尤其当企业越来越关注数据不出域、模型供应连续性和推理成本时，本地运行时、GGUF 量化、模型路由和工具调用

兼容会构成开放模型商业化的重要底座。

四、Agent 安全开始从“网关拦截”转向身份与委托关系

VentureBeat 8 月 30 日刊发的 Agent 安全分析提出一个值得企业架构团队重视的观点：一个通过身份认证的 Agent 依然可能越权、偏离任务、泄露数据或被记忆污染，因此仅在 Agent 外面加一层 API Gateway 并不足够。

文章提出六层依赖关系：先建立 Agent 清单和负责人，再为每个 Agent 建立独立身份以及“代表谁执行什么任务”的委托上下文；随后使用任务范围内的短期凭证，建立可归因追溯，再做运行时动作拦截，最后建立行为基线和可以跨系统立即撤销权限的 kill path。

这套思路的价值在于，它把 Agent 看成一种新的企业主体，而不是“拿着员工 Token 运行的脚本”。随着 Agent 进入财务、代码、客服和运营系统，企业最终需要能够回答：哪一个 Agent 做了什么、代表谁做、为什么有权限、调用了哪些工具、产生了什么结果，以及出现异常后能否立即停止。这些问题将成为下一阶段企业 Agent 平台的基础能力。

趋势观察

今天的几条信息指向同一个方向：AI 正在离开单纯的聊天窗口。支付 Agent 开始直接承担收入和回款流程，AI 视觉进入公共治理后遭遇更严格的数据边界，本地推理工具不断补齐开放模型的运行能力，而企业安全架构则开始给 Agent 建立独立身份与审计链。模型仍然重要，但企业真正需要采购的将越来越是一套“可运行、可替换、可追踪、可停止、能产生业务结果”的系统。

参考资料

Business Standard | 《Cashfree rolls out AI agents for merchants to automate payment operations》 | 2026-08-30 | 核验 Relay 正式推出、业务范围及结果计费方向。

https://www.business-standard.com/companies/news/cashfree-rolls-out-ai-agents-for-merchants-to-automate-payment-operations-126083000455_1.html Cashfree Payments | 《Relay - Your AI Teammate for Payment Operations》 | 2026-08-30 访问 | 核验 Relay Agent 能力、共享业务记忆与规则执行机制。

<https://www.cashfree.com/relay/> Cashfree Docs | 《Cashfree Relay》 | 2026-08-30 访问 | 补充 Relay 工作流、Skills 和 MCP 能力。

<https://www.cashfree.com/docs/tools-ai/relay/overview> The Verge | 《Texas Governor Abbott blocks funding for more Flock cameras》 | 2026-08-30 | 核验得州暂停州级 Flock 资金及监管背景。

<https://www.theverge.com/ai-artificial-intelligence/986541/texas-governor-abbott-flock-cameras> Texas Tribune / ABC13 | 《Lawmakers added \$1 to Texans' car insurance policies; that money paid for thousands of Flock cameras》 | 2026-08-30 | 核验资金来源、网络规模和采购情况。

<https://abc13.com/story/lawmakers-added-1-texans-car-insurance-policies-money-paid-thousands-flock-cameras/19764215/> Kobold-Cpp GitHub | 《koboldcpp-1.120 Release》 | 2026-08-29 | 核验 DirectIO、mmap+mlock、新模型与 Tool Calling 更新。

<https://github.com/LostRuins/koboldcpp/releases/tag/v1.120> AI Weekly | 《Koboldcpp v1.120 adds DirectIO load mode, Qwen3.8-Flash-Next》 | 2026-08-30 | 补充本地推理工程影响。

<https://aiweekly.co/node/11128> VentureBeat | 《AI agents that pass authentication can still drift, expose data, or get memory-poisoned》
| 2026-08-30 | 用于 Agent 身份、委托、遥测与 kill path 分析。

<https://venturebeat.com/security/ai-agents-that-pass-authentication-can-still-drift-expose-data-or-get-memory-poisoned> AccuKnox | 《AgentZ Documentation》 | 2026-08 访问 | 背景资料, 补充零信任 Agent、沙箱、运行时凭证与审计趋势。

<https://help.accuknox.com/agentz/> JetBrains | 《Modern Go Guidelines》 | 2026-08 更新 | 背景资料, 观察 AI Coding Agent 通过 Skill 补齐训练数据时效差的问题。

<https://github.com/JetBrains/go-modern-guidelines> —

联系我们，请扫描二维码



新质生产力工作委员会
官方公众号



工业智能算网
gyznsw.cn

新质生产力工作委员会：

中国高技术产业发展促进会新质生产力工作委员会，专注于推动工业人工智能、智能制造、数字化转型等前沿技术发展，为企业提供政策解读、技术咨询和产业对接服务。

工业智能算网：

专注于工业人工智能、新质生产力、工业软件 CAE、智能制造等前沿技术。提供每日动态分析、技术趋势解读、解决方案分享，推动工业智能化转型。

网站地址：<https://gyznsw.cn>