Policy Note | 系列

总 No.3 期/2026 年 5 月

代理资本与系统性风险:
Agentic AI 重构全球金融基础设施

执行摘要

当下，全球金融与科技领域的板块交界处发生了剧烈的构造运动。生成式人工智能（Generative AI）正式跨越了作为“辅助工具（Copilot）”的被动应用阶段，全面迈入具备高度自主决策、跨平台执行与复杂逻辑推理能力的“代理式人工智能（Agentic AI）”时代。这一技术跨越，并非单纯由底层算力规模的膨胀所驱动，而是由一系列密集的、极具进取性的金融垂直领域商业化部署，以及底层网络安全攻防范式的异变所共同引爆。

在这个具有历史意义的时间节点，人工智能初创巨头 Anthropic 发动了一场面对传统金融价值链的“闪电战”。2026 年 5 月 5 日，Anthropic 在纽约正式发布了 10 款专为金融服务业深度定制的 AI 代理（AI Agents）。这些代理不再是隔离在浏览器中的简单对话框，而是以插件（Plugin）和托管代理（Managed Agents）的形式，直接无缝嵌入至 Microsoft 365 生态系统（Excel、PowerPoint、Word、Outlook）的底层架构中。它们能够自主执行从起草路演材料（Pitchbooks）、跨

文件审计财务报表、动态构建与更新估值模型，到执行 KYC（了解你的客户）合规筛查与月末结账等高度复杂的、知识密集型的投行与资管核心工作流。更为关键的是，这批 AI 代理通过模型上下文协议（MCP）被赋予了访问顶级金融数据提供商（如 Moody's 的 6 亿家公司数据库和 Dun & Bradstreet 的商业图谱）的 API 权限，这意味着它们能够在合法合规的框架内，实时调用并处理真实的商业与信用数据。

伴随产品端发布的，是资本端前所未有的深度结盟。Anthropic 宣布与华尔街顶级私募股权巨头黑石集团（Blackstone）、赫尔曼-弗里德曼（Hellman & Friedman）以及全球顶级投行高盛（Goldman Sachs）达成高达 15 亿美元的战略合资协议。在该交易架构中，Anthropic、黑石与 Hellman & Friedman 作为锚定投资者各出资约 3 亿美元，高盛作为创始投资方出资约 1.5 亿美元，泛大西洋投资（General Atlantic）等机构亦参与其中。该合资企业的核心使命是作为 Anthropic 的“部署先锋”，通过指派应用 AI 工程师（Applied AI engineers）深入企业内部，将 Claude AI 代理网络无缝嵌入到全球数千家私募股权投资组合公司及中型企业的核心运营架构中，以此实现降本增效的跨越式发展。

Anthropic 的猛烈攻势绝非孤立事件，而是整个科技巨头角逐金融主导权的缩影。在同一时期，OpenAI 宣布联手普华永道（PwC），旨在围绕 CFO 办公室的财务规划、税务、资金管理等核心节奏构建企业级 AI 代理。OpenAI 自身的财务组织已成为“零号客户（Customer Zero）”，其采购代理在相同人力下处理的合同量激增了 5 倍。不仅如此，OpenAI 正在敲定一家名为“The Deployment Company”的百亿美元级合资企业，获得了 TPG、Bain Capital、软银（SoftBank）等巨头高达 40 亿美元的新一轮注资，并向私募支持者承诺了极具侵略性的五年期 17.5%的年化保底收益率，以确保其技术能深入 2000 多家传统企业的毛细血管。同时，Google Cloud 也通过推出 7.5 亿美元的合作伙伴生态基金，并将其 Vertex AI 全面升级为“Gemini Enterprise

Agent Platform”，加速争夺企业级 Agent 部署的主导权，推动所谓“Agentic Enterprise”的落地。

宏观层面的数据印证了这一趋势的狂热。根据 Evident AI 发布的 2025/2026 年度银行 AI 成熟度指数（Evident AI Banking Index），摩根大通（JPMorgan Chase）、Capital One 和加拿大皇家银行（RBC）已在 AI 应用上与同业拉开显著差距，前十大银行的 AI 能力得分年增长率是其余银行的 2.3 倍。正如摩根大通首席执行官杰米·戴蒙（Jamie Dimon）在纽约活动上与 Anthropic 首席执行官 Dario Amodei 同台时所明确背书的那样：“这项技术如此强大，即使是万亿美元级别的基础设施资本支出也是完全合理的”。此外，加密金融原生机构 Anchorage Digital 甚至联手 Google Cloud 推出了首个允许 AI 代理在无人类干预下直接访问资金和执行跨境支付的“Agentic Banking”服务。

然而，在资本狂欢、估值飙升（OpenAI 最新估值达 8520 亿美元，Anthropic 酝酿逼近 9000 亿美元的 IPO）与生产力跃升的表象之下，全球宏观金融体系正面临着前所未有的底层架构脆弱性考验。2026 年 5 月 7 日，国际货币基金组织（IMF）打破了市场的盲目乐观，针对 Anthropic 最新发布的 Claude Mythos 大模型发出了警告。IMF 正式将最新一代 AI 模型的网络安全威胁，界定为足以引发“宏观金融冲击（Macro-financial Shocks）”并可能在金融系统层面造成“相关性失败（Correlated failures）”的系统性危险源。

本报告将从资产定价重塑、商业模式崩塌、网络安全降维打击、全球监管框架演进，直至极端系统性黑天鹅危机的沙盘推演，提供一份详尽的、极具信息密度的全景式深度剖析。

我们认为，在具体的监管落地层面，资本计提的重构远不足以应对挑战，面对势不可挡的算法破防，监管哲学的底层逻辑必须发生根本性变轨。既然传统的“防堵漏洞（Prevention）”注定失败，全球金融体

系的战略重心必须全面从“防堵漏洞”向“强化被攻破后的快速恢复韧性（Resilience-first）”发生历史性的战略转移， 构筑“AI 对抗 AI”时代的宏观审慎支柱：

1. **网络压力测试与情景分析的常态化：**监管不仅要评估资本充足率，必须强制开展高度拟真的网络压力测试，模拟前沿 AI 驱动的分布式网络攻击对清算账本、支付网关造成的真实破坏。
2. **董事会层面的网络风险穿透监督：**打破技术与金融的决策隔阂，将 AI 安全治理与网络风险管理提升至最高决策层的核心议程。
3. **公私部门协同打破信息孤岛：**面对跨国境的 AI 威胁，单一机构的闭门防御注定失败。须建立央行、情报机构、头部 AI 企业与金融核心节点之间的实时威胁情报共享与事件极速响应机制。
4. **强化针对新兴市场的国际防线合作：**发达经济体须通过资金援助、技术转移与能力建设，帮助资源严重受限的新兴市场和发展中国家提升底层防御水平，修补这块足以葬送全球金融互信的结构性短板。

第一章、 商业效能跃升与传统护城河的结构性瓦解

技术范式的转移，首先最直接地体现在资本市场对原有商业模式估值的重构上。在 Agentic AI 的冲击下，依靠信息不对称壁垒的金融数据服务商与依赖线性增长的传统 SaaS（软件即服务）企业的底层逻辑，正在经历一场系统性的崩塌与重建。

1.1 “信息套利”护城河的消亡与 FactSet 们的估值重估

长期以来，如 FactSet、Morningstar、彭博（Bloomberg）及 S&P Global 等金融数据巨头，其坚固的商业护城河建立在“海量信息聚合、数据标准化处理与专有终端高价分发”的封闭闭环之上。金融机构每年支付数万至数十万美元的终端订阅费用，以获取经过清洗的宏观经济指标、财报基本面数据与市场情绪分析。然而，Anthropic 十款金融 AI 代理的问世，直接将这层维系了数十年的护城河“击穿”。

2026 年 5 月 5 日，在 Anthropic 宣布其 AI 代理可直接在 Excel 和 PowerPoint 中基于第三方数据源自动更新财务模型并撰写深度分析报告后，美国资本市场迅速做出了悲观的定价反应。FactSet 的股价在盘中一度暴跌高达 8.1%，Morningstar 回吐早盘涨幅并重挫逾 3%，S&P Global 与 Moody's 同样承受了显著的抛压。市场调研机构 Morningstar 自身的证券研究部门甚至在此前就下调了包括 FactSet 在内的多家软件与数据服务商的“经济护城河（Economic Moat）”评级及公允价值预期。

这一估值重估的底层宏观逻辑在于：AI 代理将金融数据的“提取处理”与“商业洞察生成”之间的边际摩擦成本降至无限趋近于零。在传统工作流中，华尔街初级分析师需要通过 FactSet 终端下载浩如烟海的数据，手工导入 Excel 进行 DCF（现金流折现）建模，再复制到 PowerPoint 进行排版。现在，Claude 代理能够跨平台维持上下文连贯性，在后台自动完成数据抓取、逻辑审计、可比公司选择

（Comparables selection）与可视化输出，其在金融代理基准测试（Vals AI Finance Agent benchmark）中得分高达64.37%。当AI代理成为数据消费的“第一级入口（First-level interface）”时，传统数据服务商面临着被彻底“管道化（Commoditization）”的巨大风险。市场担忧，如果AI模型能够直接从原始的SEC EDGAR数据库、低成本的API接口或非结构化新闻中抓取并推理出同等质量的数据，那些仅仅依靠“提供更好用的查询界面”来维持高溢价的传统数据商，其议价能力将被无情剥夺。

1.2 “按人头订阅（Per-seat）”模式的崩塌与定价重构

比单一金融数据商受挫更为宏大且具有破坏性的，是支撑过去二十年全球企业级软件（Enterprise Software）繁荣的底层商业基石——按人头订阅模式（Per-seat Licensing）的彻底坍塌。

在SaaS 1.0时代，软件公司的估值贴现模型（DCF）极其简单且具有线性可预测性：企业客户的业务规模扩张必然伴随员工人数的增加，进而带来软件“席位”订阅量的直线上升，形成高额的年度经常性收入（ARR）和净收入留存率（NDR）。然而，Agentic AI直接打破了“用户即人类”的假设。贝恩咨询（Bain & Company）和德勤（Deloitte）等机构的深度报告指出，当一个高度自治的AI代理能够在后台同时处理以往需要数十名员工才能完成的客服工单分发、威胁排查或客户开户审批时，企业对人类员工的操作席位需求将出现断崖式下跌。

这就是华尔街所恐慌的“席位压缩（Seat Compression）”效应。业界实证数据表明，如果10个AI代理能够承担100名销售开发代表（SDR）的工作量，企业只需保留10个Salesforce的控制塔席位，这将导致传统SaaS供应商面临高达90%的收入萎缩（Headcount Compression）。2026年2月3日，这种被称为“SaaSpocalypse（SaaS末日）”的恐慌情绪曾一度在48小时内抹去企业级软件板块近2850亿美元的庞大市值。公开数据显示，2026年第一季度，AI原生企

业级支出的年均增长率飙升至94%，而传统SaaS的增长率已停滞在8%。

定价模式演进	SaaS 1.0时代 (2005-2024)	Agentic AI时代 (2025-未来)	宏观财务影响
核心收费指标	人类员工账号数 (Per-seat / Per-user)	模型消耗量 (Tokens/Compute)或业务结果达成率	收入与人类员工规模彻底脱钩，打破线性增长模型
成本与毛利结构	极低边际交付成本，毛利率可达80%-90%	极高的大模型推理算力成本 (Inference Costs)	软件毛利率承压，从“高利润通道”向“算力成本转嫁通道”转移 [1]
客户价值感知	为“使用软件的权限与工具效率”付费	为“AI代理直接交付的最终工作成果”付费	采购话语权向买方转移，大量低效的“点方案(Point Solutions)”被退订整合 [1]

为了在这一趋势中求生，整个软件与科技行业正被迫进行痛苦的自我革命，向“基于产出的定价（Outcome-based Pricing）”或“基于使用量/消费的混合定价（Consumption-based Pricing）”剧烈转型。例如，Adobe转向了“生成积分（Generative Credits）”定价，而Intercom等客服软件则直接推出了“按AI成功解决的客户工单数量”收费的模式。彭博（Bloomberg）及Gartner估计，在未来十年内，基

于传统订阅的静态定价模式在软件行业的占比将从60%暴跌至30%以下，而混合了产出和消耗的定价模式将成为绝对主流。

1.3 代理资本回报率（ROAC）的崛起与“10倍金融机构”的组织重构

商业模式的解构直接催生了衡量企业生产力的新维度。在传统财务资本（Financial Capital）与人力资本（Human Capital）之外，“代理资本（Agent Capital）” 正式确立了其作为现代企业核心资产的地位。专为受强监管金融机构打造AI代理操作系统的初创公司Primitive（其由前银行高管创立，并受到Fin Capital与Pelion Venture Partners支持）前瞻性地提出，金融行业必须引入 “代理资本回报率（Return on Agent Capital, ROAC）”这一全新关键绩效指标（KPI）。

ROAC的本质，是要求银行的首席财务官（CFO）和首席技术官（CTO）以管理资产负债表同样严谨的纪律，来量化、监控和约束AI代理的绩效表现。对于那些在AI成熟度上领先的机构而言（如摩根大通，其已通过AI在KYC文件处理中实现了约15亿美元的降本增效价值），他们不再仅仅评估“引入AI节省了多少人工时间”，而是精确计算“部署于信贷审批、合规监控、支付路由等核心业务流中的AI集群，在扣除高昂的大模型API调用费、算力折旧与数据治理成本后，为机构带来了多少真实的净利润提升与可审计的风险缓释”。

这一指标的广泛应用，正在推动顶级金融机构向“10倍银行（10x Bank）”的战略演进。在此愿景下，极少数的高级人类“代理运维专家（AgentOps）”能够通过类似于Primitive的控制塔（Control Tower）平台，指挥庞大AI代理集群，自主完成从交易后对账到复杂商业贷款文件审查的全流程工作。然而，在实现正向ROAC之前，银行必须承受短期内庞大的IT基础设施重建资本支出（CapEx），这也是阻碍中小银行跨越数字鸿沟，加剧全球金融业“马太效应”的核心壁垒。

第二章、 风险范式转移：降维打击与系统性暗网

如果说AI代理在商业前端带来的是生产力的高歌猛进，那么在技术底层的暗网深处，前沿AI模型正在以一种冷酷的姿态，对现有的全球网络安全防御体系进行着降维打击。网络风险的性质，正在经历一场结构性的基因突变。

2.1 攻防非对称性的极致恶化：从“专家直觉”向“算法自动化”跃迁

2026年4月，两款顶尖网络安全AI模型的相继问世，彻底打破了长久以来维持着脆弱平衡的全球网络攻防态势。

首先是Anthropic极其低调但震撼业内的 Claude Mythos Preview。这是一款被官方描述为在高级推理与计算机安全任务上实现质飞跃的前沿模型。在毫无人工干预的“沙箱”自动化测试中，Mythos展现出了远超人类顶级安全专家的漏洞发掘与利用能力。Anthropic披露数据显示，Mythos能够以超越人类专家数十倍的效率，自主发现数千个高危系统漏洞，其打击范围更是覆盖了目前全球运行的每一个主流操作系统和Web浏览器。

它不仅自主发现了潜藏于以绝对安全著称的OpenBSD操作系统内核中长达27年的远程崩溃缺陷，更从被全球数千万应用（包括各大银行音视频处理模块）广泛依赖的多媒体处理开源库FFmpeg中，揪出了一个深埋16年、历经500万次传统模糊测试（Fuzzing）依然未被触发的极端隐蔽漏洞（如并发竞态条件或堆栈溢出）。在具体的基准测试中，Mythos Preview在SWE-bench Verified（软件工程漏洞修复与生成基准）上的得分高达93.9%，相比上一代Opus 4.6的80.8%实现了代际跨越。

更令全球安全界与情报机构震动的是，Mythos不仅能“发现”静态代码层面的漏洞，更能将其“武器化”。它能够全自动地将多个看似低危的、分散的系统缺陷逻辑串联起来（Chaining），动态构建出极其

复杂的“零日漏洞利用链（Zero-day Working Exploits）”，从而直接绕过主流浏览器和操作系统的多层物理防御机制。英国人工智能安全研究所（AISI）的权威评估报告证实，Mythos 是首个能够端到端自主完成高度复杂的企业内网渗透模拟（例如涵盖 32 个复杂的提权与渗透步骤）的 AI 模型，只需花费不到 50 美元的算力成本即可完成过去人类团队耗时 20 小时以上的工作。Anthropic 自身也发出严厉警告称，这种降维打击对“经济、公共安全和国家安全的冲击可能是极其严重的”。

作为竞争层面的回应，OpenAI 迅速向通过严格政府 ID 审查的安全专家和企业推出了 GPT-5.4-Cyber 模型（基于其“网络安全可信访问计划”TAC）。这款模型刻意放宽了针对恶意代码和黑客指令的“拒绝边界（Lowered refusal boundary）”，其核心杀手锏在于强大的“二进制逆向工程（Binary Reverse Engineering）”能力——允许防守方在完全没有软件源代码的情况下，直接深入编译后的机器码底层，探查恶意软件逻辑、评估架构强度并加速补丁生成。

这两款模型的出现标志着全球网络安全攻防门槛的结构性剧变。过去，发现并利用深层架构漏洞往往需要国家级黑客组织（APT）投入大量顶级安全专家、耗费数月甚至数年时间手工调试；而现在，这种稀缺的“专家直觉”已被转化为低边际成本、可大幅度压缩耗时并无限并发执行的“算法调用”。攻击能力供给曲线的剧烈右移，使得金融机构面临的高烈度“零日攻击”频率，将从“几年一遇”骤然上升至“日常高频爆发”。企业过去只需修补已知漏洞库（CVE）中约 10%的紧急漏洞即可维持安全，而在 AI 的算力穷举下，这一数学防御逻辑将被彻底击碎。

2.2 风险升格：从“操作风险”向“系统性威胁”演进

在过去二十年以《巴塞尔协议》为核心的全球金融监管框架中，由黑客攻击、系统宕机或数据泄露导致的网络风险，长期被温和地归类为

“操作风险（Operational Risk）”的一个技术子集。传统的宏观审慎视角认为，这类风险具有统计学意义上的上限，其负面影响可以通过计提充足的资本缓冲、维持高水平的流动性覆盖率（LCR）以及购买商业网络安全保险来有效吸收与对冲。然而，Agentic AI 模型的问世，颠覆了这一底层的风险假设。

该风险性质的升格与异化，深刻地体现在以下三个结构性维度：

首先是网络节点的系统性与同质化瘫痪。全球金融机构在 IT 架构上存在高度的“技术同质化（Monoculture）”。正如美国证券交易委员会（SEC）主席加里·根斯勒（Gary Gensler）在多次公开演讲中所警告的那样，未来可能出现数以千计的金融实体共同依赖于极少数几家科技巨头提供的底层大语言模型或云服务基础设施。IMF 在其最新警告中将这一威胁进行了极度具象化的剖析：金融机构普遍使用相同的软件和共享服务提供商，这意味着 Mythos 级别的 AI 模型可以“同时在众多机构中制造并利用漏洞”。一旦 AI 探针找出这些通用组件（如加密协议栈、公共基础库）的致命漏洞，攻击便不再局限于单一银行的单点爆破，而是瞬间引发全网大面积的底层软件同步崩溃或错误信号的连锁共振。

其次是风险核心向“数据完整真实性（Data Integrity）”的不可逆转变。欧洲系统风险委员会（ESRB）和国际清算银行（BIS）的分析指出，从宏观审慎视角来看，最致命的网络冲击并非系统彻底宕机（可用性受损）或客户机密泄露（保密性受损），而是账本底层数据的隐秘损毁。AI 驱动的高级威胁能够以极高的伪装技巧潜入批发支付系统（如 Fedwire 或 CHIPS）或中央清算机构（CCPs），微调或篡改交易记录、动态授信额度以及抵押品标记。在现代信用货币体系中，整个金融市场的运作完全建立在对底层分布式账本或中央清算记录“绝对信任”的基础上。

基于上述逻辑，一条经典传导路径十分清晰： 一旦多家机构同时

遭到AI并发攻击且底层数据真实性被篡改，金融网络赖以生存的信心效应（Confidence Effects）将瞬间瓦解。交易对手方因无法信任账本而立刻切断同业授信，随之而来的是大面积的支付中断（Payment Disruptions）。支付阻断使得集中的清算轧差机制失效，全额资金需求激增，进而引发极端的流动性紧张（Liquidity Strains）。为了填补流动性黑洞，机构被迫在极度缺乏买盘的市场中不计成本地抛售优质资产，最终演变为一场吞噬宏观经济的甩卖动态（Fire-sale Dynamics）。

最后是第三方供应链与集中度风险（Concentration Risk）。金融稳定理事会（FSB）在最新发布的AI系统性风险报告指出，金融机构在算力、模型和云基础设施上对少数第三方巨头（AI模型提供商和数据供应商）的过度依赖，构成了“阿喀琉斯之踵”。针对单一具有系统重要性的云计算调度引擎或AI模型的恶意攻击，其破坏力将沿着数据供应链瞬间传染，导致缺乏可替代性（Lack of substitutability）的金融机构集群瞬间瘫痪。

2.3 地缘维度的脆弱性：防御屏障的“阶级分化”

随着攻防武器库的极度不对称，全球金融网的防御态势正在经历一场深刻且极度危险的“阶级分化”。AI安全红利的分配不均，直接导致了系统性防线在地理与机构层面的断裂。

为了在Mythos级别的能力被暗网武器化之前抢占先机，Anthropic主动牵头组建了名为“Project Glasswing”的防御性联盟，投入上亿美元的算力额度来修复底层基础设施。然而，这一防御护盾的覆盖范围存在极其严重的结构性倾斜。根据披露，Mythos目前处于严格的受控发布阶段，仅仅向全球约40家以美国资本为主导的核心科技巨头与顶级金融机构开放（包括亚马逊云科技、苹果、微软、思科以及摩根大通等大型银行）。这种“特权访问（Privileged Access）”机制使得这极少数核心机构能够在网络核弹爆发前获得珍贵的预警时间，从而提前

为自身的 IT 架构打上补丁。

这种访问权限的严重不均等，将巨量的非美国本土银行、区域性金融集团以及整个新兴市场的金融基础设施，无情地暴露在了毫无遮挡的极高风险之中。IMF 在警告中对此现象进行了严厉的抨击，明确指出：“网络风险不尊重国界（Cyber risk does not respect borders）”。当 AI 能力的扩散成为不可阻挡的趋势时，那些面临严重资源约束的新兴市场和发展中国家，由于无法获得同等水平的防御工具与威胁情报，将“不成比例地暴露于针对防御薄弱地区的攻击者面前”。

在全球高度互联的现代外汇与清算体系中，物理防线的断裂往往遵循木桶效应。这对全球金融体系而言意味着一个致命的结构性软肋：由于非美国机构缺乏针对前沿 AI 漏洞扫描的防御可见性，它们不可避免地将成为黑客进行“零日漏洞利用链”测试的天然靶场。最终，防御最薄弱的新兴市场环节，恰恰将成为引爆全球核心国家金融机构系统性传染危机的始发点。

第三章、 全球监管演进与防御重构：“AI 对抗 AI”

面对 AI 引发的降维打击与危机传导机制的非线性异化，传统的静态合规检查、年度文书审计以及基于过往历史数据的风险模型已彻底失效。全球主要央行、金融监管机构以及头部科技巨头正迅速摒弃旧有思维，推动宏观监管框架向智能化演进，迫使全球金融基础设施进入一场关乎生死存亡的“AI vs AI”类军备竞赛。

3.1 极速联盟与机器速度的动态防线

在意识到前沿 AI 能力一旦不可逆转地向外扩散（通过模型权重泄露或开源生态的恶意蒸馏）将带来毁灭性后果后，防守方被迫结成了深度的跨界利益共同体。前文提及的 Project Glasswing 联盟正是这种努力的第一步。

然而，这仅仅是前哨战。在应用层，企业 IT 部门过去拥有数周甚至数月的缓冲期来测试和部署安全补丁。但在 Agentic AI 时代，当防御方发布补丁的瞬间，攻击方同样可以利用开源大模型在极短的时间（如 72 小时甚至几分钟内）逆向还原出漏洞细节并生成攻击载荷。这就要求金融机构的防御机制必须从“人工主导”彻底转向“机器速度的动态风险对抗（Machine-speed Defense）”。例如，IBM 推出的 Concert 平台及 Netskope 发布的 AgentSkope 架构，均强调部署大量的防御性 AI 代理，全天候在端点、网络层和 API 逻辑层执行异常检测、行为流量分析与毫秒级的自动微切分隔离（Micro-segmentation），以求在恶意代码横向移动的毫秒之间将其自动熔断。

3.2 宏观审慎监管：从资本重罚到韧性压力测试

各国监管层对 AI 双刃剑的认知已提升至国家安全层面。2026 年 4 月中旬，由于 Mythos 等模型展现出的破防能力，美国财政部长斯科特·贝森特（Scott Bessent）紧急召集美联储主席杰罗姆·鲍威尔（Jerome Powell）以及包括高盛等华尔街全球系统重要性银行（G-

SIBs）的CEO，于华盛顿举行了闭门危机会议。美联储负责监管的副主席米歇尔·鲍曼（Michelle Bowman）亦公开警告，监管政策正面临极其严峻的平衡考验。

在具体的监管落地层面，政策的演进首先表现为资本计提的重构：巴塞尔协议III最终局（Basel III Endgame）：2026年3月，美联储联合联邦存款保险公司（FDIC）及货币监理署（OCC）重新发布了具有历史性意义的修订草案。该草案强制要求大型银行全面废除以往依赖内部历史数据测算的复杂内部模型，转而统一采用基于实际业务收入与支出的无模型方法（Non-modeled approach）来计算并计提巨额的“操作风险资本（Operational Risk Capital）”。监管层意识到，基于过去温和时代历史数据训练的内部模型，面对AI催生的瞬时网络攻击等极端“尾部风险（Tail Risk）”时会产生严重的低估。

然而，面对势不可挡的算法破防，监管哲学的底层逻辑必须发生根本性变轨。正如IMF坦言，尽管目前受严监管的金融软件比开源基础设施更难攻击，但这层微弱的优势“随着大模型训练规模的扩展、能力的极速扩散以及不可避免的权重泄露，很快就会被彻底侵蚀”。

既然传统的“防堵漏洞（Prevention）”注定失败，IMF强烈呼吁全球金融体系的战略重心必须全面从“防堵漏洞”向“强化被攻破后的快速恢复韧性（Resilience-first）”发生历史性的战略转移。这一倡议与加拿大金融监管局（OSFI）提出的AGILE（涵盖意识、护栏、创新、学习与生态系统弹性）框架不谋而合，具体而言，须建立以下政策应对框架：

1. **网络压力测试与情景分析的常态化**：监管不仅要评估资本充足率，必须强制开展高度拟真的网络压力测试，模拟前沿AI驱动的分布式网络攻击对清算账本、支付网关造成的真实破坏。例如欧洲央行（ECB）设定的资本耗损300基点的极端断网情景现场审查（OSQAR），以及英格兰银行（BoE）假定核心节点彻底瘫痪的宏观

阻断测试。

2. **董事会层面的网络风险穿透监督：**打破技术与金融的决策隔阂，将AI安全治理与网络风险管理提升至最高决策层的核心议程。
3. **公私部门协同打破信息孤岛：**面对跨国境的AI威胁，单一机构的闭门防御注定失败。必须建立央行、情报机构、头部AI企业与金融核心节点之间的实时威胁情报共享与事件极速响应机制。
4. **强化针对新兴市场的国际防线合作：**鉴于前文所述的防御“阶级分化”，发达经济体必须通过资金援助、技术转移与能力建设，帮助资源严重受限的新兴市场和发展中国家提升底层防御水平，修补这块足以葬送全球金融互信的结构性短板。

全球主要央行与组织AI网络风险监管框架演进	核心监管机构	政策工具与演进方向	应对AI系统性风险的核心逻辑
资本惩罚与缓冲	美联储(Fed) / OCC / FDIC	巴塞尔协议III最终局修订	废除内部历史数据模型，强制采用无模型方法计提巨额操作风险资本，应对AI引发的极端尾部冲击。
实战化韧性验证	欧洲中央银行 (ECB)	2026年网络弹性与地缘政治压力测试	设定资本耗损300基点的极端断网情景，现场审查银行在核心数据库遭到破坏时的真实物理恢复能力。

防范宏观传染	英格兰银行 (BoE)	宏观阻断测试 / 弹性压力测试	假定核心金融机构彻底瘫痪，测试系统替代路径，防范单点机构异化为吸干市场资金的“流动性黑洞”。
韧性优先	国际货币基金组织 (IMF)	系统应对与国际合作	承认“金融软件防线必被侵蚀”，主张全面转向“被攻破后的快速恢复韧性”，强制推进网络压力测试与新兴市场国际协作。
生态协同治理	加拿大金融监管局 (OSFI)	AGILE 综合性框架体系	强调跨国界的数据与威胁情报共享，将防御策略从“孤岛防御”全面升级为全金融生态圈层面的系统性协作。

第四章、 极端黑天鹅虚构推演：下一次系统性危机的起点

回溯近现代金融史的百年长河，几乎每一次能够颠覆全球经济秩序的系统性危机，都源于人类对“新型结构性风险”的傲慢与认知盲区。如果说 1929 年的大萧条源于对股市极度杠杆化的失控，2008 年的大衰退源于对基于次级抵押贷款的结构性衍生品（CDO/CDS）信用的盲目崇拜，那么下一次足以撼动全球金融体系根基的系统性危机，极大概率不会单纯始于传统的宏观信用收缩、通胀失控或资产泡沫的破裂。它很可能将由一次深藏于金融系统数字底层中的“代码突破”所引爆，进而演化为一场无法通过印钞机解决的连锁崩盘。

基于对全球现代支付清算网络高度互赖与极度集中的脆弱性剖析，以及当前开源大模型极易被恶意蒸馏并在暗网扩散的现实研判，我们对未来可能的“T+0 至 T+24 小时”极端黑天鹅场景作出如下逻辑推演：

- **T+0 小时：潜行于深水区的隐秘破防与数据投毒** 全球宏观经济正处于脆弱的复苏平衡期，地缘政治局势高压。一个由利益集团或极端组织支持的高级持续性威胁（APT）黑客团体，通过暗网获取了某种经历了模型蒸馏、被移除了安全护栏且极度精通金融底层架构的开源恶意 AI 大模型（类似于不受控的轻量级 Mythos）。该组织的目标并非发起容易被云端流量清洗中心监测到的分布式拒绝服务（DDoS）攻击，而是将 AI 探针悄无声息地指向了全球金融体系中最为核心的批发外汇与大额支付清算枢纽（例如每日处理数万亿美元、连接美国各大银行的 CHIPS 网络或多币种同步交收的 CLS 系统）。经过数周的深度语义潜伏与自动化逆向测试，恶意 AI 利用底层并发竞态条件（Race Condition）漏洞成功获取高极权限。随后，AI 并未直接窃取资金（这极易触发反洗钱熔断），而是极其精确且难以察觉地篡改了系统内部针对主要系统重要性银行的动态授信额度（Credit Limits）、抵押品合格状态标记以及多币种结算

的轧差汇率。致命的“逻辑金刚砂”已被无声无息地撒入高速运转的金融齿轮中。

- **T+4 小时：异常显露、对账失效与全网相关性认知瘫痪** 随着全球各主要金融时区（从东京、伦敦到纽约）结算周期的滚动，数家核心 G-SIB 银行的自动化代理对账系统（AgentOps）同时发出最高级别的红色警报：巨额的外汇结算头寸在不同节点之间出现无法弥合的错账与逻辑悖论。由于底层核心数据已被 AI 精妙篡改并扩散，甚至连系统的灾备冗余节点也同步复制了被“投毒”的错误数据。这正是 IMF 所担忧的“相关性失败”的极致体现。整个网络瞬间失去了“唯一事实来源（Single Source of Truth）”，银行机构彻底陷入深度认知瘫痪，无法分辨账本中哪些交易指令是真实的，哪些是凭空伪造的。
- **T+6 小时：清算网络瘫痪与“流动性黑洞”的迅速形成** 面对无法修复的数据灾难，按照应急预案，受感染的核心清算系统被迫紧急物理拔管隔离，全网暂停集中清算服务。然而，这些系统日常为全球银行业提供了惊人的“流动性节约效益”（由于多边净额轧差，银行只需准备极少量的实际资金即可完成天量交易）。当集中清算网络宕机，原本通过轧差处理的交易瞬间全部转化为全额（Gross）结算需求。各家银行立刻面临原本所需资金数倍甚至数十倍的巨额美元缺口，被迫转向央行的底层 RTGS 系统（如 Fedwire）进行全额支付，在极短时间内，整个金融网络中形成了一个深不见底的“流动性黑洞（Liquidity Black Hole）”。
- **T+8 小时：囚徒困境下的“战略性流动性囤积”** 面对天文数字的临时资金缺口和极度不透明的交易对手方风险（此时无人知晓哪家同业银行的底层账本已被彻底摧毁），理性的华尔街首席风险官们基于保护自身的博弈逻辑，将不约而同地触发“自保机制”——立刻停止向同业市场拆出任何非必要的资金，进入极其保守的“战略性流动性囤积（Strategic Liquidity Hoarding）”状态。宏观层面的资金流转在数小时内被完全切断，支付活动呈现雪崩式下降。

- **T+12 至 T+24 小时：流动性危机向偿付能力危机的高频转化与甩卖崩盘** 原本平稳运行的全球同业拆借市场瞬间冰封，隔夜回购利率呈指数级飙升。高度依赖短期回购市场进行杠杆融资的非银行金融中介（如对冲基金，以及目前规模已达数万亿美元且极度缺乏监管透明度的私募信贷市场），由于无法滚动其隔夜负债，资金链在日内即宣告断裂。为了满足系统自动触发的巨额追加保证金要求（Margin Calls），这些机构被迫在极度缺乏买盘的公开市场上，以极具破坏力的“跳楼价（Fire Sales）”恐慌性抛售流动性最好的优质资产（如美国长期国债和蓝筹科技股）以换取现金救急。无风险资产价格的剧烈崩塌进一步导致全市场抵押品价值缩水，触发了衍生品市场的连环平仓。至此，一场由几行底层恶意代码引发的技术突破，完美地顺着可预见的系统性传导链条，转化为了一场摧毁全球金融市场和经济系统的浩劫。

在此次黑天鹅事件的推演中，传统的宏观金融救市工具将面临前所未有的彻底失效。中央银行固然可以无限量地提供紧急流动性注入（如量化宽松或设立紧急贴现窗口），但当危机的内核是“由于底层数据账本的真实完整性不可信，导致市场参与者彻底丧失了交易意愿”时，再多的法定货币也无法买回金融系统赖以生存的数学与密码学“信任”。只有当受损机构的底层 IT 架构，在一场漫长且痛苦的“全网 AI 逆向溯源排查”中被彻底清理并获得重新验证后，冰封的现代金融机器才有可能极为缓慢地重新启动。

结语：技术狂欢下的冷思考与“韧性优先”

以 Anthropic 的金融 AI 代理矩阵重塑华尔街业务流，以及 Mythos 与 GPT-5.4-Cyber 级前沿攻防模型的涌现为标志，全球金融系统已驶入了一片充满极大生产力红利与极度致命未知风险交织的“暗黑水域”。

在 Agentic AI 主导的全新商业范式下，传统基于人力密集型服务

和静态信息垄断的金融护城河正在不可逆地崩塌，取而代之的是由“代理资本回报率（ROAC）”驱动的智能化商业重构。然而，这场看似光鲜的生产力狂欢背后，隐藏着令政策制定者如芒在背的冷酷现实：网络安全已经完全超越了首席信息官（CIO）的 IT 运维范畴，跃升为决定一家金融机构能否生存、维系宏观经济命脉乃至保障国家核心安全的最终底线。

正如 IMF 在 2026 年 5 月的警告中所作出的论断：依靠物理隔离与封闭闭环构建的金融软件防线，其所拥有的相对安全优势只是极其短暂的幻觉。随着大模型训练的无边界扩展、能力的极速扩散以及网络暗黑地带的开源逆向工程，这道看似固若金汤的防线被人工智能彻底侵蚀并攻破，仅仅只是时间问题。

对于政策制定者、央行宏观审慎监管者以及顶级金融机构的决策者而言，彻底抛弃“毕其功于一役”的防堵思维，全面接受“防线必破”的现实，并将动态的、以机器速度运行的被动恢复韧性（Resilience）深度嵌入银行的资本计提结构、商业定价模型乃至每日高频运营的微观机制中，已不再是未雨绸缪的战略前瞻，而是关乎系统生死存亡的当务之急。下一次金融风暴的倒计时，或许已经在那数千万行支撑着全球贸易、汇兑与国家信用的开源底层代码中，伴随着大模型算力的每一次脉冲，悄然滴答作响。

联系我们：CEISD@Shanghaitech.edu.cn