安全大模型或将重塑信息安全产业格局

中央财经大学会计学院 时浩然

当前，模型技术正从通用对话、内容生成快速渗透到网络安全领域，一批具备自主漏洞挖掘、自动化攻防、智能威胁研判能力的安全大模型陆续出现。其中，美国推出的安全大模型Mythos因极强的攻防能力，被业内人士称为“网络核武级”安全大模型。它不仅能大幅提升攻击效率，也在倒逼防御体系全面升级，正在深刻改写全球信息安全产业的攻防规则、市场格局与商业模式。

Mythos的出现，标志着全球信息安全正式进入AI主导的新阶段。它在短期内会带来行业阵痛：传统业务下滑、资本市场波动、中小企业承压……但从中长期来看，它将推动整个行业从被动防御走向智能防御、从低附加值走向高价值服务，市场空间和产业质量都将迈上一个新台阶。

对国内而言，外部强攻防AI模型的出现，既是压力也是机遇。它加速了国内网络安全行业的AI转型与本土化进程，拥有自研安全大模型、落地场景丰富、服务能力突出的头部企业将显著受益。未来的信息安全竞争，本质上就是安全大模型之间的进化速度之争、智能化能力之争。

从“意外泄露”到“强到不敢公开”

Anthropic内部博客草稿因系统配置失误意外公开，提前曝光代号Capybara（水豚）的新模型，被定义为“迄今最强大的AI模型”，明确标注网络安全能力极强、风险前所未见，消息引发AI与安全圈震动。

4月7日，Anthropic放弃常规公开发布，同步推出Claude Mythos Preview（预览版）与Project Glasswing行业安全计划，以“防御优先”模式小范围开放，核心是用最强攻击能力做主动防御，避免模型落入黑客手中。命名“Mythos”源自希腊语，意为“神话、叙事体系”，象征模型能深度串联知识、重构认知，实现从工具到自主智能体的突破。

Mythos不是“安全专用模型”，而是“通用强模型的安全能力涌现”，未针对安全任务专门微调，安全能力是底层推理、代码理解、自主决策的“能力溢出”。核心定位是AI驱动的主动防御层，让防御方以接近黑客的速度、视角自主发现漏洞、构造防御，在漏洞被利用前修复。Mythos的出现是AI与网络安全融合的里程碑，标志着“大模型自主攻防”时代的到来，将会重塑全球信息安全产业格局。

Anthropic内部评估：Mythos攻击能力远超当前全球防御水平，一旦公开，会瞬间成为黑客“超级武器”，可自主攻破几乎所有主流系统、基础设施，风险完全不可控。因此采取史上最严格发布策略：不开放API、不面向公众/普通企业；仅授权12家核心科技巨头（包括苹果公司、谷歌、微软、英伟达等）以及40余家关键基础设施组织（金融、能源、通信、开源社区）。

Mythos的出现对世界来说都是信息安全格局的关键变量。美国将强化科技巨头+安全厂商的防御垄断，通过“锁仓”最强模型，掌握全球漏洞挖掘主导权，巩固网络安全霸权。对我国来说将会形成技术代差压力，倒逼国产安全大模型加速自研、落地，推动“安全可控、数据本地化”提速，重塑本土安全产业竞争格局。

冲击首先体现在资本市场

Mythos这类具备强攻防能力的安全大模型的出现，相当于在全球网络安全市场投下一颗“技术炸弹”，打破了过去几十年“攻击追着防御跑、防御跟着漏洞补”的传统节奏。资本市场对中美两国软件股、安全股的反应，既有相似之处，也有明显差异。

美国市场对Mythos的反应非常激烈。消息释放后，美股网络安全板块出现明显暴跌，多家头部厂商单日跌幅较大，市值短时间内大量蒸发。

美国软件生态高度开放、云服务普及度极高，大量系统和应用暴露在公网环境下。Mythos这类模型可以自动挖掘零日漏洞、自动生成攻击脚本，让攻击门槛大幅降低，传统依靠规则库、特征码的安全工具几乎失效。投资者担心企业会大幅削减传统安全产品支出，延迟新单、取消续约，以订阅为核心的商业模式面临挑战。

不过美股内部也有分化：本身已经在大力投入AI安全、推出AI驱动安全平台的公司，跌幅明显更小；仍然高度依赖硬件防火墙、传统网关产品的厂商，估值受到的冲击最大。

我国A股市场网络安全和软件股的表现相对温和，更多是情绪面上的跟随下跌，幅度明显小于美股，并且反弹速度更快。主要原因有两点：一是国内关键行业和政企系统相对封闭，境外攻击模型直接造成大规模现实威胁的可能性较低，基本面受冲击有限；二是美国出现“网络核武级”安全大模型，反而强化了国内对自主可控、数据安全、模型本地化的重视，后续政府采购会更明确要求安全模型自研、算法可审计、数据不出境。

Mythos将深刻改变信息安全行业结构

安全大模型真正深远的影响，是改变整个行业“赚什么钱、怎么赚钱”。目前信息安全行业收入大致分为三大类：安全产品、安全服务、安全集成。在Mythos这类AI模型普及后，低附加值业务会被大量替代，高价值智能业务会快速爆发。

被明显替代的业务包括传统防火墙、IDS/IPS、普通WAF、老式终端杀毒、常规漏洞扫描器。这类产品依靠规则和特征库，识别不了AI生成的无特征攻击和零日漏洞，未来几年需求会持续萎缩。

低附加值人工服务也会受到影响，比如简单的渗透测试、标准化等保测评、日常策略配置、基础日志审计。以前靠人工堆时间，现在AI几分钟就能完成，价格会被压得很低，相关岗位需求也会明显减少。另外，低水平安全集成商也会被淘汰，因为简单拼设备、做实施、调策略的集成业务，在AI和云化趋势下越来越没必要，行业占比持续下降，大量中小集成商会被淘汰。

有部分业务不会消失，但必须“AI化”才能活下去。比如传统SOC安全管理平台，误报多、效率低，接入大模型后变成智能态势感知、XDR平台；数据安全产品从简单脱敏、审计，升级为AI自动识别敏感数据、监测异常行为；安全运营从人工盯屏幕，变成AI自动处置、人工复核，效率和单价同步提升。

直接受益、迎来爆发的新业务应该是完全围绕安全大模型产生的新增长点，包括私有化安全大模型部署与定制，AI驱动的威胁狩猎、漏洞挖掘，针对AI攻击的专项防御服务，模型安全测评、算法合规审计，BI安全运营订阅服务。这类业务增速快、毛利率高，将成为未来行业的主要收入来源。

Mythos可能带来产业范式转移

安全大模型带来的不是简单的产品升级，而是一次完整的产业范式转移，未来3~5年行业会呈现几条非常清晰的趋势。

未来攻防将进入“AI对AI”的竞赛时代。过去是黑客挖漏洞、厂商打补丁，未来是AI自动挖漏洞、AI自动防御。攻击会更规模化、平民化，防御必须具备实时进化、自主响应的能力。人工为主的安全运维模式会逐步退出历史舞台，安全智能体、AI自动化处置会成为标配。

全球格局中美分化，美国在AI攻防原始创新上仍占据优势，而中国会在政策支持、庞大内需的推动下，快速构建自主可控的安全大模型体系。国内行业集中度会进一步提升，少数几家具备全栈AI安全能力的龙头占据大部分市场份额，缺乏AI研发能力的中小厂商逐步出清。

商业模式从卖产品转向卖AI安全能力。行业将从一次性卖硬件、卖软件授权，转向长期订阅、按防护能力付费的模式。按算力、按节点、按风险等级、按事件响应次数收费会越来越普遍，企业现金流更稳定，整体毛利率也会明显提升。

监管趋严，AI安全合规成为刚需。国内外都会加快出台安全大模型监管规则，包括模型备案、算法可解释、数据本地化、攻防能力管控等。未来能否进入关键行业采购名单，很大程度上取决于AI安全合规能力，合规本身也会成长为一个重要赛道。