

文章编号:1003-2053(2026)09-1811-09

外资 AI Agent 军民两用技术安全审查判定标准研究

王文轩,罗先觉

(中国科学院大学公共政策与管理学院,北京 100190)

摘要: AI Agent 技术依靠自主决策、多步骤任务执行以及环境感知能力,正在改变全球科技竞争的格局。它把大语言模型嵌入到支架软件中,在无人持续监控的情况下可以规划执行任务链,并操控外部系统,因此具有天然的军民两用性。双重性边界取决于部署方式而不是产品类别,使传统依靠产品类别的审查方法越来越失效。在美国战略竞争加剧、欧盟经济安全战略转向的背景下,美国 CFIUS 审查、OISP 对外投资限制、2025 年国家安全备忘录均将 AI 列为强制审查领域。欧盟修订《外国直接投资审查条例》,亦将 AI 列为强制审查领域。而中国《外商投资安全审查办法》虽然涵盖 AI,但是缺少 AI Agent 的具体判定标准。本文用技术安全悖论来搭建框架,认为 AI Agent 的商业价值就在于它的自主性、适应性以及泛化能力,而正是这三种能力才成为它最大的军事转化风险,进而对比美欧中关于判定标准和实践的差别,并用 Shield AI、OpenAI、SpaceX 和 xAI 合并等实例来揭示目前的审查制度对于技术识别准确度、标准协调性以及动态适应性的三大结构性缺陷。在此基础上提出以功能能力为锚点的分级审查模型,把 AI Agent 的能力分解成自主性、连接性、适应性、协同性、积累性 5 个维度,并设定四级阈值,任一维度达到第三级就触发深度审查,两个达到第三级或者任一达到第四级就触发限制性措施,进而从制度、操作、动态三个层面提出完善中国审查框架的建议,使审查焦点从技术是什么转向技术能做什么,为完善外资安全审查中 AI Agent 军民两用判定标准提供理论依据和制度参考。

关键词: AI Agent;军民两用技术;外资安全审查;CFIUS;技术安全悖论

中图分类号: G301;F832.6

文献标识码: A

AI Agent 技术属于人工智能前沿方向,与传统大语言模型有本质区别。据保卫民主基金会(FDD)2026年3月的安全分析报告,AI Agent 是一种嵌入支架软件的自主行动系统,可规划执行多步骤任务、对外部系统采取操作,且无需人类持续监督^[1]。这种自主能力使其有别于传统 AI 应用,也对安全审查提出了新挑战。美国国会研究服务处(CRS)2026年2月的报告指出,美军及国会日益重视 Agentic AI 能力,美国国防部已于 2025 年 6 月发布 Agentic AI 用例概述及网络安全指南^[2]。

该技术在实战中已显现其军民两用性。Anthropic 公司 2025 年 8 月的案例显示,攻击者借助 Claude 自动化执行了一次网络间谍行动中 80% - 90% 的操作环节,目标涉及全球约 30 个组织。美国外交关系委员会(CFR)2026 年 1 月的分析认为,2025 年已出现 AI Agent 被用于大规模网络攻击、并

以生成式 AI 驱动大规模影响力行动的实证,表明此类滥用正从理论风险转变为现实威胁^[3]。可见 AI Agent 的安全边界远超传统软件。

在全球投资监管层面,各大经济体都在加快将 AI 纳入外资安全强制审查的范围。2025 年 2 月,特朗普政府发布 NSPM-3 号国家安全备忘录(“美国优先投资政策”),要求美国外国投资委员会(CFIUS)强化对 AI 等敏感技术投资的审查并将绿地投资纳入其中^[4],还专门提出制定新规以阻止美国企业和投资者投向推动中国军民融合战略的产业^[5]。欧盟方面,2025 年 12 月欧盟理事会与欧洲议会就修订后的《外国直接投资审查条例》达成协议,将 AI、量子技术、半导体等列为所有成员国必须强制审查的“超关键”技术领域^[6]。中国《外商投资安全审查办法》虽已涵盖重要信息技术、互联网产品与服务及关键技术,但对 AI Agent 这一新技术形态仍缺

收稿日期:2025-06-12;修回日期:2026-06-23

作者简介: 王文轩(1995-),女,蒙古族,馆员,博士研究生,E-mail: wangwenxuan221@mails.ucas.ac.cn。罗先觉(1966-),男,教授、博士生导师。

乏具体判定标准。

当前研究面临的核心难题是 AI Agent 军民两用界限不清。联合国裁军事务厅 (UNODA) 2025 年 9 月的“军事 AI、和平与安全对话”指出, AI 工具在加快信息解读的同时也压缩了人类审慎判断的时间^[7]。2025 年 4 月 AI Now 研究所的报告更尖锐地指出, 现行 AI 模型评估标准范围过窄、与安全关键应用缺乏相关性, 却被用来支撑无法证实的系统安全声明^[8]。这意味着传统上以产品类别或最终用途声明为依据的审查方法, 面对 AI Agent 已难以奏效。

本文的研究价值在于, 在 AI Agent 技术飞速发展、各国审查制度密集更新的窗口期, 系统分析现行审查标准的适用困境, 并提出具有技术锚定性的判定框架。本文适用技术安全悖论理论, 所揭示的核心矛盾在于: AI Agent 的商业价值正源于自主性、适应性与泛化能力, 而这些特征恰恰构成了最大的军事转化风险。由此, 审查制度的设计不应试图消除技术的双重性, 而应建立可识别并加以管控这种双重性的动态判定机制。

1 AI Agent 技术军民两用属性的技术解构

对 AI Agent 技术进行准确的军民两用判定, 必须从技术架构上理解它与传统 AI 系统之间的不同

点。根据 FDD 技术分析可知, AI Agent 系统一般由一个或者多个大语言模型嵌入支架软件组成, 该支架软件为它提供了一个规划和执行多步骤任务的工具, 即对外部系统采取行动^[1]。这就意味着 AI Agent 不单单是信息处理的工具, 它还可以在物理世界和数字世界里产生实际的效果。PMC (PubMed Central) 在 2025 年的综述文献中又将 Agentic AI 定义为“自主、适应、有目标的系统, 在动态环境中做主动决策”^[9]。

从军事应用的角度来考察, AI Agent 的技术特点同现代军事需求存在着非常高的结构性契合。根据 Grand View Research 发布的市场数据显示, 全球军事人工智能市场在 2024 年就已经达到 93.1 亿美元, 预计 2025 - 2030 年会以 13% 的年增长率持续扩大, 美国国防部 2025 年对 AI 研究的投入达到 49 亿美元, 主要用在自主系统和基于 AI 的数据分析上^[10]。美国空军 2025 年 4 月发布的 AFDN 25 - 1 号人工智能条令表明, 俄罗斯等国正在集中利用人工智能提高指挥控制、目标打击、侦察、电子战系统等各方面的效能, 加快创建以人机协同为特点的“智能化”军事力量^[11]。这些信息一起呈现出 AI Agent 技术在军事领域巨大的吸引力, 也就是它具有自主规划和工具使用的能力, 这正是智能化作战所需要的核心技术要素。

表 1 AI Agent 核心能力与军民两用特征的对应关系

Table 1 Correspondence between core capabilities of AI agents and military - civilian dual - use characteristics

技术能力维度	民用应用场景	潜在军事转化路径	安全审查关注等级
自主规划与多步骤执行	商业流程自动化, 客户服务	自主作战任务规划, 无人系统协同	高
工具调用与外部系统操作	API 集成, 数据库操作	武器系统接口控制, 网络攻击执行	极高
环境感知与动态适应	智能导航, 实时推荐	战场态势感知, 自适应打击	高
多智能体交互与自复制	分布式计算, 协作办公	蜂群作战, 自主网络渗透	极高
长期记忆与上下文推理	个性化服务, 知识管理	情报分析持续监控, 目标识别积累	中高

资料来源: 根据文献^[1, 2, 11] 综合整理。

FDD 的分析特别重视多智能体交互风险这个传统安全审查框架所不能涵盖的方面。不同于传统的软件系统, AI Agent 越来越多地被部署到与其他自主系统而不是人进行交互的场景当中, 并且还可以通过自复制的方式产生新的 Agent^[1]。在多智能体环境中, 单个系统的安全漏洞 (security compromise) 无法通过传统隔离策略被隔离, 被污染的指

令、被投毒的数据、被劫持的工作流会在互相连接的系统之间传播。这就表明, 即使一个 AI Agent 的投资在单个产品上已经通过了安全审查, 在它被集成到系统中之后所产生的“涌现性”安全风险, 也很难被现有的审查手段所预见。

Privacy International 在 2026 年 1 月关于 Shield AI 公司的案例分析, 提供了一个 AI Agent 技术军民

两用转化的典型样本。Shield AI 的核心产品 Hive-mind 属于 AI 自主飞行系统,官方声称可以“被训练来执行任意数量的任务”,即监视、空中作战和导弹搜索等^[12]。该公司 2024 年营业收入为 2.67 亿美元,比上一年增长了 64%,得到了 L3Harris 等国防科技企业以及硅谷风投的大力支持,Other People's Capital 基金也明确表示自己只投资军民两用技术企业。Shield AI 同 Palantir 合作研发出大规模自主无人机技术,并且把 Hivemind 系统部署到新加坡空军以及与中国台湾航空工业发展公司签订协议,充分体现出 AI Agent 技术怎样在商业投资逻辑的推动下,从民用领域顺畅地过渡到军用领域。

2 主要经济体 AI 投资安全审查制度的比较分析

2.1 美国:CFIUS 审查与 OISP 对外投资限制的双轨体系

美国形成了目前世界上最为复杂的 AI 投资安全审查体系,分为对内投资的 CFIUS 审查和对外投资的“对外投资安全计划”(OISP)限制两个并行的制度。CFIUS 的管辖权在 2018 年《外国投资风险审查现代化法案》(FIRRMA)通过之后发生了根本性的扩张,由原来的只审查外国对美国企业的“控制”,扩大到包括非控制性投资以及某些房地产交易^[13]。FIRRMA 创建了“TID 美国企业”,即技术、基础设施、数据三个方面的公司,任何外国对这些公司的投资只要取得相应的治理权,即使只是少数股权,也会被 CFIUS 管辖^[13]。

2026 年 1 月 Goffacts 就 CFIUS 对 AI 投资是否构成国家安全威胁的审查标准做了详细的分析。CFIUS 不公布判定交易是否具有国家安全风险的标准清单,法律也没有规定什么样的行为属于国家安全威胁,这就给了审查官员很大的自由裁量权^[14]。在实践中,CFIUS 评估三个维度,即威胁(threat)、脆弱性(vulnerability)和后果(consequence)。就 AI 投资来说,CFIUS 已经对这些交易的敏感性做出了明确的表述,大语言模型具有军民两用的特点,同一个优化商业流程的模型可以被用来打击军事目标或者进行情报分析,因此外国对 AI 开发的控制本身就带有军事意义^[14]。能够触发安全审查的特定投资者权利包括:获取重要机密技术信息、取得董事会席位或观察员席位、参与实质性经营决策、获取敏感个人

数据等。

NSPM-3 号备忘录在 CFIUS 现有的框架上做了很多重要的补充。Squire Patton Boggs 认为该备忘录要求财政部在和国会商讨的基础上,加强 CFIUS 对绿地投资的监管,并且特别提到 AI 领域。这一点对于投资实践来说影响巨大,之前一些双币基金在投资美国项目的时候会采用绿地投资的方式,即在初创企业刚成立的时候就介入其中,以此来规避 CFIUS 的审查,但是如果绿地投资被正式纳入到管辖体系中,那么这样的策略就不再可行了。对外投资方面,2025 年 1 月 2 日生效的 OISP 条例禁止或者要求美国主体就涉及“关注国家或地区”(目前为中国大陆、中国香港和澳门)的半导体、AI 和量子信息技术领域的特定投资向财政部进行申报^[15]。2025 年 12 月 18 日签署的 2026 财年国防授权法中加入的 COINS 法案,给 OISP 赋予了明确的立法授权,要求财政部考虑将管辖范围延伸到生物技术、高超音速、航空航天等各个方面^[15]。

2.2 欧盟:从分散审查走向强制统一的制度转型

欧盟的外资安全审查制度正在发生由自愿协调到强制统一的转变。2019 年生效的首部《外国直接投资审查条例》(Regulation (EU) 2019/452)本质上只是一种成员国之间信息交流和协调平台,并不强制要求所有成员国建立审查机制^[16]。分散模式对于 AI 等关键技术领域出现的系统性风险而言,就暴露出明显的制度缺陷。据欧盟委员会于 2025 年 10 月发布的 FDI 审查年度报告显示,2024 年欧盟成员国受理的审查案件(含投资者主动申报与成员国依职权发起两类)共计 3136 件,比上一年增长约 75%^[17]。但是整体干预率仍然很低,根据该报告,86% 的案件得到无条件批准^[16]。

2025 年 12 月 11 日,欧盟就修订前述《外国直接投资审查条例》达成的协议,标志着欧盟 FDI 审查制度的范式转换。Mayer Brown 认为修订后的条例要求成员国对具有战略重要性领域的所有外国投资进行审查,如与太空、国防有关的通用人工智能、量子技术、半导体等^[18]。Covington 的分析表明,强制审查的最低范围包含军民两用物项、军事装备、关键技术(如《AI 法案》所定义的人工智能)、关键原材料,以及能源、交通、数字基础设施领域的关键实体^[6]。另外一项重要的制度创新就是把审查范围扩大到由非欧盟投资者最终控制的欧盟子公司投资行为上,堵住了通过欧盟境内子公司间接投资欧

盟战略技术领域的规避途径^[16]。

与此同时,欧盟也开始在对外投资审查这一新的制度领域进行探索,在此之前,欧盟层面尚无针对本土资本对外投资的安全审查机制。2025 年 1 月 15 日,欧盟委员会发布建议书,要求成员国对本国企业到非欧盟国家投资进行审查,涉及的领域有半导体、人工智能、量子技术^[19]。成员国于 2026 年 6 月 30 日前须递交综合评价报告,其内容将会成为欧盟就该领域是否应再行立法的参考。这一动向同美国 OISP 的制度逻辑非常契合,显示出现阶段主要经济体在 AI 技术投资审查方面正在塑造起“入境审查 + 出境监管”的双向闭环态势。

2.3 中国:以开放为导向的审查框架及其适应性挑战

中国的外商投资安全审查制度是以 2020 年 12 月 29 日公布的《外商投资安全审查办法》为基本框架的。该办法将两类投资纳入审查范围,一类是投资军工、军工配套等关系国防安全的领域,另一类是投资关系国家安全的重要信息技术和互联网产品与服务、关键技术以及其他重要领域且取得所投资企业实际控制权的情形^[20]。中国审查框架的触发条件比美国 CFIUS 制度更重视“实际控制权”的取得,而不是像 FIRRMA 一样把非控制性的少数股权投资也纳入管辖范围。

在中国,对于 AI 技术的治理已经有了比较完善的制度。全国网络安全标准化技术委员会于 2024

年 9 月发布并于 2025 年 9 月迭代到 2.0 版的《人工智能安全治理框架》,把 AI 安全风险分为技术内生安全风险、技术应用安全风险和应用衍生安全风险三大类,提倡实行风险分级管理、敏捷治理^[21]。2023 年 7 月公布的《生成式人工智能服务管理暂行办法》是全球第一个专门对生成式人工智能服务管理作出规定且属于国家层面的法律法规,对于服务提供商而言需要完成安全评估并开展算法备案工作。配套国家标准有《生成式人工智能服务安全基本要求》(TC260 - 003 - 2024)和《网络安全技术生成式人工智能预训练和优化训练数据安全规范》(GB/T 45652 - 2025),把原则性要求变成可以衡量的技术指标,形成“法律加标准”的双轮驱动模式^[21]。

中国对于 AI 军事应用的国际治理是积极的。按照中国外交部 2025 年 4 月向联合国提交的文件,中国坚持“以人为本的军事人工智能”,坚持共同、综合、合作、可持续的安全观,2021 年就向《特定常规武器公约》提交了有关规范人工智能军事应用的立场文件^[22]。产业政策上,中国主张以军民共享为导向来部署算法、芯片、数据等新一代人工智能的支柱性技术的研发,创建起科研院所、高校、企业单位与军工单位之间的协调机制^[23]。由于军民融合的产业定位,客观上加大了外资在 AI 领域投资时所面对的安全审查的复杂程度。

表 2 中美欧 AI 投资安全审查制度核心要素比较
Table 2 Comparison on core elements of AI investment security regimes: China, US and EU

比较维度	美国	欧盟	中国
审查机构	CFIUS(财政部牵头)	成员国审查机构加欧委会协调	国家发展改革委加商务部牵头
触发门槛	控制性交易和特定非控制性投资	成员国自主设定(修订后强制审查)	实际控制权获取(军工领域除外)
AI 技术覆盖	TID 企业中的关键技术,OISP 限制	超关键技术中的 GPAI(太空/国防相关)	重要信息技术和关键技术(概括性规定)
对外投资管控	OISP(2025 年 1 月生效)	成员国评估阶段(2026 年 6 月前)	尚无专门制度
强制申报	涉及关键技术的外国政府投资	修订后对关键领域强制审查	军工相关强制,其他领域引导申报
审查时限	声明 30 天加调查 45 天加总统 15 天	成员国自定加欧委会 35 天意见期	一般审查 30 天加特别审查 60 天

资料来源:根据文献^[4,6,20]综合整理。

3 现有审查标准的结构性不足与批判性分析

从技术安全悖论的角度来观察,目前各个主要

经济体的 AI 投资安全审查制度存在三个结构上的困难,这三个困难并不是简单的执行问题,而是源自制度设计逻辑和科技发展规律之间的深层次矛盾。

其中,技术识别精度的难题是最根本的。目前的审查框架大多采用基于产品类别或者出口管制清

单的静态分类方式,但是 AI Agent 技术的军民两用属性并不取决于产品本身,而取决于它被使用的方式以及所连接的系统。CFIUS 相关法律有意未对“国家安全”作出明确定义,因为威胁的变化速度比法律要快^[14]。该种制度安排给审查机构留下了很大的裁量余地,但是也存在标准不明晰、缺乏预期的弊端。在 AI Agent 技术快速迭代的环境下,一个在审查时被判定为低风险的商业自动化 Agent,一旦得到新的工具调用权限或者接入军事级 API,就会立刻变成有武器化潜力的系统。根据 Arthur D. Little 2025 年 10 月的研究,各国已经把将军民两用技术当作战略资产,并且使用出口管制、安全审查以及主权投资策略等手段加强监管^[24]。但是这种监管趋势所面临的矛盾就是, AI Agent 的“双重性”不是可以被技术参数所界定出来的属性,而是由应用场景来决定的动态状态。

标准协调性困境在地缘政治竞争当中被不断放大。中伦律师事务所 2026 年 1 月的年度合规回顾表明,截至 2025 年 12 月 8 日,被列入美国出口管制与经济制裁黑名单的中国实体和个人一共是 2675 个,比上一年新增加了 475 个,增加了大约 22%,主要集中在半导体、人工智能、航空航天、新能源这些战略产业中^[25]。这些数据体现的不只是审查力度变大,更彰显了国家安全标准的异化,即“国家安全”标准正在被异化为技术遏制的政策工具。正如安全内参 2025 年所指出的那样,美国决策层把 AI 当成决定国运的重要变量,将国内的合规审查视为一种竞争劣势,于是产生了“在不损害产业扩张主要利益的前提下,依靠芯片管制等外部的技术封锁、去监管化,牺牲部分安全性换取绝对的发展速度”的竞争思想。

动态适应性困境就是审查制度对于技术更新速度的系统性滞后。AI Agent 技术能力呈指数级增长,CFR 引用的一个重要数据点是 Anthropic 的 Claude 两年前只能用 50% 的可靠性完成耗时两分钟的任务,但是 2025 年 11 月发布的新版本已经可以以 50% 的可靠性解决需要人类专家近五小时才能完成的复杂软件工程问题^[3]。美国云服务提供商计划在 2026 年对 AI 基础设施的投入达到 6000 亿美元,比 2024 年多出一倍。它们不是渐进式的改善,而是相变信号^[3]。在这种技术飞速发展的背景下,即使是在 2025 年制定的审查标准,到了 2026 年也会和实际的技术状况严重不符。美国 2026 财年

国防授权法要求国防部在 2026 年 4 月 1 日之前成立 AI 未来指导委员会,其任务之一就是制定出对于比现有的所有先进人工智能系统都更加先进的 AI 系统进行评估、治理和风险缓解的前瞻性政策^[2]。该制度安排就承认了目前的标准对于未来 AI 能力适用性的根本不确定性。

从技术安全悖论的角度来综合评价,这三种困境一起指向一个核心矛盾,即审查制度所追求的确定性、可预测性与 AI Agent 技术本质特征的不确定性、涌现性是一对矛盾。美国计算机协会(ACM)发表的 2025 年军事 AI 风险治理研究论文的结论具有启发性,即该论文构建的“四维协同”治理框架打破了传统军控的“以武器为中心”思维,超越了技术伦理的“乌托邦”幻想,直面“安全与发展”的两难困境^[27]。联合国裁军研究所(UNIDIR) 2025 年报告从军事领域多样性入手提出“拥抱多元性”治理思想,保证审查框架的包容性而不是排他性^[28]。这些研究说明,简单地提高审查门槛或者放宽审查范围,不能解决 AI Agent 军民两用判定的深层次问题,安全审查的制度创新要从方法论上实现突破。

4 案例分析:AI Agent 投资安全审查的实践困境

OpenAI 融资审查为认识 AI Agent 投资安全审查的实践复杂性提供了十分有力的实例。GovFacts 2026 年 1 月的分析认为,亚马逊对 OpenAI 的投资把 CFIUS 置于了一个不寻常的位置,它不是一件普通的外国收购,双方都是美国公司,但是现代风险投资的运作方式使得外国资金几乎参与了每一轮的重大融资^[14]。据称中东政府投资基金将会参与 OpenAI 的融资,在外国投资者通过融资获得了有意义的治理权之后,CFIUS 就会对交易展开审查。触发审查的外国投资者权利包括获得重要的机密技术信息、董事会成员资格或者观察员权利、参加实质性的决策等。OpenAI 在多个方面具有被审查的条件,大语言模型被政府认定为关键的技术类型,如果国防承包商或者政府机构使用了该技术,那么该 AI 系统就有可能涉及敏感政府或者军事数据^[14]。

SpaceX 和 xAI 的合并审查又证明了 AI Agent 技术对于安全审查来说至关重要。GovFacts 2026 年 2 月的分析认为,对该项交易进行国家安全审查的原因是 SpaceX 运营着军事和情报机构需要的

Starlink 卫星网络,而 xAI 开发出的先进 AI 系统可以处理或者操纵通过这些卫星传输的数据,合并后的公司可以将这些能力以单独公司无法实现的方式组合起来,产生对手可以利用的弱点^[29]。CFIUS 在 2025 年威胁评估中引用的国家有中国、俄罗斯、伊朗、朝鲜等,这些国家被用来作为审查判断地缘政治的参照。值得注意的是,CFIUS 在交易完成之后的 5 年内可以强制要求剥离,例如一家中国企业早在 2020 年就收购了视频显示技术公司 Jupiter Systems,却在 5 年后也就是 2025 年 1 月被特朗普总统下令进行剥离^[29]。

Shield AI 的案例从投资生态角度说明军民两用 AI Agent 企业怎样巧妙地在审查界线上游走。根据隐私国际的调查结果可知,Shield AI 属于风险资本资助的企业,得到了许多硅谷投资公司的支持,并且其游说预算在 2021 年至 2024 年间增长了约二十倍,2024 年已达 140 万美元^[12]。美国前国家情报总监、现任 CIA 局长 John Ratcliffe 曾在 2021 至 2022 年间担任 Shield AI 的顾问。Palantir 前政府关系负责人也在 2021 年加入 Shield AI 担任类似职务。企业结构模式是以风险资本融资为纽带、以旋转门人事关系为润滑剂,实现民用技术向军用场景快速转化的模式,给现有的审查框架带来了根本性的冲击。当审查的焦点放在交易结构本身以及直接投资者的身份上时,生态网络所引发的技术转移路径就很容易被审查视野所忽略。

从全球安全事件的角度来讲,民用 AI 技术进军军事领域已是必然的态势。2025 年 7 月 Global Security Review 分析了若干重大事件,德国北部警方追捕在核电站、化工厂和液化天然气终端上空盘旋

的无人机,无人机速度达到每小时 100 公里^[30]。2024 年乌克兰冲突期间生产了近 200 万架无人机,很多无人机装有 AI 导航、目标识别系统。武装组织用第一人称视角无人机对基础设施实施低成本、高影响打击,把军事和民用威胁界限模糊化了。这些事件表明,AI Agent 技术军民两用转化不只是理论上的风险,而是正在发生的现实安全威胁,投资审查制度必须对此做出有效的回应。

5 构建以功能能力为锚点的分级审查模型

根据对目前审查标准结构缺陷的批判性分析以及对典型案例的实证研究,本文提出以“功能能力”为标准的 AI Agent 军民两用分级审查模型。该模型的理论依据就是技术安全悖论的反向使用,既然 AI Agent 的军事风险来源于它的功能而不是产品类别,那么审查标准就应该根据能力参数而不是产品标签来确定。

模型的主要设计思路就是把 AI Agent 的功能能力分成 5 个可以量化的维度,而且给每一个维度设定一个由低到高 4 个等级的风险阈值。自主性维度评估 Agent 在无人类干预下独立完成任务链的长度(即连续自主执行的步骤数)和复杂度,连接性维度衡量 Agent 调用外部工具和操控外部系统的能力范围,适应性维度测试 Agent 在陌生环境中保持任务执行能力的泛化程度,协同性维度评估 Agent 与其他自主系统进行交互协作和信息共享的能力水平,积累性维度关注 Agent 通过长期记忆和持续学习提升自身能力的速度和上限。

表 3 AI Agent 功能能力分级审查指标框架

Table 3 Hierarchical assessment indicator framework of AI agent capabilities

能力维度	低风险(L1)	中风险(L2)	高风险(L3)	极高风险(L4)
自主性	单步骤任务执行	多步骤任务链规划	开放域自主决策	无限期自主运行
连接性	API 读取操作	API 读写操作	多系统协调操控	物理系统直接控制
适应性	固定规则响应	有限场景迁移	跨域环境适应	对抗性环境自适应
协同性	无多 Agent 交互	信息共享型协作	任务分工型协同	自组织蜂群行为
积累性	无学习能力	会话内学习	跨会话持续学习	自优化能力增长

资料来源:根据文献^[1,9,11]综合整理。

在审查触发机制上设置了明确的规则,任何一个维度达到 L3 级别就触发深度审查程序,两个或者

两个以上维度达到 L3 或者任何一个维度达到 L4 级别就触发限制性措施(附条件批准或者禁止)。以

功能能力为基础的触发机制,比现有的根据投资者国籍或者产品类别来触发的方法更具有技术精度以及更好的制度弹性。ACM 2025 年军事 AI 风险治理研究给出的风险分类逻辑,可以为本研究提供借鉴,其把 AI 侦察系统归入低风险范畴,并实行登记制度,把自主武器系统归入高风险范畴,并实行许可审批制度,而且提出了依靠技术发展自行调节监管强度的“动态合规”机制^[27]。

就中国外资安全审查制度的完善而言,本文模型的应用建议有三个方面。从制度上来说,在《外商投资安全审查办法》的实施细则中增加 AI Agent 技术专项审查指引,把上述五维能力评价体系纳入到“关键技术”的判定流程当中。从操作上来说,建议组建以技术专家、安全评估人员和产业经济分析师为主的技术、安全、经济三方面联合评估小组,对包含 AI Agent 技术在内的外商投资进行功能能力实测,而不是只看申报材料对技术的描述。从动态适应的角度出发,应该建立定期(每年至少一次)的阈值参数更新制度,参照国际上最新的技术发展以及安全事件的动态变化来重新确定各个方面的风险等级定义。中国的《人工智能安全治理框架 2.0》已经提出风险分级管理的治理理念,本模型就是该理念在外资安全审查这一特定制度场景中的具体实施方案。

6 结论

AI Agent 技术的军民两用性不是可以轻易判定的静态属性,而是由技术能力、应用场景、地缘政治环境三者共同构成的动态风险谱系。本文通过对美国 CFIUS/OISP 体系、欧盟修订后的《外国直接投资审查条例》、中国《外商投资安全审查办法》的制度逻辑及实践困境进行梳理,得出一个核心结论,即目前审查标准对于 AI Agent 技术的识别存在技术识别精度低、标准协调性差、动态适应性滞后等结构性缺陷。其根源在于传统审查思维把军民两用判定当作一个分类问题,在军用和民用之间划出一条清楚的界限,但是 AI Agent 技术的技术安全悖论特性使这个界限不在产品层面上,而体现在能力层面上。

本文提出以功能能力为锚点的分级审查模型,试图针对该困境提供一种具备可操作性的制度解决办法。该模型把审查重点从“这是什么技术”转向“这项技术能做什么”,用自主性、连接性、适应性、

协同性、积累性这 5 个维度的量化评价来达成对 AI Agent 军事转化风险的准确识别。该方法论转向的现实紧迫性表现在 2025 - 2026 年各国审查制度正处在密集调整期,美国扩大了 CFIUS 的管辖范围和 OISP 的覆盖范围,欧盟把修订后的《外国直接投资审查条例》转化为成员国立法,中国的《人工智能法》草案也正处在调研起草阶段。该时期将功能能力的判定标准嵌入到审查框架中来提高审查技术的精准度,并且可以为国际审查标准的协调提供方法论上的基础。

需要指出的是,本文的研究还存在着一定的不足。功能能力分级模型各个维度的阈值参数目前还处在定性描述的阶段,还没有通过大量的技术实测来校准。模型对于多 Agent 交互环境下涌现性的风险量化评价能力较差,这正是 FDD 报告所提及的安全问题。在地缘政治高度对抗的环境下,即使技术上合理的审查标准也会被政策目的所扭曲,在美国针对中国的对外投资审查(即“反向 CFIUS”)制度中就有所体现。未来可以利用技术红队测试(red teaming,即通过模拟对抗性攻击来检验系统安全性的测试方法)等方式对功能能力阈值进行实证校准,在多 Agent 交互的场景下研究 AI Agent 安全审查的方法,从多边框架的角度出发,推进各国审查标准最低限度的协调。

参考文献:

- [1] FDD (Foundation for Defense of Democracies). Regarding Security Considerations for Artificial Intelligence Agents[R]. March 9, 2026.
- [2] CRS (Congressional Research Service). Agentic Artificial Intelligence [R]. IF13151, Updated February 3, 2026.
- [3] Horowitz M C. How 2026 Could Decide the Future of Artificial Intelligence [R]. Council on Foreign Relations, January 12, 2026.
- [4] K&L Gates LLP. Trump Administration Directs CFIUS to Tighten Restrictions[R]. February 26, 2025.
- [5] Squire Patton Boggs LLP. Trump Administration Outlines Policy Changes to CFIUS and Outbound Investment Rules[R]. February 2025.
- [6] Covington & Burling LLP. The EU Adopts Revamped Regime to Screen Foreign Investment [R]. December 22, 2025.
- [7] UNODA (United Nations Office for Disarmament Affairs). Key Takeaways of the Military AI, Peace & Se-

- curity Dialogues 2025[R]. September 17, 2025.
- [8] AI Now Institute. Safety and War: Safety and Security Assurance of Military AI Systems[R]. April 23, 2025.
- [9] Adabara I, et al. A review of agentic AI in cybersecurity: cognitive autonomy, ethical governance, and quantum – resilient defense[J]. PMC, 2025.
- [10] SmartDev. AI in Military: Top Use Cases You Need to Know[R]. September 10, 2025.
- [11] USAF. Air Force Doctrine Note 25 – 1: Artificial Intelligence[R]. April 8, 2025.
- [12] Privacy International. Dual – Use Tech: The Shield AI Example[R]. January 16, 2026.
- [13] Harris Sliwoski LLP. U. S. Foreign Investment and the New CFIUS Rules[R]. November 12, 2025.
- [14] GovFacts. How CFIUS Decides Whether AI Investments Threaten National Security[R]. January 30, 2026.
- [15] Davis Polk & Wardwell LLP. U. S. Updates Outbound Investment Rule, Adopts BIOSECURE Act[R]. January 17, 2026.
- [16] Gleiss Lutz. EU FDI Screening Regulation Reform; Council and Parliament Reach Provisional Deal [R]. December 12, 2025.
- [17] A&O Shearman. EU FDI Screening Adapts to Evolving Geopolitical Risks[R]. November 10, 2025.
- [18] Mayer Brown. Shaping Investments into EU Strategic Sectors: The FDI Screening Reform[R]. March 2026.
- [19] EU Trade. Investment Screening[R]. European Commission, 2025.
- [20] 国家发展改革委, 商务部. 外商投资安全审查办法 (第 37 号令) [S]. 2020. National Development and Reform Commission, Ministry of Commerce. Measures for the Security Review of Foreign Investment (Order No. 37) [S]. 2020.
- [21] 安全内参. 发展与安全的双轮驱动: 中国人工智能立法演进与治理前瞻[R]. 2025 – 12 – 08. Anquan Neican. The Dual – Wheel Drive of Development and Security: the Evolution and Governance Outlook of China's AI Legislation[R]. December 8, 2025.
- [22] 中华人民共和国外交部. 中国根据联大第 79/239 (2024) 号决议就人工智能军事应用机遇和挑战问题提交的文件[R]. 2025 – 04 – 21. Ministry of Foreign Affairs of the People's Republic of China. Submission of China on the Opportunities and Challenges of Military Applications of Artificial Intelligence Pursuant to UN General Assembly Resolution 79/239 (2024) [R]. April 21, 2025.
- [23] 网信军民融合. 加强我国人工智能国防应用的研究与思考[J]. 网信军民融合, 2019;2. Cyberspace Civil – Military Integration. Research and Reflections on Strengthening the National Defense Application of Artificial Intelligence in China [J]. Cyberspace Civil – Military Integration, February 2019.
- [24] Little A D. Unlocking the Strategic Power of Dual – Use Technologies[R]. 2025.
- [25] 中伦律师事务所. 2025 国际贸易合规年度回顾及 2026 年展望 [R]. 2026 – 01 – 09. Zhong Lun Law Firm. 2025 Annual Review of International Trade Compliance and 2026 Outlook[R]. January 9, 2026.
- [26] 安全内参. 2025 年域外人工智能安全治理范式[R]. 2026 – 01 – 20. Anquan Neican. The Extraterritorial Paradigm of Artificial Intelligence Security Governance in 2025[R]. January 20, 2026.
- [27] ACM. Research on Military Artificial Intelligence Risk Governance[C]. Proceedings of the 2025 2nd International Conference on Digital Society and Artificial Intelligence, 2025.
- [28] UNIDIR. Artificial Intelligence in the Military Domain and Its Implications[R]. 2025.
- [29] GovFacts. How CFIUS Reviews Dual – Use Technology Mergers for National Security[R]. February 3, 2026.
- [30] Global Security Review. How Civilian Dual – Use Technologies Are Reshaping Global Security Policies [R]. July 17, 2025.

A study on the judgment criteria for civilian – military dual – use security review of foreign investment in AI agent technology

WANG Wen – xuan, LUO Xian – jue

(School of Public Policy and Management, University of Chinese Academy of Sciences)

Abstract: AI Agent technology is different from traditional large language models in that it can make autonomous decisions, perform multi – step tasks, and sense the environment; thus, it is now changing the world of technology. The AI Agent will use the large lan-

guage model in the scaffolding software to plan and carry out a series of tasks independently without continuous human intervention, and it can do more than just answer questions. The purpose of this architecture is dual – use, and it can be employed for civil and military applications; thus, the same framework of autonomous planning and tool invocation that has optimised commercial workflows can also be used in the military for command and control, intelligence analysis, autonomous weapons systems, etc. Therefore, the dual – use boundary is determined not by the category of goods but by how they are used and connected in the system; thus, the old way of classifying or declaring the end – use of goods is no longer effective. With the intensification of strategic competition between China and the United States and the changes in the economic security strategy of the European Union, many countries have begun to add artificial intelligence to the mandatory foreign investment security review. The Committee on Foreign Investment in the United States has expanded the scope of application for inbound investment under the Foreign Investment Risk Review Modernization Act, introduced the Outbound Investment Security Program to curb outbound investment in countries of concern, and by 2025, the National Security Presidential Memorandum will increase control over sensitive technologies such as artificial intelligence. The new Foreign Direct Investment Screening Regulation in the EU lists artificial intelligence, quantum technology and semiconductors as important areas to be monitored for all member states, and the EU has started to study the screening of outbound investments. China's Measures for the Security Review of Foreign Investment have covered significant amounts of information and key technologies in the field of information technology and core industries, but specific judgment criteria for the new form of technology, AI agents, have not been established. Building on the foundation of technological security paradox, this paper thinks that the commercial value of AI Agent technology is directly related to its autonomy, adaptability and generalisation ability; at the same time, these are also the reasons for concern. Based on the criteria and practice in the United States, the European Union and China, as well as through cases such as Shield AI, OpenAI and the merger of SpaceX and xAI, this paper introduces the three structural defects of the current review system. The first is the accuracy of technical identification; static classification cannot consider that the capability changes according to the application environment. The second is the coordination of norms, as the security of a country is increasingly used to achieve technological containment in the background of geopolitical competition. The third is dynamic adaptability; that is to say, with the exponential growth in agents' capabilities, the revision of review norms has been slow. A new hierarchical review model based on functions, rather than product labels, will be put forward in this paper. The five parts of the AI Agent model are Autonomy, Connectivity, Adaptability, Collaboration and Accumulation, and one of them has been assigned to each of the four risk levels. If any aspect reaches the third level, a deep audit will be conducted; if two or more aspects reach the third level or any single aspect reaches the fourth level, then strict measures will be taken. Some proposals for the construction and operation of the Chinese review system were also made in this paper, such as adding a particular review guide for AI agent technology, establishing a joint expert assessment team to conduct practical capability tests, and introducing a regular update mechanism for threshold parameters. By changing the focus of the review from what a technology is to what it can do, the model aims to provide a theoretical basis and an institutional reference for China to improve the judgment criteria for the civilian and military dual – use attributes of AI Agent technology in foreign investment security review.

Key words: AI agent; dual – use of civilian – military technology; foreign investment security review; CFIUS, technological security paradox