我国人工智能安全标准体系加速构建

《经济参考报》4月7日刊发记者叶健、吴蔚采写的文章《AI安全焦点追踪丨我国人工智能安全标准体系加速构建》。文章称，随着我国“人工智能+”行动的深入推进，各类智能体及AI应用广泛渗入生产生活场景。而近期频发的AI安全事件，不仅引发公众关注，也成为产业界与学界协同攻坚的重要方向。近日，全国网络安全标准化技术委员会（以下简称“网安标委”）正式组建“人工智能安全标准工作组”（WG9），标志着我国人工智能安全标准体系建设进入系统性推进阶段。

AI安全事件频发 攻防之战升级

近期，全球人工智能行业安全事件频发。3月底，人工智能公司Anthropic旗下AI编程工具Claude Code源代码泄露，这一事件被视为AI行业首次核心代码泄露事件。

奇安信安全专家章磊认为，综合各类公开信息以及Anthropic的官方回应分析，此次源代码泄露是典型的发布流程中的人为失误，属于供应链安全事故。“好比原本只该给顾客成品，结果把全套制作图纸一起送出去了。”

“产品的核心逻辑和防护底线一旦公开，整个产品的运作方式就变得透明。竞争对手可以直接研究它的架构、功能设计、智能体逻辑，能快速模仿、追赶甚至优化。同时，安全规则暴露后，也更容易被人找到漏洞、绕过约束、破解使用限制。”章磊表示。

今年以来爆火的智能体工具OpenClaw（俗称“龙虾”）也接连被曝出存在多重安全隐患。4月3日，国家信息安全漏洞库（CNNVD）发布通报称，自3月10日至4月2日，共采集OpenClaw漏洞155个，其中超危漏洞11个、高危漏洞53个，OpenClaw多个版本受到漏洞影响。

“我们只花了一个下午，就攻破了OpenClaw。”国内知名白帽安全团队DARKNAVY安全创新总监陆晨表示，目前，国内主流的“龙虾”方案分为两类：一类是在OpenClaw基础上套壳提供对话框，另一类是提供服务器供用户自行配置。相比较而言，前者风险更高，一旦被攻破，黑客就能直接获取服务器权限，甚至访问内网大模型。

上海交通大学安泰经济与管理学院副院长刘少轩透露，近期一家制造业企业因为仓促上马OpenClaw，导致产线停产72小时，直接损失可能超过2000万元。还有一家法律服务企业，因为没有做好风险防范和数据安全，导致大量客户隐私数据泄露。

亚信安全相关负责人也指出，当前网络攻击正在向智能化、自动化演进，黑客利用AI实现勒索软件载荷的动态生成、高仿真钓鱼内容制作，使得攻击效率与隐蔽性大幅提升。AI自主攻击智能体、基于深度伪造的商务诈骗，将成为2026年最紧迫的安全挑战。

AI安全供给发力 需求创造机遇

国资委1月底发布的数据显示，中央企业已在工业制造、能源电力、智能网联汽车等重点领域，打造了超过1000个AI应用场景，AI赋能产业转型的态势日益明显。与此同时，AI安全问题引发的行业担忧，也催生了全新的安全需求，推动AI安全供给侧持续发力。

对此，东莞证券认为，近期OpenClaw等智能体技术快速落地，催生全新安全需求场景，叠加网络安全领域政策利好持续释放，行业有望迎来新的增长机遇。

长江证券则预测，2026年国内网络安全市场规模有望突破1500亿元，2030年可达3000亿元，年复合增长率达18%至20%，行业正处于发展黄金期。

同时，AI安全新品与服务也在持续发布。近日，上海人工智能实验室推出高安全产业级智能体平台SafeClaw，聚焦高安全需求的产业智能化转型，以推动行业从“事后安全”迈向“内生安全”的路径。同时，上海人工智能实验室还开源了能快速诊断风险的智能体守卫模型，并探索将安全准则内嵌至智能体决策层的“内生进化”治理框架。

“最危险的并非已知风险，而是‘没有想到的危险’，因此，当前的核心任务是，在AI能力飙升的同时，前瞻性地构建内生安全体系。”上海人工智能实验室领军科学家胡侠表示，“这些工作旨在将安全能力深度融入AI发展全链条，为智能体时代的‘内生安全’提供系统性解决方案。”

AI治理持续完善 安全标准加速制定

随着人工智能被广泛应用，人工智能治理也越发受到重视。今年政府工作报告明确提出“完善人工智能治理”，全国人大常委会工作报告提出“加强人工智能领域立法研究”。

在此背景下，人工智能安全标准正在加快制定。3月25日，工信部公开征求《人工智能安全治理模型上下文协议应用安全要求》等行业标准计划项目意见。

4月初，人工智能安全标准工作组（WG9）表示，将重点推动《网络安全技术 人工智能安全能力成熟度评估方法》《网络安全技术 人工智能应用安全分类分级方法》及《网络安全技术 人工智能技术涉及未成年人应用安全指南》等核心标准的落地实施。同时，在全国网安标委统一部署下，集中力量攻坚内生安全与数据基座、新形态与服务安全、系统与应用安全及科学评测等领域的国家标准。

“针对AI带来的新型风险，需从政策法规、技术标准、实施机制三个层面协同推进。”奇安信副总裁张勇认为，展望未来，第一，安全将从“可选配”升级为“必标配”，安全合规将从推荐性转向强制性，可以设定“AI安全投入不低于AI应用总投入15%”这样的行业基准；第二，网络安全从“单点防护”走向“全链条协同”，实现“一处发现攻击，全网自动免疫”；第三，从“人防”走向“技防+智防”，AI对抗AI成为攻防常态；第四，从“被动应急”走向“主动免疫”，构建韧性防御体系，实现“即使遭受攻击也能快速恢复、核心数据不丢失”。

新华社北京4月7日电

惩治“机闹”犯罪 两高发文守护民航飞行平安

新华社记者 齐琪 刘硕

民航安全无小事，法治之力护平安。

4月8日，最高人民法院、最高人民检察院联合发布办理危害民航飞行安全刑事案件的司法解释，用严明的法律标尺衡量行为边界，守护飞机上万千乘客的安全。

依法惩治“机闹”犯罪行为——

违规开启民航飞机应急出口舱门，在机舱内打架斗殴，对乘务人员使用暴力……这些行为严重影响飞行安全。

如何判定这些行为属于行政违法行为，还是构成刑事犯罪？

对此，司法解释明确，在民航飞机处于依靠自身动力移动期间或者空中飞行期间违规开启舱门、足以引发危害公共安全危险的情况下，以刑法中的“以危险方法危害公共安全罪”定罪处罚；对于飞机尚未依靠自身动力移动等情况下违规开启舱门的行为，可以根据有关规定给予行政处罚，并由行为人承担相应的民事赔偿责任。

最高法刑四庭庭长罗国良介绍，司法解释采用列举式规定，对在飞行中的民航飞机上使用暴力行为构成暴力危及飞行安全罪的定罪量刑标准作出了规定，特别明确乘务员即通常所称的“空姐空少”属于“飞行安全保障人员”，是暴力危及飞行安全罪的犯罪对象，对飞机乘务员使用暴力的行为可能构成暴力危及飞行安全罪。

从严惩治“造谣”涉民航飞行安全犯罪——

口无遮拦谎称“飞机有炸弹”，会有哪些严重后果？

2023年8月，陈某波的妻女因超过值机时间而未能登机。陈某波同日报警要求为其妻女办理机票改签，未果后心生不满，在报警电话中编造航班上有炸弹的虚假恐怖信息，导致机场启动一级应急预案，后续多个航班延误。最终，陈某波以编造虚假恐怖信息罪被判处有期徒刑一年。

陈某波虽被依法惩处，但在司法解释出台前，相关法律适用问题比较复杂，编造、故意传播涉民航飞行安全虚假恐怖信息在什么情况下可以判处五年以上有期徒刑没有明确标准，影响执法司法实践中对此类行为打击的精准度和惩治的威慑力。

对此，司法解释突出对编造、故意传播涉民航飞行安全虚假恐怖信息犯罪的从严惩治，规定行为人的行为影响民航航班、民用机场正常运行，或者致使公安、武警、消防救援、卫生检疫等部门采取应对措施的，应作犯罪处理；结合民航飞行安全领域的实际情况及危害行为的特点，明确列举六种情形，具有这六种情形的，属于造成严重后果，处五年以上有期徒刑。刑事惩处更为严格了。

司法解释同时明确，无论是采取明示还是暗示的方式编造、故意传播涉民航飞行安全虚假恐怖信息，符合相关条件的，均可构成编造、故意传播虚假恐怖信息罪。

坚持依法从严，贯彻宽严相济刑事政策——

编造、故意传播涉民航飞行安全虚假恐怖信息行为严重影响社会公众安全感和民航行业健康发展，甚至可能造成一定范围内的社会恐慌心理，必须坚持依法从严惩处总体原则不动摇。

据介绍，对出于仇视报复社会等动机和目的，经周密预谋策划后实施编造、故意传播涉民航飞行安全虚假恐怖信息行为，意图制造社会不稳定因素的行为人，以及多次实施此类行为屡教不改的行为人，依法予以严惩。对于行为情节一般，且具有自首、坦白、认罪认罚等法定或者酌定从宽处罚情节的行为人，可以依法适度予以从宽处理，确保罪责刑相适应。

民航飞行安全底线不容触碰。人人都应敬畏法律、遵守规则、文明出行，共同维护民航飞行安全和社会公共安全，让每一次起降都顺利、平安。

新华社北京4月8日电