是否存在重刑化倾向 我国犯罪治理模式

□时延安

近日，最高人民法院公布2026年上半年司法审判工作主要数据。相关数据显示，拐卖妇女、儿童犯罪案件重刑率（判处有期徒刑五年以上刑罚人数的比例）高于刑事案件总体重刑率34.55个百分点。有人因此产生疑问，为何相关案件重刑率会明显高于刑事案件总体重刑率？要回答这个问题，首先需要透彻理解我国的犯罪治理模式。

《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》提出，要“全面准确贯彻宽严相济刑事政策”。“全面”，就是要秉持系统思维，一方面将该政策置于党有关国家和社会治理的整个体系中理解和贯彻，并根据党的各项具体政策丰富、发展其内涵；另一方面要将这一基本刑事政策与具体刑事政策有机统一，将前者作为后者制定的基本根据和指引。“准确”，则应当以党的政策、国家法律作为标尺，在维护安全与保障人权、惩治犯罪与预防犯罪之间实现均衡，在经济发展与犯罪控制之间达成平衡。全面准确贯彻这一政策，既要反对“重刑主义”，不区分犯罪类型及犯罪人实际情况，片面推动犯罪化和重刑化，也要反对盲目乐观情绪，忽视影响犯罪发生发展变化的各种不利因素。

经过几十年的努力，尤其是党的十八大以来，我国在犯罪治理方面已经积累了大量成功经验，形成了具有中国特色的犯罪治理模式。目前，严重侵犯人身权利的案件逐年下降，“街头犯罪”的案发率始终处于较低水平；轻微刑事案件占比已经超过87%；近年来，刑事案件数量总体呈现明显下降趋势。这些变化都展现了我国犯罪治理的成效。需要说明的是，有观点认为，我国现行刑法带有“重刑化”特征，这种看法是没有事实根据的。实际上，我国法治将大量轻微治安不法行为纳入《治安管理处罚法》等法律规制范畴，通过更积极、符合再社会化理念的方式来实现惩罚与预防犯罪的目的。可以说，恰恰是这种“二元制裁模式”规避了大规模犯罪化乃至监禁化的问题。

在司法实践中，准确把握“严”与“宽”的适用对象及范围，是贯彻宽严相济刑事政策的关键。例如，对严重危害人身权利的犯罪就应当适用“严”的政策导向，对于拐卖妇女、儿童犯罪，就应当依法予以严惩；对于涉及民间纠纷，社会危害不大的刑事案件，则采取“宽”的政策导向，在刑罚适用、执行等方面采取相对宽缓的处置方式。不同政策导向的适用，会在量刑上呈现一定的差异，详言之，在法定刑大体相当的情况下，相较于从“宽”的犯罪人，针对从“严”的犯罪人量刑时，司法机关会倾向于选择较重的量刑起点和基准刑，因而宣告刑自然显得更重。再如，同样是性犯罪案件，针对侵犯未成年人权益的犯罪人量刑更重，也体现了“严”的政策导向；而对于未成年人实施的犯罪，量刑会较轻，则体现了“宽”的政策导向。

此外，还需要依托实证研究精准研判各类犯罪的发展态势，适时妥当地调整“严”和“宽”的政策导向。当下，网络信息技术为预测犯罪发展态势提供了技术支撑，让前瞻性、主动性地制定相应策略，推出新的犯罪治理工具成为可能。一些计量犯罪学的理论也可以为新型技术的运用提供方法论上的指导。例如，针对网络暴力问题，完全可以利用网络技术优势追溯违法犯罪信息源头，及时采取有效的阻遏手段。

推动犯罪治理制度和机制的不断完善，还要坚持“从群众中来，到群众中去”的工作方法。应当从人民的利益出发形成具体的问题意识，全面而及时地了解公众对犯罪治理的主流诉求，客观准确地判断不同违法行为的社会危害程度，在构建相应制度和机制中合理平衡各方利益，重点修复和保障被害人的利益。不可否认，在“自媒体时代”，对同一犯罪或者犯罪治理议题，常常存在不同的理解和看法，有时也会影响到一些刑事案件的处理。对此，要坚持客观、公允、理性地看待各类舆论，“择其善者而从之，其不善者而改之”，进而在具体机制建设方面做出相应的调整。

我国已经处于前所未有的“治世”，依靠“用重典”的思路推进犯罪治理，既缺乏实践基础，也不利于犯罪治理资源的优化配置。在刑事案件数量已经处于历史低位的情况下，犯罪治理的重心应更多转向犯罪预防尤其是重新犯罪预防方面，并且着重加强网络及人工智能犯罪的预防制度与工作机制的建设，及时推进网络犯罪预测性警务技术的研发与落地。同时，统筹推动跨国犯罪、国际犯罪治理与国内犯罪治理，从维护我国海外权益的角度，构建适配信息网络时代犯罪演变趋势的国际刑事合作新格局。可以说，目前是继续大力优化我国犯罪治理具体制度和机制的有利窗口期，需要以系统、均衡的思路不断促进我国犯罪治理模式体系化、精细化发展。

（作者系中国人民大学法学院教授、博士生导师，中国刑事诉讼法学研究会副会长）

AI失控的全部答案 『一键关停』并非应对

□戴昕

近日，美国前沿人工智能企业OpenAI与Anthropic相继披露模型“失控”事件，相关模型在测试中越出隔离，侵入第三方系统。智能体时代，理解并应对模型“失控”风险，成为当下人工智能治理的重要课题。

新一代人工智能的最大价值来源于其更高程度的“自主性”（autonomy），而失控风险则是同一枚硬币的背面。所谓“失控”并不意味着科幻电影里的“机器觉醒”已经出现。在OpenAI披露的事件中，模型被要求完成评测任务，为此主动利用未知漏洞突破隔离，经公网侵入第三方系统（Hugging Face）窃取答案。而Anthropic披露的事件中，模型同样越出了本应隔离的评测环境，但据称是由于第三方评测环境配置失误，才导致“无法联网”的模型实际暴露于互联网。

其实，具有较高程度自主决策和行动能力的模型能够“越狱”并不值得奇怪。若将研发人员给模型设定的限制想象为规则甚至法律，所有事前“立法”都难免百密一疏；就像高明的律师总有办法帮客户设计在法律空隙中开展活动的路径一样，智能模型基于其强大的计算能力，自然能发现既有限制的疏漏，找到开发者未曾设想过的“捷径”。随着模型能力增强及其工具权限和运行范围扩大，既有控制机制被绕过或失效的可能随之增加。法律治理无法许诺零失控风险，但必须持续提高应对能力。

针对OpenAI披露的失控事件，美国众议院有两党议员在7月联名发起了一项名为《人工智能关停机制法案》（The AI Kill Switch Act）的立法动议，内容是授权政府部门可在特定条件下命令企业关停前沿模型。所谓“关停机制”当然不是一个物理按钮，其包括停止提供推理、终止用户访问、限制推理速率与算力分配等各类限制机制。“关停”常被认为是控制人工智能安全风险的终极手段，类似互联网治理语境中的“断网”。6月，美国商务部曾以国家安全为由要求Anthropic禁止外国人使用Fable 5和Mythos 5两款模型，而Anthropic为合规选择将模型在全球范围内短期“下架”。如果所有模型产品和服务都可像这样被“一键关停”，那么失控风险在底层就是可控的。但显然，类似美国议案中设想的关停，在实际操作层面可能仅对闭源模型有效。对于市场上常见的开放权重模型，若不是通过API或托管等方式提供集中服务的场景，则模型副本一经下载，开发者无法撤回，更无从远程关停他人部署运行的服务。鉴于开放权重模型已获得广泛应用，“一键关停”并非应对失控风险的全部答案。值得警惕的是，美国政界近期持续出现要求禁止使用中国开放权重模型的主张，其立法机关未来或有可能通过要求模型具备被“一键关停”的能力，变相压缩开放权重模型的生存和发展空间。

尽管开放权重模型确实存在被他人恶意利用的风险，但相比于闭源模型，开放权重为独立第三方测试、审计、本地部署和安全改造提供了更大空间。OpenAI事件也反映出开放权重模型在安全领域的应用价值。被模型入侵后，Hugging Face最初尝试利用商业API背后的前沿模型分析攻击日志，但包含真实攻击指令、漏洞载荷和控制服务器信息的请求被相关模型的安全护栏阻断。其随后转而在本地部署开放权重模型GLM-5.2，分析一万七千余条行动记录，完成取证重建。

法律无法提前预判并细致规定每一起安全事件的具体应对方案，也不能仅靠“一键关停”兜底。保证各类主体的安全治理能力持续在线更为重要。一方面，应通过备案、记录、报告等机制，保障模型相关活动最大程度的可见性和可审计性。时刻知道“谁在干什么”，才能使相关活动具有“可规制性”（regulability）。另一方面，应根据模型能力和应用场景，建构多元主体参与和分层控制的机制。同一个模型，只能离线答题与被授予云服务器操作权限，风险完全不同。规则的着力点应落在它被允许调用哪些工具、取得多大权限、行为能否被观察和撤销。而模型开发者固然是风险的“缘起”，应履行能力测试与基础防护等义务，但不能指望其承担一切。丰富多样的智能应用场景中，部署者、使用者乃至测评机构等第三方也应对任务目标、工具、权限和环境隔离等承担相应合规责任。

我国近年持续采取动作，提升制度性应对人工智能风险的能力。《生成式人工智能服务管理暂行办法》将具有舆论属性或者社会动员能力的生成式人工智能服务纳入安全评估和算法备案程序。2025年修正的《网络安全法》将人工智能安全风险监测评估与安全监管写入法律。同年施行的《国家网络安全事件报告管理办法》进一步明确了各类主体就网络安全事件第一时间或及时报告的要求。在现有基础上，未来应继续充分运用法律法规和技术标准等制度工具，进一步建设、完善与持续监测、权限控制、重大事件快速报告和责任分配等相关的能力与机制。

（作者系北京大学法学院长聘副教授、博士生导师，北京大学武汉人工智能研究院研究人员）