

管控“强人工智能” 面临法律悖论

□季卫东

ChatGPT 的问世，是人工智能发展进程中的重大转折点，甚至可以称之为科技奇点，从此 AI 由弱转强并进入物种大爆发阶段。特别是今年 3 月 15 日发布的 GPT-4，在很多方面俨然达到甚至超越人类的聪慧水准。但是，两天后斯坦福大学教授米哈尔·科辛斯基公开了他的惊人发现：GPT-4 不仅具有更明显的机器觉醒的迹象，还暴露出试图摆脱人类控制的主观动机和潜力。这意味着阿西莫夫关于机器人研发的第一定律（不许伤害人）和第二定律（服从人命令）正面临极其严峻的挑战。

进入 4 月，一个更加令人不寒而栗的事实被披露，加州大学洛杉矶分校尤金·沃洛克教授的实验性研究发现，ChatGPT 居然能够捏造法学家性骚扰的丑闻及其信息来源，指名道姓进行有鼻子有眼的诽谤。舆情瞬间铺天盖地，让强人工智能继续进化还是就地止步，突然成为法律判断的一个重要问题。

尽管如此，与 ChatGPT 类似的强 AI 研发活动并未停步。中国也正在涌现出各种各样的大模型，例如百度的“文心”、华为的“盘古”等等。显然，在这场强 AI 研发竞争中，并不缺少后来居上的积极性，各种新兴势力紧紧抓住弯道超车的契机不放。然而，ChatGPT 毕竟带来了一系列重大问题和风险，例如取代人类工作、自主武器伤害、社会工程攻击、能源消耗代价等等，都需要人类认真对待。据斯坦福大学 HAI（人本人工智能研究所）2023

年报告，73%的 AI 专家相信 AI 将引发社会革命，但也有 36%的人认为 AI 有可能造成核弹级别的不可逆灾难。正因为存在小概率、大危害的恐怖，才必须采取有力举措，防患于未然。

GPT-4 以及类似“强人工智能”的本质是大规模语言生成模型和 AI 生成内容。相关研究表明，这类新型人工智能将在 2026 年之前耗尽高质量人类语言数据、在 2030-2040 期间耗尽所有人类语言数据。这意味着在今后十余年里 AI 合成数据的占比越来越高，直到彻底取代现实世界的语言数据。如果此后喂食人工智能的主要数据养料是人工智能合成数据，AIGC 的可信度究竟如何、会不会垄断人类沟通的语境就成为一个非常严重问题。

何况深度的机器学习会使 AI 逐步远离人类的介入和掌控，算法也将变得难以解释、不可理解甚至无法控制，从而导致大模型反噬人类的噩梦。这意味着人工智能治理的重点势必从算法治理转向大模型治理，即人工智能治理标准将从算法偏见最小化转向模型滥用最小化——预测的准确度将决定一切。

以 2021 年《新一代人工智能伦理规范》为标志，中国不断加强对人工智能开发活动的治理，强调可信、可控、可问责以及敏捷应对的基本原则。在这个规范性文件指导下，有关部门着手研究制定算法伦理规范并推动企业形成共识，并试图通过算法社区、行业联盟等方式搭建内部治理框架和责任体系。

而我国人工智能产业的“清朗行动计划”，也已从当年的算法滥用专项治

理推进到了 2022 年的算法综合治理阶段，计划在三年时间里把相关的伦理原则、法律规则以及技术性测试、监管、评价方法落到实处。2022 年 3 月 1 日颁布的《互联网信息服务算法推荐管理规定》还确立了算法备案制，以确保监管部门在源头上控制人工智能风险。

以联合国科教文组织推出《人工智能伦理问题建议书》为背景，中共中央、国务院在去年 3 月 20 日颁发《关于加强科技伦理治理的意见》。随着 ChatGPT 引发的全球性争议日趋尖锐，今年 4 月，科技部公布《科技伦理审查办法（试行）》；网信办也拿出《生成式人工智能服务管理办法》草稿征求意见，希望通过备案制、专家复核程序、明确责任主体、拟订风险清单、进行抽查核验等方法加强对 AI 大模型研发和应用的监管，防止人工智能系统失控。

但是，当 ChatGPT 已经开始应用于司法和法律服务场景时，算法会不断形成并加强尼克拉斯·卢曼所强调的不受人类控制的“自反身机制”，并在完全没有理解涵义、没有进行思考的状况下与人类进行法律沟通，影响人类的判断和决策。因此，制定和实施关于强人工智能的法律规则，势必面临更加复杂的问题。针对这类悖论，计算法学研究者以及人工智能治理机制设计者应通盘考虑，尽早未雨绸缪。

（作者系上海交通大学文科资深教授、博导、中国法与社会研究院院长、人工智能治理与法律研究中心主任，中国计算机学会计算法学分会会长，中国法学会法学教育研究会副会长）

监管AIGC,我们要立怎样的法？

□朱福勇

与以往的专家系统或者领域人工智能不同，ChatGPT 展现出人工智能不亚于人类的智力水平，用户与之对话，在互动中通过程序创造新内容，激发创造力，产生难以置信的人机交互效果。但巨大红利呼啸而至的同时，普遍性挑战也附随而来。尤其是不当使用内容生成式人工智能（AIGC）可能会进一步加剧社会不公、固化或者扩大偏见与歧视、侵害个人隐私、提供误导性信息、威胁数据安全、人类安全和成为违法活动或不道德行为的工具。如比利时公民皮埃尔因使用 ChatGPT 诱发自杀，同时还出现类似软件性骚扰用户、泄露用户数据等情形，均引发重大的伦理关切。意大利个人数据保护局认为 ChatGPT 违反了欧盟《通用数据保护条例》，以国家名义将其禁用。

各国对人工智能的立法监管正如火如荼地展开，今年被认为是人工智能立法监管的元年。尽管不同的国家或法域采取的立法与监管模式不同，但是依循算法治理——深度合成治理——人工智能立法的思路，如何确保人工智能技术安全、有序、规范和可持续向善发展，是一个需要人类共同直面的课题。

早上个世纪，阿西莫夫提出“三大定律”，已成为人工智能伦理和道德研究中的重要参考，也是立法监管 AIGC 的重要规范指引。从全方位规范人工智能研发、应用等主体行为以及技术规范与进步平衡的角度出发，相关立

法应遵循如下原则：

创新导向原则。鼓励创新，减少对人工智能研发的限制，以创造更多的科技进步机会和空间，推进人类的发展。当然，自由研发并非是恣意妄为，基于国家和社会安全以及人类伦理等因素考量，对于诸如军事、生物技术和可能威胁人类基本权利等敏感领域的人工智能技术研发与应用，则应强化监管力度以保障人类自身的发展和福祉。

合法必要原则。综合考量国家安全，社会稳定，个人权益保护等，注重立法的系统与周延。特别要明确风险规制的目标、规制机关的权力、职责、参与主体的权利义务关系，科技企业的合规监管；对风险防控的具体操作，设定自愿性标准供企业选择；合理平衡各主体间的权益，避免对创新过重苛责。对在既有法律框架下能够解决的风险问题，则无需单独、重复立法。

安全可信原则。为保护人类和环境免受人工智能技术潜在风险的危害，必须确保 AIGC 技术和系统的安全、可信。即增加人工智能系统的透明度，确保能被理解和解释；需要找到相关责任主体时，能被追踪和追溯；不会造成严重不良影响和危害、不被滥用，提高人工智能设计和运行的安全性；保护用户个人信息和隐私，确保不被非法收集、利用或泄露；避免产生歧视和不公平情形，保证公正平等待遇；确保技术可控，技术成果共享，不被少数人垄断。

风险防范原则。在技术研发、应用场景选择、数据采集和隐私保护等方

面，立法设置监督关口，预防潜在的风险和危害，确保技术的安全和可靠性。最大程度从源头降低技术可能带来的风险和危害，同时加强事前风险评估和风险管理，对可能的风险和危害事前预备应对措施，保障技术有序发展。

多主体全领域监管原则。人工智能技术更新迭代迅猛，立法跟不上技术前进的脚步。需要建立一个多元化的监管体系，赋予不同主体监管权力和义务，从不同维度对技术加以规范和引导。政府是主要的监管主体，但不应是唯一的主体，以技术治理技术是较好的监管方式，需要发挥行业自律和企业合规建设的功效，使立法监管更加灵活高效，有效促进和规范技术的不断进步，同时保障公共利益和社会安全。

总体而言，AIGC 的技术监管与创新对立而统一。监管在确保技术向着符合人类社会和道德准则方向发展，尽可能减少负面影响，为人类带来更大福祉的同时，还可以促进技术发展和主体间的竞争，推动 AIGC 技术的进步。受技术发展水平、应用领域和场景、社会影响和公众关注度等因素制约，通常在人工智能技术发展初期，不宜采取过于严苛的监管措施；当技术成熟、市场需求增长、风险和问题日益突出的时候，则应加大监管强度。当前情势下，坚守价值引领与技术“纠错”并行、秩序保障与创新提升协同、权益维护与义务承担对称，无疑是最恰当的平衡之道。

（作者系西南政法大学人工智能法学院教授、博导）

□皮勇

日前，苏州某大学学生赵某因在网上发布虚假 P 图侮辱女性被处以行政拘留十天，此事引发民众讨论。社交网络上代表性意见认为处罚太过轻微，司法机关应当追究赵某刑事责任，提供的参考案例是 2020 年杭州“女子取快递被造谣案”的造谣者被认定构成诽谤罪。这种观点折射的公众情绪可以理解，但若要适用刑法治理网上违法发布信息的行为，必须严格遵守罪刑法定、罚当其罪的刑法基本原则。

侮辱罪、诽谤罪是侵犯公民人身权利的犯罪，是否构成情节严重关系到是民事侵权行为还是应适用刑法制裁的认定，应以对被害人人身权利的侵犯程度进行评价，同时考虑对社会公共秩序的影响。

赵某通过文字配图片的方式捏造事实侮辱他人，并通过互联网传播侮辱信息，使得信息接收者更易轻信。较之传统社会环境下的侮辱行为，其传播范围广泛且难以消除影响，给被害人声誉造成更严重的损害，同时还扰乱信息网络空间秩序，损害社会公序良俗。当被害人发现是赵某所为找其理论时，后者并未道歉悔改以争取谅解，而是置若罔闻不予理睬，给被害人身心造成再度伤害，表明其主观恶性较大。如果能够证实以上侮辱信息的传播范围达到较广的程度，如浏览、转发、评论次数达到千次以上，或者给被害人的正常生活、学习、健康造成严重损害后果，综合上述主客观方面的事实，即可评价为符合情节严重的构成要件，进而可以诽谤罪追究他的刑事责任。

但需要注意的是，该罪须告诉才处理。被害人念及与其曾是同学好友及其他考虑，没有向司法机关提起诉讼，且赵某的行为尚不属于“严重危害社会秩序和国家利益的行为”，不符合以公诉的方式追究其刑事责任的法定程序条件，因此司法机关不能越界提起公诉。

网络空间是新的社会空间，当一些用户故意发布虚假、侮辱性信息，造成恶劣的社会影响后，就有必要对这些传播违法信息的行为进行法律规制。但是应当遵守法治原则，对不同类型、不同危害程度的行为依法合理适用刑法或行政法。

侮辱罪、诽谤罪是侵犯公民人身权利的犯罪，应指向特定的受害者。如果相关虚假信息没有明确指向特定人，或综合其他信息也不能明确特定人，若有“被害人”认为该信息对其构成侮辱或诽谤的，由于社会实际并未对其声誉产生消极评价，此时“被害人”的精神伤害及由此形成的其他后果在客观上不能归因于该虚假信息，就不能将散布该信息的行为定性为侮辱、诽谤行为，进而追究其侮辱罪、诽谤罪的刑事责任。

有观点认为，以上行为扰乱信息网络秩序，可按寻衅滋事罪追究刑事责任。2013 年“两高”颁布的《关于办理利用信息网络实施诽谤等刑事案件适用法律若干问题的解释》也规定对其中“造成公共秩序严重混乱的”行为以寻衅滋事罪定罪处罚。但是根据刑法规定，寻衅滋事罪只能适用于扰乱公共场所秩序的行为，不能适用于前述网上传谣行为。至于认为可以适用《刑法》第 291 条之一第二款的规定、故意传播虚假信息罪予以定罪处罚的观点，属于望文生义，该罪处罚的是编造、故意传播虚假的险情、疫情、灾情、警情信息或恐怖信息的行为，如果所传之谣不属于以上五类虚假信息，就不能按该条定罪处罚。

作为最严厉的法律，刑法的适用应当保持谦抑性，遵守罪刑法定原则，坚守刑事法治边界。刑法的目的是预防犯罪，要求在公正的基础上对犯罪人进行特殊预防，而公正则意味着刑法的适用应当严守罪刑法定原则。刑法相比于其他法律更为严厉，实际是以刑罚之恶去除恶之恶，因此，难免会导致消极后果。如给受刑人打上“罪犯”的标签，影响其刑满释放后回归社会，而短期自由刑则不利于教育改造等。刑法是社会治理的重要法律手段，但不是唯一手段，更不是万灵万应的手段，不当扩大刑法的适用不仅达不到预防犯罪的目的，反而会给社会造成严重的伤害，这方面国内外都有沉重的历史教训。因此，面对汹涌的“民意”，刑事立法与司法仍要保持谦抑的定力，明晰与行政法及其他部门法管辖范围的界限，不将应由行政法或民法规制的一般违法行为纳入处理范围。

（作者系同济大学上海国际知识产权学院教授、博导，中国犯罪学学会副会长）

