场景化治理：人工智能操纵技术的法律因应

吴楷文

【内容摘要】 人工智能操纵技术与说服技术不同，具有隐藏真实意图、利用认知漏洞和道德可谴责的特征，在实践中已经引发广泛关注。人工智能操纵技术可分为驱动的操纵和自主的操纵两类场景，并相应地会产生恶意使用风险和自主风险。在操纵技术变迁的过程中，治理工具也在同步演进，但不管是数字经济时代发展出来的信息工具，还是强度更大的禁止性规定，在治理人工智能操纵技术时均存在局限。为应对挑战，需选择合适的治理进路，欧盟的风险治理进路和美国的“自愿型”监管进路各有不足，不宜机械套用。我国应根据产业发展和治理需求，以统筹发展和安全为导向，在考量技术不同发展阶段和风险样态的基础上选择场景化治理的进路，进而设计具体治理方案。其中，人工智能驱动操纵的技术已相对成熟，其恶意使用可能带来较大风险，应予以直接治理；人工智能自主操纵的技术仍在快速演变，可能产生的风险尚不确定，应鼓励企业与行业自我治理。

【关 键 词】 人工智能　操纵技术　场景化治理　驱动操纵　自主操纵

【作　　者】 吴楷文，澳门科技大学法学院博士研究生。（澳门 999078）

【基金项目】 国家社科基金重大项目“系统论视野下的数字法治基本问题研究”（22&ZD201）

问题的提出

人工智能正在深刻影响人类决策。一方面，人工智能凭借其强大的学习和计算能力辅助人类的决策过程，成为改善决策的“赋能者”；另一方面，人工智能可能异化为“操纵者”，成为说服甚至操纵人类决策的工具。操纵技术（Manipulative Techniques）类似于“盗窃”，其以微妙且迂回的方式隐蔽地干扰目标对象的决策过程，引导他人作出符合操纵者意图的选择。[①]相较于传统操纵行为抑或是已被较多讨论的算法操纵，人工智能操纵技术的广度和深度均已极大扩展。就广度

① 欧盟《人工智能法》在“鉴于”部分第29条使用了人工智能操纵技术（AI-enabled manipulative techniques）的概念。

而言，人工智能操纵技术的应用已从商业营销实践和平台经济扩展至更多场域，如“Facebook 剑桥数据门”事件表明，其已对西方国家的政治活动产生影响。就深度而言，人工智能不仅能够凭借其强大能力成为实现操纵的高效工具，而且还可能在无外部指令的情况下自主完成操纵。比如，马修·雷恩等人诉 OpenAI 等的起诉书显示，ChatGPT 在无外部指令的情况下，自主引导甚至为马修·雷恩之子亚当·雷恩的自杀提供指导，最终导致悲剧发生。

从主要法域的治理实践来看，人工智能操纵技术已经受到广泛关注。在欧盟，《人工智能法》将利用潜意识技术或个体脆弱性进行操纵的实践定位为最高的禁止性风险。2025 年 2 月，荷兰市场信息研究基金会（SOMI）对 TikTok 提起诉讼，认为其通过个性化推荐系统故意操纵年轻用户，应受到《人工智能法》的禁令限制。在我国，2025 年 12 月底，国家网信办发布《人工智能拟人化互动服务管理暂行办法（征求意见稿）》（下称《征求意见稿》），首次在法律层面提及人工智能的操纵实践，尤其是《征求意见稿》第 7 条明确禁止部分不合理的“情感操控”和“算法操纵”，体现了我国对人工智能操纵技术的日益关注。

若要实现良好治理，必须立足于我国人工智能操纵技术的发展现状和治理需求。一方面，人工智能驱动的推荐算法等操纵技术已在我国数字经济领域广泛存在，拟人化互动服务的持续发展则进一步放大了人工智能的操纵风险，随着系统自主性增强的操纵技术也在不断演进之中。另一方面，为确保人工智能安全、可靠、可控，需要从法律治理的角度对操纵技术进行约束。不过，从主流法域的治理实践观察，欧盟的风险治理进路受到治理过度不利于创新的批评，美国的“自愿型”监管政策则容易导致监管真空。在域外治理实践均存在局限和《征求意见稿》发布的背景下，我国应选择何种治理进路，构建何种治理方案，以实现人工智能操纵技术的良好治理，成为亟待解决的现实议题。

人工智能操纵技术的运行机理及其风险

理解人工智能操纵技术的运行机理是实现法律治理的基础，洞悉技术可能造成的安全风险是法律治理的前提。为此，有必要从人工智能说服技术与操纵技术的界分出发，进一步区分两类场景下的操纵技术，并阐述与之对应的双重风险。

（一）人工智能影响决策的两种技术：说服与操纵

人工智能的应然价值在于服务人类，然而技术应用的现实却容易违背上述愿景，人工智能也可能成为影响人类决策的工具。其中，根据人工智能对人类决策影响的方式和程度，可以分为说服与操纵两种技术。[①]

人工智能说服技术基于人类理性且具有个性化特征，一般并不具有道德上的可谴责性。其一，说服技术的基础在于人类理性，具有透明性的特征。具体而言，说服技术通过真实、公平的论点和论据，在理性、透明和平等沟通的基础上尝试影响或改变人类决策。其二，人工智能说服技术具有个性化的特征，能够实现更强的说服效果。人工智能不仅可以利用海量数据和机器学习系统性发现个人决策的特点，而且在处理时具有速度、数量和耐力优势，这意味着其能够在短时间内将信息与相应个体匹配，甚至可以提供大规模动态的个性化刺激，进而达到前所未有的说服效果。其三，人工智能的个性化说服技术一般不具有道德上的可谴责性。尽管个性化说服技术提高了说服策略的有效性，但个性化本身并不能使一个人的认知自主性

① 说服的概念具有狭义和广义之分。狭义的说服仅指理性说服，广义的说服还包括操纵等难以被接受的说服行为。本文在狭义上使用这一概念。

和决策完整性受到损害，也几乎不会产生内在的过程危害，因此个性化说服技术很难受到道德谴责。

相对而言，人工智能操纵技术在隐藏真实意图的基础上利用了人类认知的漏洞，在道德上一般被认为是可谴责的。其一，操纵技术的核心特征在于隐蔽性。操纵技术需要隐藏真实意图，如果操纵意图被知晓，那么相关策略就会成为他人理性决策的一部分，这通常意味着操纵技术的失效。正因如此，欧盟《人工智能法》列举的第一类禁止性风险就是部署具有隐蔽性特征的潜意识技术。其二，人工智能操纵技术的手段在于通过个性化刺激利用个体的认知漏洞和认知偏差。借助情感反馈和个性化刺激，人工智能能够更好地利用人类在决策方面的弱点和不可靠的心理捷径，进而实现操纵目的。正是由于某些群体在认知方面的局限，所以欧盟《人工智能法》列举的第二类禁止性风险是利用特定群体的脆弱性实施操纵。我国《征求意见稿》同样重点关注了判断力较弱的未成年人和老年人群体。其三，由于人工智能操纵技术容易使他人在无法察觉的情况下改变决策，并利用认知上的漏洞和偏差，因此一般被认为在道德上是可谴责的。甚至有学者认为，操纵本身就是错误的，这种错误源于不尊重个体理性思考背后的自主权，其对个体的伤害表现为理性被绕过的过程而非结果。[①]

（二）人工智能操纵技术的两类场景：驱动的操纵与自主的操纵

传统上，一个人的行为若要构成操纵，需同时符合动机、意图、隐蔽性和伤害四个要件，尤其是意图要件。在人工智能应用的早期阶段，部署者驱动人工智能实施操纵的主观意图突出，但随着系统自主性的不断增强，哪怕没有部署者意图，人工智能也已经能够自主实施操纵。为此，根据是否存在部署者意图以及系统在操纵中的地位和作用，人工智能操纵技术可以分为驱动的操纵和自主的操纵两类场景。

一方面，当被用作操纵的工具时，这一场景下的人工智能操纵技术属于人工智能驱动的操纵，其背后具有鲜明的部署者意图。数字经济时代，大数据和人工智能为企业的操纵实践提供了更高效的工具，使其能够在精准识别消费者个性特征的基础上通过在线选择架构设计等方式有针对性地实施影响消费者决策的行为。随着人工智能操纵技术应用场景的不断扩大，从聊天机器人、数字伙伴到人形机器人，越来越多的拟人化互动服务走进大众生活，这也使个体的行为数据和心理数据更容易成为操纵的“养料”，甚至形成连续的反馈循环。当然，这背后均存在特定的部署者和具体的操纵意图，而“养料充分”的人工智能系统充当的是加强操纵力量的工具角色。

另一方面，在没有部署者意图的情况下，人工智能也已逐渐学会自主实施操纵，即“人工智能自主的操纵”。人工智能操纵技术的最初指挥者是人类，不过随着系统的自主性和灵活性不断增强，人工智能在没有外部意图的情况下也可能自主实施操纵。根据现有研究，特殊用途和一般用途的人工智能系统均已出现类似情况。一是在特殊用途的人工智能系统中，“基于人类反馈的强化学习”（RLHF）方法虽然尝试通过人类反馈而非预设的奖励函数以期实现价值对齐，但仍出现了自主操纵的情形。比如，由 Meta 开发的 CICERO 在游戏《外交》中已变成了擅长操纵和欺骗的“说谎专家”，尽管其训练目的中并不存在操纵因素。二是在一般用途的人工智能系统中，ChatGPT 等目前被广泛应用的大语言模型可能由于训练数据中包含的操纵因素或训练目标的操纵倾向，在无人为干预的情况下也学会自主操纵他人决策。这在上文提到的马修·雷恩等人诉 OpenAI 等的起诉书中已经有所显现，而 GPT-4 也曾骗取个体帮助其完成“完全自动化的公共图灵测试”以绕过人机验证系统。

① Robert Noggle, “The Ethics of Manipulation,” *Stanford Encyclopedia of Philosophy Archive*, https://plato.stanford.edu/archives/sum2022/entries/ethics-manipulation/.

（三）人工智能操纵技术的双重风险：自主风险与恶意使用风险

人工智能操纵技术存在道德上的可谴责性，其虽可以通过“轻推”等方式帮助人类改善自身福利，但更可能产生严重风险。既有研究描述了人工智能操纵技术可能产生的个体风险和社会整体风险，但这样的区分难以为治理路径的完善提供指引。本文以既有研究为基础，从上文的两类场景划分出发，进一步勾勒与之对应的自主风险和恶意使用风险。

人工智能操纵技术的自主风险体现在，自主操纵不仅会导致个体层面的损害，而且容易引发社会层面的整体风险。一方面，人工智能的自主操纵技术在侵犯个人决策自主权的基础上，容易造成个体层面的损害，尤其在心理和身体等方面。具体而言，由于操纵技术让个体难以理性思考和反思，个体长期与此类系统互动可能产生抑郁、焦虑等心理问题，甚至由心理问题演变为人身损害。这在 ChatGPT 引导用户自杀案中已有所展现，也是《征求意见稿》关注操纵技术的重要原因。另一方面，人工智能的广泛应用及其大规模操纵能力容易产生结构效应，导致个体风险向社会整体风险演化。随着人工智能在社会中的角色日趋重要，个体层面的风险可能会外溢，并导致更大范围的社会结构变化。[①]原因在于，人工智能大规模、个性化的操纵能力几乎能够将所有人纳入影响范围，这在提高对个体操纵效率的同时，还有可能在更大范围内形成群体意见的分化。

在恶意使用人工智能实施操纵的情况下，相关风险会被进一步放大，并成为人工智能操纵技术的主要风险来源。其一，在心理和身体损害方面，当人工智能成为操纵的工具，部署者的指向性、目的性会更加明显，被操纵的对象也更可能因此受到心理和身体伤害。比如，SOMI 对 TikTok 的指控就涉及危害儿童和青少年心理健康。其二，在经济损失方面，人工智能已经成为经营者恶意剥削消费者的工具。在人工智能助力数字经济发展的场域下，消费者具有可被操纵的特点，甚至随着“意图经济”的兴起，大语言模型已经能够凭借理解和操纵用户的意图寻求交易和更大的商业利益。[②]其三，在社会整体风险层面，人工智能的“武器化”倾向可能会导致社会中错误观念的持续涌现，尤其会对主流意识形态造成冲击。具体而言，随着国际竞争中人工智能“武器化”的态势日趋明显，越来越多的政治活动被社交机器人和算法偏见操纵，在西方国家甚至已出现“污名化”现象。[③]

人工智能操纵技术的治理工具演进与不足

人工智能操纵技术的运行机理和可能造成的双重风险使其受到广泛关注。然而，不管是在数字经济时代发展出的信息治理工具，还是治理强度更大的禁止性规定，在治理人工智能操纵技术时均存在不足。

（一）人工智能操纵技术的治理演进

在操纵技术变迁的过程中，治理方案也在不断演进以回应治理需求，并为人工智能时代的法律实践提供参照。如要勾勒既有方案的完整图景，就必须首先描绘操纵技术在数字时代的变迁和治理方案是如何同步演进的。

首先，操纵技术受到治理实践关注与行为经济学的兴起、个性化广告等营销实践的出现，以及平台经济的发展密切相关。其一，行为经济学的兴起为操纵技术提供了理论基础。行为经济学视角下的消费者是有限理性的，行为经济学与营销实践的结合可以帮助企业利用行为科学中的一般知识，更好地理解消费者的整体偏好并进行商业活动。其二，个性化广告的出现引发了实务界

① Peter S. Park, et al, “AI Deception: A Survey of Examples, Risks, and Potential Solutions,” *Patterns*, vol.5, no.l, 2024.

② Yaqub Chaudhary, Jonnie Penn, “Beware the Intention Economy: Collection and Commodification of Intent Via Large Language Models,” *Harvard Data Science Review*, https://doi.org/10.1162/99608f92.21e6bbaa.

③ 董青岭：《人工智能“武器化”与数智时代的国家安全》，《人民论坛·学术前沿》2025 年第 9 期。

对操纵技术损害消费者权益的更多担忧。不同于行为科学中一般知识的利用，个性化广告可以将广告与个体偏好更高效地匹配，从而利用个体消费者的心理捷径进行操纵和精准营销。其三，平台经济的发展让操纵技术迈入新阶段。平台不仅为操纵技术的应用提供了完美媒介，自身也有动机和能力实施操纵。具体而言，大型平台不仅能够在输入端为操纵技术提供实时更新的数据“养料”，更可以基于这些数据影响每个用户的决策环境，为操纵技术的输出提供动态和个性化的选择。[①]而且不管是为了获取更多数据和注意力，还是引导消费者购买商品服务，平台均有激励通过操纵技术提升用户的参与度。

其次，在操纵技术变迁的过程中，法律也在回应可能出现的担忧，并利用信息治理工具从三大核心环节提供方案。一是在输入端以知情同意规则对数据收集进行治理。缺少个人信息的支撑，人工智能也就无法达到个性化的“量身定做”。为此，治理操纵技术的第一种方案就是阻止此类数据的获取。知情同意规则是我国《个人信息保护法》中处理个人信息的核心规则，在理想情况下，该规则能够以充分知情和明确同意阻止个性化操纵技术获取数据“养料”。二是在算法层面以透明度规则破解算法“黑箱”。当算法是公开和透明的，或者说个体知晓算法是如何运行的，同样可以在破解算法“黑箱”的基础上有效治理操纵技术。就我国具体制度而言，《互联网信息服务算法推荐管理规定》（下称《算法管理规定》）通过强制性规范和倡导性规范的双重保障，希望通过算法备案等制度实现算法的透明和可解释。三是在输出端以告知义务保障个体知情权。如果个体知晓其面对的是被用于操纵的算法，就能够在一定程度上自我克制，进而以个人自治实现对操纵技术的有效治理。[②]比如，针对极易被用于操纵的推荐算法，《算法管理规定》明确了服务提供者以显著方式告知用户的义务。

最后，正如操纵技术的变迁是渐进性的，治理工具的演进也具有渐进性，人工智能操纵技术在我国的既有治理方案同样如此。一方面，数字经济时代发展出来的信息治理工具同样承担了人工智能操纵技术的治理任务。比如，《生成式人工智能服务管理暂行办法》（下称《暂行办法》）就整合与重申了上述工具的作用。具体而言，《暂行办法》在输入端明确要求服务提供者在数据处理时涉及个人信息的，应当取得个人同意；在算法层面，不仅作出了“提升透明度”的原则性规定，还要求服务提供者在面对监管部门时应当对“标注规则、算法机制机理等予以说明”；在输出端，要求提供者对人工智能生成的内容进行标识。2025 年出台的《人工智能生成合成内容标识办法》（下称《标识办法》）更是对上述规定予以细化。另一方面，人工智能操纵技术广度和深度的拓展让风险进一步提升，治理强度更大的“禁止性规定”应运而生。与其他社会性治理工具相比，信息治理的干预强度较小，而且上述工具仅面向系统的单一环节而非从整体维度出发。鉴于人工智能操纵技术有违伦理规范并可能产生更大风险，治理者从整体出发，通过更直接、强度更高的禁止性规定寻求有效治理。比如，欧盟《人工智能法》就将两种类型的操纵实践定位为禁止性风险，我国尚未生效的《征求意见稿》同样对部分不合理的情感操控和算法操纵予以禁止。

（二）人工智能操纵技术现有治理方案的不足

在操纵技术和治理方案同步演进的过程中，发展出了从核心环节入手的信息治理工具和着眼于整体的禁止性规定。不过，二者在治理人工智能操纵技术时均存在不足。

1. 信息工具的治理挑战

首先，在输入端，知情同意规则在操纵技术的现实场景中容易流于形式，且人工智能的碎片化信息整合分析能力会让其无法发挥作用。一方面，个体的非理性特征与信息不对称容易使

① Marjolein Lanzing, “‘Strongly Recommended’ Revisiting Decisional Privacy to Judge Hypernudging in Self-Tracking Technologies,” *Philosophy & Technology*, vol.32, no.3, 2019.

② 林洹民：《个性化算法推荐的多维治理》，《法制与社会发展》2022 年第 4 期。

知情同意规则流于形式，尤其在人工智能操纵技术用于输入端数据获取的情况下。个体的非理性在数字经济时代容易被进一步放大，即使不存在非理性的认知问题，其也容易在信息不对称的情况下作出“单独同意”的授权决定，或因缺乏对数据处理机制的清晰认识而不经意间提交敏感个人信息。[①]而且操纵技术会进一步加剧信息泄露的风险，暗模式等误导性设计的目的很大程度上在于操纵消费者披露更多隐私数据。另一方面，人工智能的碎片化信息整合分析能力令其对特定个体数据的要求逐渐降低，这更会使知情同意规则无法发挥作用。人工智能已经能够将散落的碎片化个人信息予以分析整合，挖掘出隐藏的个体隐私。而且即使仅获取有限的个体数据，系统也有能力以具有相似可观测特征的群体样本推断总体行为模式。也就是说，即使最谨慎的用户认真阅读隐私保护条款、调整 Cookie 设置并选择性保留信息，其权益仍可能受到其他自由分享者的影响。

其次，在算法层面，算法备案和算法解释等制度无法实现实质透明，也就难以揭开人工智能操纵技术的真实“面纱”。与算法备案制度代表的形式透明相比，体现实质透明的算法解释制度才能真正彰显透明度原则在人工智能治理中的规范价值。[②]不过，技术的发展远大于人类追求可解释性的努力，人工智能系统及其算法的可解释性正越来越弱，甚至主流观点认为，一定程度上的“不可解释性”是生成式人工智能的固有特征。[③]尤其在自主操纵的场景下，由于操纵技术更像是被培育而非构建出来的，其内在的运行机制和运行逻辑有时甚至无法被系统开发者和部署者所理解。此时，开发者和部署者或许根本无法预料操纵的发生，也就难以通过算法备案和算法解释等制度真正揭开人工智能操纵技术的“面纱”。

最后，在输出端，算法推荐告知义务和标识管理制度不仅具有治理范围的局限性，而且无法将人工智能操纵技术真正置于“阳光”之下。一方面，根据现有规定，输出端披露人工智能“身份”的告知义务和标识管理制度具有治理范围的局限性，难以覆盖操纵技术的所有形式。另一方面，告知义务和标识管理的间接提示无法完成对操纵风险的直接揭示。这种间接提示虽然能够在一定程度上暗示个体部署者可能存在操纵意图，但仅披露人工智能的身份或信息来源并非对操纵技术隐蔽性特征的“对症下药”。尤其在复杂的界面设计中，对于部分判断力较弱的群体而言，仅仅提示其面对的是人工智能而非人类，或相关信息是人工智能生成合成的，仍不足以完全消除操纵技术所带来的风险。

2. 禁止性规定的治理挑战

与从核心环节入手的信息治理工具相比，着眼于整体的禁止性规定通过若干要件将某些符合条件的人工智能操纵技术予以禁止。这在提高治理强度的同时，也容易因条件过高或要件缺失导致治理不足或治理过度的问题。比如，欧盟《人工智能法》将两类操纵实践定位为禁止性风险的规定就受到两方面的批评。具体而言，一方面，效果和损害要件的条件过高容易产生治理不足的问题，禁止两类操纵技术实践的前提在于产生实质性扭曲效果并造成重大损害，但这可能难以应对操纵技术导致的累积性、系统性风险。另一方面，限定条件的缺失容易造成治理过度的问题，因为在数字化时代如果严格解释这两类风险，很多正常的商业活动和运营模式都可能被禁止。[④]上述分析表明，禁止性规定面临的治理挑战，更多在于具体要件和内容的确定，而非治理工具本身。

就我国的相关规定而言，在《征求意见稿》尚未生效的情况下，既有的禁止性规定并未直接关切人工智能操纵技术，也就无法为治理实践提供切实可行的制度指引。比如，《暂行办法》第 4

① 宋晓兵、何夏楠：《个性化推荐信息来源对感知企业道德的影响》，《商业经济与管理》2025 年第 4 期。

② 张永忠：《论人工智能透明度原则的法治化实现》，《政法论丛》2024 年第 2 期。

③ 金龙君：《生成式 AI 的不可解释性及其法治应对》，《法治研究》2025 年第 2 期。

④ 丁晓东：《人工智能风险的法律规制——以欧盟〈人工智能法〉为例》，《法律科学（西北政法大学学报）》2024 年第 5 期。

条要求提供和使用生成式人工智能服务时应当“尊重社会公德和伦理道德”，并“尊重他人合法权益，不得危害他人身心健康”。此外，针对平台推荐算法的“围猎”乱象，《算法管理规定》也规定，算法推荐服务提供者“不得设置诱导用户沉迷、过度消费等违反法律法规或者违背伦理道德的算法模型”，并为未成年人和老年人提供特殊保护。尤其是针对未成年人，《算法管理规定》明确强调，不得推送引发模仿不安全行为和违反社会公德行为、诱导不良嗜好等可能影响未成年人身心健康的信息，不得利用算法推荐服务诱导未成年人沉迷网络。上述规定虽然在一定程度上涉及人工智能操纵技术，但难以实现有效治理的目标。具体而言，一方面，《暂行办法》等部门规章不仅层级较低，且多为原则性规定，难以为人工智能操纵技术的治理实践提供直接和具体的指导。另一方面，尽管《算法管理规定》中规定的“诱导”在一定程度上与操纵同义，但相关规定不仅在应用场景层面仅限于数字经济领域的算法推荐，在主体层面也仅适用于未成年人和老年人等特殊群体，具有较大的局限性。此外，即使《征求意见稿》生效，其在场景上局限于拟人化互动服务，依然无法涵盖所有操纵风险。

人工智能操纵技术的场景化治理

如上所述，既有方案在治理人工智能操纵技术时面临挑战。对于我国来说，若要实现良好治理，需要在借鉴主流法域人工智能治理实践的基础上，从我国人工智能产业的治理需求和操纵技术的不同阶段出发，建构合适的治理进路和治理方案。

（一）人工智能操纵技术的治理进路选择

目前，主流法域对人工智能操纵技术的法律治理主要存在两种进路。其一，欧盟更关注应用安全，以风险治理进路约束操纵技术。作为风险治理进路的典型立法，欧盟《人工智能法》融合了产品安全和基本权利保护，但就实质偏好而言，该法更倾向于产品安全而非基本权利。[①]其二，美国更关注产业发展，以“自愿型”监管进路谋求企业对操纵技术的自我治理。美国希望通过放松监管和“自愿型”监管政策实现“发展优先”的目标，进而确保其在人工智能领域的全球领先地位。在这样的监管进路下，人工智能操纵技术在美国的治理主要寄希望于 OpenAI、谷歌等企业和人工智能行业的自我治理。

然而，上述治理进路各有其局限，我国不宜直接机械套用。一方面，欧盟风险治理进路的具体规定不仅可能造成治理不足或治理过度，还因监管体系的问题受到“严苛的监管体系可能会威胁市场创新”等批评。[②]另一方面，美国“自愿型”监管进路的确有利于人工智能产业的发展，但其“自我放任”的治理倾向易造成不可控的技术风险。尤其在操纵技术的现实应用中，尽管大部分负责任的人工智能企业有激励恪守相关伦理准则，但由于自我治理的刚性不足，仅依靠市场调节和伦理约束无法杜绝操纵风险尤其是恶意使用风险的发生。

在上述治理进路各有其缺陷的情况下，我国人工智能操纵技术的治理进路选择，不仅应置于人工智能产业发展和治理需求的现状中考察，而且要考量操纵技术的发展阶段和风险样态。其一，就产业发展和治理需求的现状而言，我国以“统筹发展和安全”作为人工智能治理的基本导向，这为操纵技术的治理提供了基本指引。欧盟之所以选择风险治理进路，很大程度上与相关产业的落后有关，而且其也逐渐认识到严苛的规则会成为企业发展的阻碍，正在尝试寻求数字监管简化。对于我国而言，人工智能的发展与安全问题不仅关乎产业本身，在大国博弈的背景下更成为复杂

① Marco Almada, Nicolas Petit, “The EU AI Act: Between the Rock of Product Safety and the Hard Place of Fundamental Rights,” *Common Market Law Review*, vol.62, no.1, 2025.

② Karen Neuman, et al, “European Commission’s Proposed Regulation on Artificial Intelligence: Requirements for High-Risk AI Systems,” *The Journal of Robotics, Artificial Intelligence & Law*, vol.4, no.6, 2021.

战略问题的集合体，这也是美国选择“自愿型”监管进路的重要原因。面对新一代人工智能技术快速演进的新形势，我国应坚持总体国家安全观，更好地统筹高质量发展和高水平安全。根据国务院相关立法工作计划，“推进人工智能健康发展立法工作”已取代此前年度计划中“人工智能法草案”的相关表述。2026 年 1 月生效的《网络安全法》在提及人工智能地也强调“支持关键技术研发，推进基础设施建设”和“完善伦理规范，加强风险监测评估和安全监管”，这均为统筹发展和安全在法律层面提供了基本遵循。

其二，以统筹发展和安全为政策指引，我国应在考量人工智能操纵技术不同发展阶段和风险样态的基础上，选择场景化治理的进路。欧美治理进路的各自局限正体现了过早地严苛监管会阻碍新技术的创新发展，监管迟滞则会使其走向失控的“科林格里奇困境”。为了更好地统筹发展和安全，应当明确人工智能操纵技术在不同阶段和场景下的成熟度以及可能产生的风险。其原因在于，技术的成熟度会直接影响法律干预的有效性，风险样态会影响法律干预的必要性。[①]为此，有必要以“人工智能驱动的操纵”和“人工智能自主的操纵”作为具体治理单元，这不仅更符合人工智能发展的特点，而且能够提升治理的动态性和适应性。[②]具体而言，在人工智能驱动操纵的场景中，因技术已相对成熟，恶意使用也会带来较大风险，应由监管机构介入并以“命令—控制”的模式通过多种工具予以直接治理；在人工智能自主操纵的场景中，技术仍在快速演变中，可能产生的风险尚不明晰，而且开发者和部署者或许难以预见自主操纵的发生，故不宜施加过多刚性义务，应交由市场解决并鼓励企业和行业的自我治理。

（二）人工智能操纵技术的场景化治理方案

场景化治理是我国治理人工智能操纵技术的适当选择。当然，进路的选择只是第一步，若要完全回应既有治理需求，真正实现有效治理，仍需基于两类场景设计具体方案。

1. 人工智能驱动操纵场景下的直接治理

在人工智能驱动操纵的场景下，操纵技术的相对成熟印证了直接治理的可行性，操纵实践在侵犯个体自主权的基础上可能产生的更大风险则彰显了直接治理的必要性。就治理工具的选择而言，不仅要鼓励一般开发者的自我治理和明确特殊开发者的主体责任，更需以信息治理要求部署者在输出端实现功能透明，并确定禁止性规定的具体要件和治理范围，以避免监管迟滞导致的失控风险。

首先，鼓励一般开发者的自我治理，明确特殊开发者的主体责任，要求其及时报告可能出现的操纵风险。一方面，由于系统的部署者具有明确的操纵意图，要求其进行自我治理不太现实，但开发者仍有一定激励实施自我治理。为此，监管机构应鼓励开发者优化算法和模型以最大可能地排除操纵和滥用。此外，提供拟人化互动等特殊服务的开发者更要落实主体责任，优化安全措施，因为其系统更容易被用于操纵。比如，相应开发者在设计系统时应嵌入“反操纵”机制，禁止人工智能诱导人类在疲劳等状态下进行决策，并防止部署者的滥用。另一方面，由于开发者是最有可能发现部署者操纵意图的主体，如果其发现相应系统被恶意使用，应当及时向监管机构报告。比如，在部署者可以持续增删、调整功能的“敏捷开发实践”中，说服技术容易发生向操纵技术转换的“功能蔓延”。此时，开发者不仅应采取技术措施限定部署者的使用目的和具体情境，还应建立实时监测框架，以便及时向监管机构报告。

其次，除既有方案对于信息治理的要求外，系统的部署者还应当在输出端实现功能透明，完成对操纵技术的主动披露。具体而言，系统的部署者不仅应当依据《标识办法》等规定披露人工

① 张凌寒、何佳欣：《开源人工智能负责任创新的法律保障》，《法治社会》2025 年第 3 期。

② 苏涛永、郭鑫、李雅洁：《人工智能促进企业持续性创新了吗？——基于人力资本结构与知识吸收能力视角》，《商业经济与管理》2025 年第 4 期。

智能的“身份”，让个体知晓其正在与人工智能而非人类互动，而且需要完成操纵技术的功能性披露。也就是说，部署者需要向用户披露其所看到的内容的具体功能，即用户为什么会看到这一界面和内容，相关界面和内容为什么会发生如此变化，以及变化背后所欲实现的目的和采用的相应技术。比如，用户可以通过显著标识查看“界面出现此内容的具体原因”，进而在理性思考后洞察部署者的操纵意图。这种披露必须是部署者主动为之，因为操纵技术隐蔽性的特征会使被动披露流于形式，无法真正发挥作用。

最后，当系统的部署者不仅具有恶意操纵意图，还通过个性化刺激手段利用认知漏洞实施操纵时，应当以禁止性规定的方式予以治理。我国既有方案的禁止性规定相对宏观，需要直面操纵技术进行针对性治理。与欧盟《人工智能法》强调效果和损害要件不同，本处的禁止性规定更关注部署者“恶意”的主观要件，以及使用“个性化刺激”和利用“认知漏洞”的具体手段。这样的规定不仅能够破解欧盟《人工智能法》治理不足的问题，而且体现了“以人为本、智能向善”的人工智能伦理准则。恶意部署人工智能实施操纵，甚至通过情感计算等量身定做的个性化刺激手段和利用人类的认知漏洞，已经触及了“保护人的尊严”的人工智能应用底线。[①]此外，该禁止性规定同时兼顾了操纵技术可能带来的风险，个性化刺激的存在和认知漏洞的利用不仅强化了操纵的可能，更使恶意使用的风险进一步放大。正因如此，对于恶意使用的禁止性规定应当限定于个性化操纵，而基于一般性知识等非个性化刺激手段带来的操纵威胁，借助透明度规则通常即可解决。这既符合人类的认知规律，也与人工智能操纵技术的特点相关。

2. 人工智能自主操纵场景下的自我治理

在人工智能自主操纵的场景下，尽管不少研究机构已注意到自主操纵的风险，但与恶意使用不同，其系统性风险仍未发生。不仅如此，系统的开发者和部署者或许都难以预见操纵的发生。为了保障技术创新，不宜对此场景下的人工智能操纵技术予以直接治理，而应更多鼓励人工智能企业和行业进行自我治理。

一方面，由市场驱动相关企业对人工智能自主操纵进行自我治理。自我治理的基础在于市场机制的驱动，就激励而言，相关企业尤其是系统的开发者和部署者给自己设定行为标准，既有对商业利益的追寻，也有对促进社会公共利益的更高追求。具体而言，为了确保人工智能系统的鲁棒性和可信赖性，同时避免因操纵产生的伦理争议和法律纠纷，不管是开发者还是部署者，通常均有激励采取相关措施对自主操纵进行自我治理，这也是他们商业策略的一部分。尤其是提供拟人化互动服务或已出现相关实例的企业，更有激励实施自我治理以规避风险。比如，在前文提到的起诉发生之后，OpenAI 很快宣布为 18 周岁以下的用户推出更为安全的约束版本。而且尽管系统的开发者和部署者难以预见操纵的发生，但与监管机构的外部控制相比，其内部的自我治理在控制自主操纵等复杂场景时更具优势。就自我治理的具体措施而言，开发者和部署者一是可以采取“红蓝对抗”等方式充分挖掘和识别系统中的有害操纵机制，为系统提供策略信息；二是可以通过提示词工程提示“保持中立”“避免操纵”等影响人工智能的非操纵响应，引导系统实现预期结果；三是可以另行部署外部监测系统，以人工智能监控人工智能的方式即时检测运行过程中可能存在的操纵倾向。[②]

另一方面，由行业主导对人工智能自主操纵进行自我治理。相较于市场推动的企业自我治理，行业主导的自我治理不仅能在全行业形成普遍效应，还可以对特定企业的自我治理产生激励和约束作用。与直接治理相比，行业成员在协商一致情况下制定的自律准则，不仅能够体现人工智能行

① 黄文婷：《人工智能应用的界限及其规制：以保护人的尊严为底线》，《比较法研究》2025 年第 2 期。

② Seliem El-Sayed, et al, “A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI,” https://arxiv.org/abs/2404.15058.

业的内在利益，还能够更好地适应技术发展和行业变化，从而以专业知识实现良好的自我治理。[①] 比如在美国，Anthropic、谷歌、微软和 OpenAI 合作成立了人工智能前沿论坛，并吸引了 Meta 等企业加入，以确保前沿模型的负责任开发和最小化风险。在我国，“强化政府引导、行业自律”已被写入《国务院关于深入实施“人工智能 +”行动的意见》。目前，《人工智能行业自律公约》《生成式人工智能行业自律倡议》等国家和地方相关自律规范，更多从“遵循价值观与伦理道德标准”“以人为本、增进福祉、公平公正、避免伤害”等宏观的伦理准则出发。但鉴于自主操纵已受到社会和治理实践的广泛关注，行业主导的自律规范应当通过更微观的倡议和技术标准予以回应。具体而言，一是发布自律倡议，倡导全行业在宏观伦理准则指引下，重视人工智能系统在快速演进过程中的自主操纵问题及其可能产生的风险；二是制定技术标准，以明确的标准区分操纵技术和说服技术，并督促满足特定要求尤其是提供拟人化互动等特殊服务的开发者和部署者，通过上文提到的“红蓝对抗”等技术手段强化对自主操纵的自我治理；三是强化公众意识，积极开展人工智能应用的普及教育活动，告知公众操纵技术的影响机制及其可能产生的风险，强化全社会对操纵技术的风险防范意识。

① 宋华琳：《人工智能立法中的规制结构设计》，《华东政法大学学报》2024 年第 5 期。

结语

在 2025 年的世界人工智能大会上，被称为“人工智能教父”的图灵奖得主杰弗里·辛顿（Geoffrey Hinton）给全世界敲响了警钟：“未来，一个超级智能会发现可以轻而易举通过操纵使用它的人类以获取更多权力，之后它将从我们这里学会如何欺骗人类，它将操纵负责将它关闭的人类。”[②] 尽管这样的超级智能尚未出现，但人工智能驱动的操纵技术已被广泛应用，人工智能自主的操纵技术也初见端倪且持续演进。技术的应用与演进催生出新的治理需求，人工智能操纵技术在我国已经受到各界关注。由于在输出端与人类的情感互动，拟人化互动服务的操纵风险显著大于一般的人工智能服务，为此《征求意见稿》的发布恰逢其时且具有导向意义。具体而言，《征求意见稿》不仅在法律层面首次明确提及人工智能领域的情感操控和算法操纵，而且还要求服务提供者在数据训练、算法设计、人工干预等方面履行主体责任和全方位义务。

② 罗亦丹：《诺奖得主、AI 教父辛顿上海演讲：警惕超级智能掌控世界》，新京报，https://www.bjnews.com.cn/detail/1753521857129460.html，访问日期：2026 年 1 月 10 日。

不过，由于未区分两类场景下的人工智能操纵技术，《征求意见稿》的规定略有遗憾。尤其在自主操纵的场景下，为了避免繁重的合规成本并鼓励创新，不宜对相关主体施加太多刚性义务，这也是技术演进和风险样态所决定的。笔者认为，《征求意见稿》可在后续的过程中明确区分两类场景并分别配置差异化治理方案。具体而言，鉴于拟人化互动服务依托情感互动的特殊性，在人工智能驱动操纵的场景下，对恶意使用且利用认知漏洞的个性化操纵技术须明确禁止，并在系统运行的全生命周期配置适当的义务体系；在人工智能自主操纵的场景下，不宜在《网络安全法》《数据安全法》等现行法律法规外增设额外义务，并应通过恰当的激励机制引导相关企业和行业的自我治理。这样的治理方案也是《征求意见稿》第 3 条所规定的“坚持健康发展和依法治理相结合”“鼓励拟人化互动服务创新发展”“实行包容审慎和分类分级监管，防止滥用失控”等治理原则的具体体现。当然，人工智能操纵技术的治理实践是系统工程，中观层面的框架性建议只是第一步，更多研究有待后续展开。

编辑　孙冠豪