

人工智能时代中国未成年人网络保护制度的完善

徐 婧

(福州大学 法学院,福建 福州 350108)

摘要:中国未成年人网络保护规范体系体现为限制型保护模式,与长期形成的以刚性限制为核心的家长主义路径依赖存在一定的内在冲突。人工智能时代,未成年人网络保护模式转型的核心在于实现从限制型向调控型的转变,以最有利于未成年人原则矫正异化的国家亲权本位治理逻辑,从科层制治理转向多元主体协同治理,秉持未成年人的生存权与发展权均衡保障的全面权利观。通过实现监护人同意规则的体系构建,细化未成年人同意能力与行为能力的认定规则。明确多元网络平台权责统一的标准,在以风险为基础进行责任分级的同时,设定严格的比例原则约束。强调生成式人工智能面向未成年人的价值观对齐法定义务,促进数据利用与公共福祉的良性互动,从而实现人工智能专门规范与未成年人保护规范的制度弥合。

关键词:人工智能;未成年人;网络保护;平台责任

中图分类号:D922.183 **文献标志码:**A **文章编号:**2096-028X(2026)03-0093-10

随着以 ChatGPT、DeepSeek 等为代表的生成式人工智能技术的爆发式发展,人类社会正经历着自互联网诞生以来最为深刻的数字技术革命。截至 2023 年底,中国未成年网民规模已突破 1.96 亿,互联网普及率已达 97.3%。^① 生成式人工智能不仅提升了传统网络风险的烈度、扩大了风险的覆盖面,更重塑了风险生成与传导的底层逻辑,使未成年人由信息接收者转变为交互共创者。这种人工智能原生代的成长环境的形成,不仅意味着技术载体的更迭,更标志着未成年人网络行为模式、认知发展机制与保护范式的系统性变革。面对人工智能技术引发的范式变革,当保护制度的规制对象从静态信息流转向动态人机对话,既有的未成年人网络保护模式在理念、机制与效能上均面临严峻挑战。

一、对中国现行未成年人网络保护制度的审视

中国未成年人网络保护政策历经多个阶段迭代,核心治理目标从单一限制向多维保障演进,政策工具也从零散规制逐步走向体系化构建,呈现出与互联网技术发展高度协同的阶段特征。^② 未成年人网络保护规范体系的架构,既是对科学技术快速发展背景下未成年人网络行为特征变化的回应,也是数字时代国家亲权理念与家庭监护协同治理理念的具象化体现,并始终在技术环境与治理需求的动态平衡中寻求突破。

(一) 中国现行未成年人网络保护制度的体系架构

随着科技的进步与时代的发展,中国未成年人网络保护法律体系从问题导向型治理的萌芽期,到风险防控型治理的探索期,再到权益保障型治理的体系期,完成了从局部管控到全域治理、从行政主导到法律主导、从被动防御到主动保障的结构性升级(如表 1 所示)。在技术变革与治理需求的动态博弈中,中国未成年人网络保护法律体系不断调整价值理念与工具选择,^③治理导向渐趋权益保障型,已经逐步构建兼顾风险防控与发展需求的多维度的未成年人网络保护体系。

作者简介:徐婧,女,福州大学法学院副教授。

^① 参见刘言:《守护好 1.96 亿未成年人的网络童年》,载《中国青年报》2025 年 6 月 11 日,第 8 版。

^② 参见冯嫫、高慧军:《技术治理视阈下我国未成年人网络保护的政策分析——基于 2004—2024 年的国家政策文本》,载《中国青年研究》2025 年第 5 期,第 52 页。

^③ 参见袁秀挺、严驰:《人工智能激励型立法审思:以激励规制理论为视角》,载《中国海商法研究》2025 年第 4 期,第 37 页。

表 1 中国未成年人网络保护法律体系不同阶段概述

阶段	时间范围	核心政策文件	治理导向
萌芽期	2004 年至 2010 年	《中共中央、国务院关于进一步加强和改进未成年人思想道德建设的若干意见》(2004 年)、《网络游戏管理暂行办法》(2010 年)	问题导向型治理,聚焦网络色情、暴力等显性风险的初步规制,构建保护框架雏形
探索期	2011 年至 2019 年	《中华人民共和国未成年人保护法》(简称《未成年人保护法》)(2012 年修正)、《国家新闻出版署关于防止未成年人沉迷网络游戏的通知》(2019 年)	风险防控型治理,以网络游戏为核心规制场景,通过时长限制、身份核验等手段阻断未成年人网络沉迷的风险
体系期	2020 年至今	《未成年人保护法》(2020 年修订)、《未成年人网络保护条例》(2023 年)	权益保障型治理,覆盖全网络场景,兼顾风险防控与发展需求,构建多维度保护体系

时至今日,中国未成年人网络保护规范体系的不同层级之间各司其职、相互衔接,呈现金字塔式的层级结构。在根本法层面,《中华人民共和国宪法》(简称《宪法》)第 46 条确立的受教育权与第 47 条保障的文化活动自由,不仅为未成年人发展权与受保护权的数字延伸提供了根本依据,也奠定了国家亲权原则的宪法基础。在部门法层面,《未成年人保护法》设立网络保护专章,明确了网络产品和服务提供者的特殊保护义务,构建了以最有利于未成年人原则为核心的制度框架;《中华人民共和国个人信息保护法》(简称《个人信息保护法》)设立未成年人个人信息保护专条,强化特殊保护机制,确立了单独同意原则的法定要求。在行政法规层面,《未成年人网络保护条例》作为第一部专门性综合立法,系统构建了网络素养促进、网络信息内容规范、个人信息网络保护、网络沉迷防治四大制度单元,标志着中国未成年人网络保护法治建设进入体系化阶段。在部门规章与规范性文件层面,国家网信办联合多部委出台了一系列配套规范,形成了从内容分类到平台认定、从技术标准到建设指南的精细化规则体系。其中,《可能影响未成年人身心健康的网络信息分类办法》以八部门联合发文的形式,系统列举了可能影响未成年人身心健康的网络信息,针对算法推荐与生成式人工智能服务设置了具有约束力的规范条款。《未成年人用户数量巨大和对未成年人群体具有显著影响的网络平台服务提供者认定办法》进一步细化了大型平台的特别义务,建立了定期认定与专项启动相结合的动态认定机制。值得关注的是,2024 年国家网信办发布的《移动互联网未成年人模式建设指南》,虽属规范性文件而非严格意义上的部门规章,却在事实上发挥着准规制功能。其首次引入分龄推荐机制,试图在不同年龄节点构建差异化的内容供给标准。

从规范内涵来看,中国未成年人网络保护规范体系的治理导向从单一防控向多元保障转型。一是网络素养促进制度,通过学校教育、家庭引导、社会参与的协同机制,提升未成年人与家长的网络认知与风险防范能力;二是网络信息内容规范制度,围绕网络信息生产、传播全过程,明确涉及未成年人不良信息的界定标准与处置流程,强化对于人工智能生成内容的监管;^①三是个人信息网络保护制度,针对未成年人信息的敏感性,设置收集、存储、使用、传输的全流程严格规范;四是网络沉迷防治制度,从前期的网络游戏时长限制措施,升级为未成年人模式下的内容筛选、时长管控、消费限制等综合性措施。然而,立法进程的体系化并不意味着治理逻辑的与时俱进。深入分析政策文本可以发现,在“保护”与“素养”等前进性话语的背后,“限制”“管控”“不得”等具有刚性约束色彩的表述仍然占据显著比重。中国现行的未成年人网络保护体系体现为一种限制型保护模式,尽管制度设计已明确迈向涵盖发展权保障的多元体系,但长期形成的以风险阻断为核心的防控型治理逻辑仍呈现出强大的路径依赖特征。这种限制型保护模式并非国家亲权理念的必然逻辑展开,而是在特定技术条件下被窄化理解后的阶段性产物。当网络风险形态单一、未成年人数字素养普遍较低时,以刚性限制为主要手段的防控型治理具有一定的历史合理性。然而,这种特定历史条件下形成的路径依赖也带来了实践层面的问题,使得许多新制度在具体落地时,仍不自觉地沿用或依附于传统的限制性思路与管控性工具。

(二)生成式人工智能技术对传统未成年人网络保护范式的解构

长期形成的以刚性限制为核心的家长主义路径依赖仍深深嵌入未成年人网络保护的制度运行之中,而在人工智能技术迅速发展的时代,大模型技术的深度渗透,则使得这一内在张力趋于显性化。

① 参见李润生:《未成年人保护视阈下“不良信息”的法律规制》,载《法律科学》2024 年第 2 期,第 111-112 页。

其一,未成年人网络保护的规范对象从静态信息转变为动态生成信息。既有制度体系建立在内容可分类、风险可枚举的预设之上,无论是对危害未成年人身心健康信息的禁止性规定,还是未成年人模式的内容过滤机制,均指向对预设信息池的静态管控。然而,生成式人工智能的输出内容并非源于可索引的内容库,而是依托大语言模型概率预测生成的涌现性文本,具有高度的情境依赖性与不可复现性。传统的依赖关键词屏蔽、黑名单匹配与人工审核的事前防控机制,面对生成内容的动态流变会发生根本性的技术失效。更为关键的是,人工智能生成内容的拟人化特质,使其难以被简单地归入不良信息或有害内容的既有分类框架内,造成规范对象在可识别性上的模糊。

其二,未成年人网络保护的规范机制从外部行为限制转变为交互过程调控。传统保护模式以监护人同意与平台限制为两大支柱,前者通过监护人代为决策阻断风险接触,后者通过时长管控与功能禁用压缩行为空间,二者共同构成国家亲权逻辑下的二元保护结构。然而,生成式人工智能的深度交互特征呈现出高度的私密性与渐进性。此类交互体现为持续性的关系建构,监护人既无法实时介入情感反馈循环,也难以传统的“知情—同意”框架规制未成年人的情感依赖。^①由此,既有制度在外部行为与内部认知之间出现了调控盲区:未成年人模式的刚性限制虽可切断使用入口,却无法破解算法驱动与情感驱动的成瘾机制,更遑论对未成年人认知固化与价值观闭环进行深层干预。

其三,未成年人网络保护的规范标准从静态年龄分层转变为动态发展评估。现行法秩序将以年龄划定的同意能力门槛,作为未成年人网络行为能力的核心界分标准,呈现出全有或全无的刚性特征。此种静态分层预设了未成年人的认知能力随年龄变化的线性增长,却忽视了生成式人工智能通过反馈学习对个体认知的差异化塑造。人工智能大模型能够基于用户交互数据精准塑造特定未成年人的偏好结构与情感模式,使得同一年龄段的未成年人可能面临截然不同的风险暴露与发展机遇。在此语境下,以年龄为单一变量的标准化保护,既可能导致对低龄高素养者的过度限制,也可能造成对高龄低识别力者的保护不足,无法回应人工智能时代未成年人发展的差异化与动态性需求。

概言之,中国未成年人网络保护的限制型保护模式虽在一定程度上回应了移动互联网时代的场景分化,却未能充分预见生成式人工智能带来的交互范式变革。生成式人工智能的技术跃迁,不仅对传统的限制型保护模式构成挑战,也揭示了以刚性限制为核心的家长主义保护模式在人工智能时代所遭遇的现实危机。

二、中国现行未成年人网络保护规范体系的内在缺陷

宪法层面的国家亲权仅赋予国家兜底保护、补足家庭监护缺位的公法义务,本身具有正当性。然而,中国未成年人网络保护规范体系在长期实践中,过于放大国家管控的权责属性,强调对未成年人的保护而忽视其发展权益,并以刚性限制为最重要手段的运行逻辑,不可完全与宪法上国家义务履行的价值理念等同。换言之,中国未成年人网络保护规范体系的限制型底色,难以适应人工智能时代新的挑战,存在一定的内在缺陷。

(一) 同意能力与行为能力的年龄标准衔接缺失

《中华人民共和国民法典》(简称《民法典》)第19条和第20条以8周岁为界,确立了限制行为能力人与无行为能力人的二分结构,以意思能力为核心,承认8周岁以上的未成年人可独立实施与其年龄、智力相适应的民事法律行为。然而,《个人信息保护法》第31条将14周岁设定为未成年人个人信息处理的同意能力分界线,规定14周岁以下未成年人的个人信息处理原则上须取得监护人同意。此规定让人不禁反思:既然8周岁未成年人已具备有限的意思能力,为何在个人信息这一防御性权利领域,其自主决定权却被推迟至14周岁?从法理层面剖析,个人信息处理的同意在性质上属于防御性权利能力,其制度功能在于防范信息自主权益受侵害,而非从事复杂的财产交易行为。同意能力作为权利防卫能力,其认知门槛应低于一般民事行为能力,而非更高。^②《个人信息保护法》第31条设定的14周岁标准,实际上是以家长主义逻辑替代了《民法典》的意思能力逻辑,将未成年人在数字空间中的主体地位向后推迟,这可能会造成一定的司法适用冲突,

^① 参见高志宏:《大数据时代“知情—同意”机制的实践困境与制度优化》,载《法学评论》2023年第2期,第126页。

^② 参见李永军、张兰兰:《未成年人信息同意能力的双重功能及其法律实现》,载《南京社会科学》2022年第4期,第91页。

并引发对未成年人数字主体性的认知差异。

与此同时,《个人信息保护法》中14周岁的一刀切的身份限制,既无法适配未成年人多元化的网络行为需求,也与未成年人数字素养提升与创造性发展的需求相违背。在现有未成年人个人信息保护的监护人同意规则下,监护人无法准确理解数据收集的范围及用途,从而作出最有利于未成年人的同意与否的判断。其同意的意思表示往往退化为流于形式的程序性点击,难以体现最有利于未成年人的实质标准。^①此种形式化的同意机制,不仅未能真正实现信息风险防范,反而将监护责任直接转嫁,压缩了未成年人在数字实践中自主决策能力的成长空间。在人工智能深度嵌入网络生活的背景下,静态单一的年龄阈值难以适配算法环境下信息处理的即时性、场景化特征,也与民事法律体系所倡导的意思能力分层、个案适配的法理精神相冲突。^②

(二) 网络平台责任规范之间的归责困境

“清朗·2024年暑期未成年人网络环境整治”专项行动期间,国家网信办共清理拦截涉未成年人违法违规信息430万余条,处置账号13万余个,关闭下架网站平台2000余个。^③诚然,中国未成年人网络保护离不开网络平台的介入和监管,现有的法律体系也对网络平台应当承担的未成年人网络保护责任提出了一定要求。《未成年人保护法》第78条要求网络产品和服务提供者建立“便捷、合理、有效”的投诉和举报渠道;《未成年人网络保护条例》第20条进一步要求“未成年人用户数量巨大”的平台履行特别义务。然而,将上述规定放置到更宏大的数字法治规范体系之下进行横向对比,会发现脱离了对未成年人这一特殊主体的特别关注之后,前述规范与《中华人民共和国电子商务法》(简称《电子商务法》)第38条规定的平台“知道或应当知道”侵权而未采取必要措施时的连带责任及《中华人民共和国网络安全法》(简称《网络安全法》)第49条要求平台“加强对用户发布信息的管理”,在归责标准上存在一定的差异。前者以特殊保护义务为核心,倾向于结果导向的严格责任;后者则以注意义务为核心,保留了一定的过错判断空间。正因如此,当平台在未成年人保护场景中因算法推荐的隐性风险或人工智能生成的诱导性内容而引发损害时,现行法律规范未对究竟适用何种归责标准进行协调和细化,导致平台在合规预期实践中陷入抉择困境:全面限制则牺牲用户体验与发展权,精细干预则缺乏法律确定性,放任自流则面临追责风险。

传统互联网时代,未成年人网络保护的责任主体主要是网络产品和服务提供者,限制型保护模式通过强化平台义务,能够实现有效的责任传导与规制落地。但在人工智能时代,内容生产模式从“专业生产”“用户生产”转向“算法生产”。风险来源呈去中心化特征:人工智能开发者负责算法设计,平台负责人工智能应用部署,包括未成年人在内的用户负责触发内容生成,甚至人工智能自身通过深度学习实现自主生成。当出现有害内容侵害未成年人权益时,责任归属难以精准界定。^④其责任究竟归因于人工智能开发者的算法缺陷、平台的审核缺位、未成年人自身的使用不当,还是人工智能本身的自主行为,现有法律框架与限制型保护模式均无法给出明确答案,使得平台和开发者均缺乏足够的合规动力,监管部门也难以实现精准追责。^⑤

(三) 人工智能专门规范未系统性嵌入未成年人保护逻辑

中国现有的人工智能专门规范的制度设计,并未系统性嵌入未成年人保护逻辑。例如,《生成式人工智能服务管理暂行办法》第14条规定了提供者在违法内容出现后的应急处置、模型整改与报备义务,但对违法内容的判断仍以成年人视角为基准,未针对未成年人的认知脆弱性设定专门的过滤标准。例如,一条“如何高效逃课”的生成内容,可能侵蚀青春期未成年人的规则意识。《未成年人网络保护条例》第23条要求平台“以显著方式提示”生成内容的风险,但并未赋予平台针对未成年人用户的价值观对齐义务,即通过技术手段调整模型输出,使其符合社会主义核心价值观与未成年人身心发展规律。此外,现行规范对人工智能模型的训练数据中是否包含未成年人个人信息、如何对这些信息进行去标识化处理、未成年人在其数据被用于

① 参见丁晓东:《人工智能背景下个人信息保护制度的挑战与应对》,载《政治与法律》2026年第1期,第105页。

② 参见商建刚:《生成式人工智能风险治理元规则研究》,载《东方法学》2023年第3期,第6页。

③ 参见《“清朗·2024年暑期未成年人网络环境整治”专项行动曝光典型处置案例》,载国家互联网信息办公室网站2024年10月9日, https://www.cac.gov.cn/2024-10/09/c_1730163316422064.htm。

④ 参见谢潇、罗世杰:《人工智能大模型训练数据供给的激励性法律规制》,载《广东财经大学学报》2025年第6期,第112页。

⑤ See Bernd W. Wirtz, Jan C. Weyerer & Ines Kehl, *Governance of Artificial Intelligence: A Risk and Guideline-Based Integrative Framework*, Government Information Quarterly, Vol.39: 1, p.1 (2022).

模型训练后是否享有退出权或收益权等问题,均未作出细致的规定。

限制型保护模式针对网络游戏、短视频等特定场景设置未成年人模式,通过限制使用时长、筛选内容及管控消费等方式实现保护,本质上是一种围墙式的被动防御。但人工智能技术已渗透至未成年人网络生活的全场景,未成年人无须进入特定模式,即可随时随地接触人工智能的算法服务。在此过程中,人工智能可根据未成年人的认知水平与兴趣爱好,生成看似无害实则有害的内容进行不良引导。^①例如,针对青春期未成年人的叛逆心理,生成“反抗原生家庭、逃离应试教育”等引导性内容;甚至可生成极具迷惑性的虚假信息,如一些急功近利的学习方法、不科学的虚假健康知识,误导未成年人的行为。这些内容没有明确的关键词与违规标识,难以通过关键词过滤或者人工审核等方式识别,却能对未成年人的价值观念和行为习惯形成潜移默化的负面影响,其危害程度远超传统显性不良信息。^②时长管控只能限制未成年人使用网络的时间,却无法破解算法驱动与情感驱动的成瘾机制,未成年人网络沉迷问题呈现低龄化的发展趋势。^③

综上所述,当前中国未成年人网络保护体系仍以限制型保护为核心运行机理,即通过身份识别划定保护范围,通过权限设置阻断未成年人的风险接触。限制型保护模式在规范逻辑上体现为一种简化主义的规制思路:将连续发展的未成年人认知能力简化为年龄门槛,将多元主体参与的风险治理简化为平台归责,将流动生成的交互过程简化为静态内容过滤。这种规制思路在法律体系中起到统领性作用,也使得既有未成年人网络保护法律制度在人工智能带来的复杂性面前存在一定程度的失灵,而非仅仅是若干规则的滞后或缺失,既有制度亦非一个协调自洽的整体。^④随着生成式人工智能技术的深度渗透,未成年人的网络行为已从被动接收信息转向主动生成内容,从单一场景使用转向全生态沉浸体验,从人机工具交互转向人机情感共生。既有的限制型保护模式内在缺陷逐步显现,亟待在法理基础与制度设计层面进行反思与重构。

三、从限制到调控:最有利于未成年人原则对国家亲权内涵的规范重构

国家亲权理念是中国未成年人网络保护制度不容剥离的公法权源基础,宪法所承载的未成年人受教育权与文化发展自由,为国家介入数字场域履行兜底监护职责奠定根本法正当性,由此确立国家防范未成年人网络侵害的法定权责。然而,权源正当并不等同于运行逻辑具备合理性。传统互联网治理进程中,国家亲权理念在制度落地中被异化为以刚性管控为核心的限制型家长主义保护模式,将未成年人简化为被动接受保护的客体,侧重消极风险防御而忽视数字赋能与自主发展需求。生成式人工智能的普及与渗透,已从根本上改变了未成年人网络行为的底层逻辑。以限制为核心理念的保护模式,在应对具有高度交互性的数字环境时,其结构性缺陷日益凸显。在此困境下,可将最有利于未成年人原则作为审视现有规范体系的重要标准,对异化的国家干预权实施价值约束与法理矫正,推动传统限制型家长主义完成范式迭代。质言之,中国未成年人网络保护制度的完善绝非消解国家兜底监护的宪法权责,而应当以儿童最大利益原则为标尺重构国家监护权的行使方式,在保留国家亲权正当权源的前提下摒弃过于绝对化的治理逻辑,实现国家亲权内涵从管控型向调控型的规范重构。

(一)从限制型保护模式转向调控型保护模式的法理基础

国家亲权理论源于罗马法中“国家是未成年人的最终监护人”的理念,其核心内涵是当家庭监护出现缺位或不足时,国家有权介入并行使监护权,以保障未成年人的合法权益。在传统互联网时代,网络技术的普及度较低、风险形态较为单一,且未成年人数字素养普遍不高,国家作为终极监护人,通过制定刚性规则及施加统一限制,对未成年人的网络行为实施干预。传统限制型保护模式中所显化的未成年人模式的强制设置、时长限制等制度设计,均是国家亲权本位法理逻辑的直接体现。国家通过行政监管与法律规制划定刚性边界,将未成年人视为需要被全面保护的脆弱个体,替代未成年人与监护人作出选择。这种法理逻辑在应对传统显性网络风险时具有一定的合理性,但在人工智能时代,其内在缺陷日益凸显:一方面,管控型家长主义的

① 参见赵丽、马子煜:《与孩子聊天的AI人设是“出轨对象”?记者调查AI剧情聊天软件乱象》,载《法治日报》2024年11月23日,第4版。

② See Alicia Solow-Niederman, *Information Privacy and the Inference Economy*, Northwestern University Law Review, Vol.117:357, p.357(2022).

③ 参见林维、吴贻森:《网络保护未成年人模式:立法跃升、理念优化与困境突破》,载《吉林大学社会科学学报》2022年第5期,第8页。

④ 参见张凌寒:《生成式人工智能的法律定位与分层治理》,载《现代法学》2023年第4期,第127页。

过度扩张,挤压了未成年人的权利空间与自主发展空间,与尊重未成年人主体地位的要求相违背;另一方面,国家亲权主导下的刚性限制,难以适配人工智能时代风险形态的隐蔽化、个性化特征,国家作为单一治理主体,难以实现对网络风险的精准管控。

国家亲权理念下未成年人网络保护法律体系的具体内涵与实现方式,需随技术环境与社会条件的变化而不断调适。人工智能时代的未成年人网络保护模式转型,并非要完全抛弃国家亲权理念,而是以最有利于未成年人原则为价值坐标,重构国家亲权理念的内涵与实现方式:从替代儿童作决定转向提高儿童及其责任主体作出明智决定的能力,从直接屏蔽风险转向调控产生风险的环境与系统。换言之,国家亲权理念与最有利于未成年人原则并非相互排斥的对立项,而是分属不同层次的法理范畴。国家亲权理念解决的是谁负有保护责任的问题,最有利于未成年人原则解决的则是如何保护的问题。前者是制度安排,后者是价值尺度;前者是手段,后者是目的。

因此,人工智能时代的未成年人网络保护模式转型,需要在原有的国家亲权理念基础上,进一步加深对于最有利于未成年人原则的理解,并将二者进行充分结合。《儿童权利公约》第3条第1款明确规定:“关于儿童的一切行动,不论是由公私社会福利机构、法院、行政当局或立法机构执行,均应以儿童的最大利益为一种首要考虑。”《未成年人保护法》第4条也明确将最有利于未成年人原则确立为基本原则,这一原则作为未成年人保护的核心法理依据,也是调控型保护模式的价值内核——其并非一种单一而刚性的国家干预权,而是一项兼具复合性、动态性与权衡性的治理理念。^① 该原则内在地包含三个相互关联、逐层递进的核心意涵。

首先,最有利于未成年人原则体现为一项目的性原则,要求一切制度与行动以未成年人的福祉与发展为终极目的,摒弃将管控便利置于优先地位的惯性思维。在人工智能治理中,技术开发与应用的正当性不仅在于商业成功或效率提升,更在于是否促进而非损害未成年人的身心健康、人格尊严及受教育权与社会性发展。^② 评估一项人工智能产品或服务,必须前置性地进行未成年人影响评估,系统分析其对不同年龄、不同处境的未成年人可能产生的短期与长期效应。

其次,最有利于未成年人原则体现为一项权衡性原则,要求在冲突的权益与价值间进行审慎衡量并作出最优选择。人工智能时代,保护与发展、安全与自由、个体权益与公共利益之间的张力更为突出。最有利于未成年人原则要求,不能为追求绝对安全而牺牲必要的探索与成长空间,也不能为技术发展而容忍可预见的重大伤害。^③ 其要求引入比例原则的精细考量:干预措施是否为实现保护目的所必需? 是否存在对未成年人权利限制更小的替代方案? 干预收益是否显著大于其所造成的损害? 调控型模式所倡导的动态化措施,正是这一权衡性原则在实操层面的体现。

最后,最有利于未成年人原则体现为一项参与性原则,强调尊重未成年人不断发展的能力与主体意见,是对传统国家亲权本位将未成年人纯粹客体化的修正。根据《儿童权利公约》第12条,有主见能力的儿童有权对影响其本人的一切事项自由发表意见,并应获得与其年龄和成熟度相称的考量。在人工智能愈发强大的网络环境中,保护制度的设计与评估应建立有效的未成年人参与机制。例如,在未成年人模式的功能设置、教育类人工智能的内容生成规则制定、算法推荐偏好的调整机制中,应通过安全有效、符合伦理要求的方式,倾听并吸纳未成年用户群体的真实反馈,使其从被动的保护接受者,转变为保护体系建设的积极协作者。

(二) 从限制型保护模式转向调控型保护模式的基本要求

基于对人工智能时代未成年人网络保护模式转型下最有利于未成年人原则的内涵解读,可以进一步推出从限制型保护模式转向调控型保护模式的基本要求。调控型保护模式要求国家的角色从直接的管制者与替代决策者,转型为更高级的框架设定者、能力培育者与利益平衡的仲裁者,通过构建激励相容的法律与标准框架,督促并支持家庭、学校、平台履行其赋能与保护义务,并在多元利益发生冲突时,依据最有利于未成年人原则进行最终判断。^④ 在人工智能带来的高度复杂性面前,调控型保护模式正从通过简化规则来控制

^① 参见苑宁宁:《最有利于未成年人原则内涵的规范性阐释》,载《环球法律评论》2023年第1期,第147-153页。

^② 参见邓丽:《最有利于未成年人原则的实践基础与制度理性》,载《政治与法律》2024年第6期,第27页。

^③ 参见刘征峰:《以比例原则为核心的未成年人国家监护制度建构》,载《法律科学》2019年第2期,第119页。

^④ 参见刘晓春:《从限制到调控:人工智能背景下未成年人保护模式的转型与重塑》,载《财经法学》2025年第5期,第118页。

风险的传统家长主义,迈向一种更具反思性、包容性和适应性的智能型家长主义。其仍然承认国家作为未成年人最终监护人的宪法地位与终极责任,但在实现方式上从管控转向调控。换言之,国家亲权理念的正当性基础未变,但其操作逻辑发生了根本性转变,不是去直接屏蔽风险,而是去调控产生风险的环境与系统;不是替代儿童作所有决定,而是致力于提高儿童及其最亲近的责任主体作出明智决定的能力。^①

具体而言,人工智能时代未成年人网络保护的理念转型,需要充分考虑以下两个方面的内容。一方面,强调权利本位。将未成年人视为独立的权利主体,而非被动的保护对象,尊重其自主意志与参与权。调控型保护模式需要实现从消极保护到积极赋权的权利转型。传统限制型保护模式体现为一种消极保护的权利观,即通过限制未成年人的网络行为,避免其受到网络风险的侵害,核心是保障未成年人不受侵害。而在人工智能时代,最有利于未成年人原则要求实现权利保护的全面化,通过积极赋能,为未成年人的数字素养提升、创造性发展提供保障,实现消极保护与积极赋能的有机统一。另一方面,注重利益均衡。在风险防控与发展权保障之间实现动态平衡,既要防范网络风险对未成年人的侵害,也要保障未成年人数字素养提升与创造性发展。调控型保护模式强调治理模式的算法正义与数据正义:一是算法正义理论要求公平性与透明性。人工智能算法的设计与应用,应当遵循公平、公正与透明的原则,不得存在歧视性设计,也不得通过算法推荐或者内容生成等方式侵害未成年人的权益;二是数据正义理论要求安全性与合规性。未成年人个人信息的收集、使用、存储,应当遵循合法且必要的原则,不得过度收集或者滥用未成年人的个人信息,并且应当保障未成年人的个人信息权益与数据主权。

从限制到调控的模式转型,体现了人工智能时代背景下最有利于未成年人原则对国家亲权内涵的规范重构。其要求打破单向度保护思维,搭建安全保障与数字发展之间的平衡机制,既要守住国家作为未成年人最终监护人的宪法根基,又需要借助儿童最大利益标尺,克服长期以来国家亲权本位的管控异化所带来的弊端,为人工智能时代未成年人网络保护制度的现代化转型提供法理支撑。

四、人工智能时代未成年人网络保护模式转型的具体路径

最有利于未成年人原则的适用并非仅仅是一种价值宣言,而是实质性地提出了未成年人网络保护制度重构的规范性要求。这些要求能否从抽象原则转化为可操作的制度安排,取决于能否针对人工智能时代未成年人网络保护的规范困境给出技术适配性的解决方案,从而实现调控型保护模式的治理范式革新。

(一) 监护人同意规则的体系化

《个人信息保护法》第31条将14周岁设为监护人同意的刚性分界,其隐含的家长主义逻辑在于:14周岁以下的未成年人一律不具有独立的信息自决能力。这一预设不仅在法理上与《民法典》以意思能力为核心的行为能力体系有所冲突,更在人工智能的深度交互场景中暴露出执行困境。监护人既无法对动态和隐性的数据处理行为作出有意义的知情同意,未成年人也因僵化的年龄门槛而被剥夺与其实际认知能力相匹配的自主权。因此,依据最有利于未成年人原则中的参与性原则,应当尊重未成年人不断发展的能力与主体意见,构建以意思能力为核心、以年龄为推定依据的弹性化同意规则。

欧盟《通用数据保护条例》(General Data Protection Regulation,简称GDPR)第8条关于信息社会服务的儿童同意规则已形成成熟的规范范式,为中国未成年人网络保护法律体系中弹性化的同意能力构造提供了一定借鉴。GDPR第8条第1款规定,针对信息社会服务的个人信息处理,默认以16周岁为未成年人独立同意的年龄门槛,同时允许成员国通过国内立法将年龄下限降低至不低于13周岁,从而在13至16周岁之间形成可由各国自主设定的弹性区间。从条文文义看,GDPR第8条在立法思路与价值取向上,明显摒弃一刀切的刚性思维,转而认可未成年人意思能力的渐进性与差异性,并允许各国根据本土民事行为能力制度、数字发展水平与保护传统作出适配性安排。成员国在设定本国同意年龄门槛后,其法律体系通常允许在特定情形下对未成年人的实际理解能力进行考察,以确定其是否具备独立同意资格。

这种将同意年龄与儿童实际发展能力相挂钩的做法,实际上体现了一种更具弹性的立法逻辑。GDPR的规范构造为中国修正《个人信息保护法》第31条提供了参照。中国虽然不必照搬13至16周岁的具体数

^① 参见王广聪:《论最有利于未成年人原则的司法适用》,载《政治与法律》2022年第3期,第146页。

值,却可以借鉴其年龄推定与个案弹性调整的规范技术,将意思能力实质判断引入同意规则,从而实现与《民法典》民事行为能力体系的内在统合,真正契合最有利于未成年人原则所要求的主体性尊重与发展权保障。

一方面,确立推定加推翻的法律技术结构。《民法典》第19条确立的限制民事行为能力制度,以意思能力的连续发展为前提,允许8周岁以上的未成年人独立实施与其年龄、智力相适应的法律行为。《个人信息保护法》第31条将信息同意能力推迟至14周岁,实际上是以形式年龄标准取代了实质意思能力标准,由此形成财产理性早熟、人格理性迟滞的规范悖论。同意能力作为防御性权利能力,其认知门槛应低于一般民事行为能力。据此,应当以“14周岁加意思能力”的复合标准重构同意规则:以14周岁为原则性分界点,14周岁以上未成年人原则上具有独立的信息同意能力,无需监护人代为同意;8至14周岁未成年人原则上由监护人同意,但若平台能通过合理评估手段,证明该未成年人已具备充分理解数据处理风险的能力,则应承认其独立同意的效力。

另一方面,以场景化分层同意替代一次性概括同意,赋予未成年人及其监护人事后退出权与删除权。信息自决权的核心不仅在于事前的同意选择,更在于事后的控制能力。鉴于数据被用于训练大语言模型后在技术上完全删除极为困难,对此应当区分不同层级的义务:对于尚未投入训练的数据,平台应在合理期限内完成物理删除;对于已投入训练的数据,平台应采取技术手段尽量消除其影响,或至少停止在后续模型迭代中使用该数据;对于技术无法实现完全删除的情形,应赋予未成年人屏蔽权,即要求平台在输出中不再基于该数据作出个性化响应。与此同时,对于已同意但后续不愿继续共享的数据,尤其是已用于模型训练的历史对话数据,应允许未成年人随时撤回同意,阻断数据在模型中的持续影响,并给予其删除相关负面信息的权利。^①

(二)从单一平台责任到多元责任分配

最有利于未成年人原则中的权衡性原则要求,在多元主体之间进行审慎的责任分配,既不过度苛责单一主体以致抑制创新,也不放任任何一方以技术中立为借口逃避保护义务。现行规范在平台归责标准上存在一定的模糊地带:《未成年人保护法》《未成年人网络保护条例》以特殊保护义务为核心,倾向于结果导向的严格责任;《电子商务法》《网络安全法》以注意义务为核心,保留了过错判断空间。在人工智能时代,这一冲突因风险来源的去中心化而进一步凸显:人工智能开发者负责算法设计与训练数据筛选,平台负责应用部署与内容审核,用户负责触发内容生成,人工智能自身通过深度学习实现自主生成。当有害内容侵害未成年人权益时,责任究竟归属于算法缺陷、审核缺位、使用不当还是自主行为,现有法律框架无法给出明确答案。

欧盟《人工智能法》是全球首部全面监管人工智能的综合性法规,于2024年8月正式生效。该法以风险为基础建立了四级分类框架:不可接受风险、高风险、有限风险和最小风险。在未成年人保护问题上,该法第5条明确将利用因年龄而产生的弱势性的人工智能系统列为不可接受风险类别,予以绝对禁止。对于高风险人工智能系统,该法则规定了严格的全生命周期合规义务:在投放市场前,开发者必须对系统可能对健康、安全及基本权利造成的风险进行系统性评估,并建立符合性评估体系。除此之外,该法还确立了从基础模型开发到应用部署的链条式责任,即在设计前评估、运行中透明、投入后监测的三段式监管框架内,将保障人的基本权利与安全的理念贯穿于人工智能系统的全生命周期。这种框架设计,对中国人工智能领域未成年人保护制度中风险分类的精细化构造,具有一定借鉴意义。

一方面,应当构建以风险为基础的分级责任框架,区分基础模型开发者与应用部署者的不同义务层次。多元协同归责的首要任务在于打破平台概念的模糊性,在人工智能产业链中识别具有实质风险的关键节点,并配置与其控制能力相匹配的注意义务。^② 具体而言,应将责任主体区分为基础模型开发者与应用部署者两个层次:基础模型开发者作为技术源头,对其训练数据的质量及已知安全漏洞负有源头治理义务;应用部署者作为场景化服务的直接提供者,对其面向未成年人的具体交互场景负有实时注意义务。基础模型开发者应尽可能在模型层面消除系统性风险,应用部署者则针对具体场景中的残余风险进行动态管控。二者对

^① 参见王贞会、周梓睿:《未成年人数字权利的倡导及其实现》,载《中国青年社会科学》2024年第3期,第125页。

^② 参见张欣:《生成式人工智能的算法治理挑战与治理型监管》,载《现代法学》2023年第3期,第110页。

外就对未成年人的损害承担法定连带责任,即未成年人及其监护人可向任一主体主张赔偿,对内再按照过错比例分配最终责任,通过责任分配倒逼源头防控与过程合规。

另一方面,应当明确主动技术的法律授权与比例原则约束。人工智能时代主动技术的使用,如实时监控对话情感、分析语义倾向,构成对未成年人隐私的深度干预。因此,应在授权的同时设定严格的比例原则约束。一是目的正当。任何对未成年人网络行为的干预措施,都应当服务于直接且重大的保护目标,如防止性引诱、自杀自残等严重伤害,不得以保护之名行普遍监控或商业利用之实。二是最小必要。在多种可实现保护目标的技术方案中,平台应当选择对未成年人及其他用户权利限制最小且隐私影响最低者,优先采用本地化处理而非云端上传、匿名化分析而非个体识别,并在风险消除后及时停止干预。三是公开透明。平台使用主动技术或实施干预前,须以清晰易懂的方式向监护人及未成年人披露所采用的数据范围及用户权利行使方式。^①从另一个角度说,并非所有人工智能应用都需要接受同等强度的监管,而是应当根据其未成年人权益的潜在侵害程度配置差异化的合规义务;对极端高风险行为予以绝对禁止,对具有潜在系统性影响的人工智能系统施加强化的全生命周期合规义务,对一般性人工智能应用设置有限的透明度义务,从而实现风险预防与创新空间的平衡。

(三)人工智能专门规范与未成年人保护的规范适配

中国现行的未成年人网络保护规范体系,一方面要求防范违法内容的产生,但判断标准以成年人视角为基准,未针对未成年人的认知脆弱性设定专门的过滤标准;另一方面又要求平台显著提示风险,但未赋予其价值观对齐义务,未成年人作为数据贡献者的权利保障近乎阙如。最有利于未成年人原则中的目的性原则要求,一切制度设计应以未成年人的福祉与发展为终极目的。在生成式人工智能领域,这意味着不能仅满足于不违法,而应主动追求有益于未成年人。

值得关注的是,欧盟在推进未成年人网络保护的过程中,依据《数字服务法》拟定的保护未成年人指南草案,围绕年龄验证、内容过滤等强监管手段的正当性与技术可行性引发了一定的争议。^②这也给中国立法者传递出两个值得深思的信号。其一,保护未成年人免受网络风险的正当目的,并不能自动为任何干预手段提供充分正当性,不宜贸然对全体用户普遍施加强制性身份验证或统一年龄管控。其二,技术方案本身并不具有中立性或无害性,宣称隐私友好的设计可能隐藏着更严重的安全风险或执行漏洞。设计不当的技术工具不仅可能无法实现保护目的,反而会加剧数据安全风险、对匿名性侵蚀及制度体系的内在冲突。基于上述反思,中国在推进人工智能专门规范与未成年人保护规范的制度弥合时,应当把握以下两个方面的核心面向。

一方面,确立生成式人工智能面向未成年人的价值观对齐法定义务,实现从“不违法”到“有益”的标准升级。建立合法但有害内容的识别与干预机制,这一框架可借鉴英国《在线安全法》中关于合法但有害内容的治理逻辑。尚未违法并不足以豁免平台的保护义务,监管者需要进一步判断:某些内容即使不触犯刑法,是否仍可能因其性质而非传播事实对儿童造成心理或身体伤害。该法通过引入内容性质这一判断维度,突破了传统法律中违法性作为唯一规制门槛的局限,为应对人工智能生成内容中非显性价值渗透风险提供了规范依据。对于构建人工智能时代中国未成年人网络保护制度而言,平台应当承担比“不违法”更高的注意义务,将价值观对齐作为强制性合规义务。同时,针对人工智能生成内容隐性价值渗透、认知诱导、非显性不良引导等传统规制无法覆盖的风险,平台应对识别为未成年人的用户实行模型输出特殊校准。^③在提示词层级中嵌入正向价值引导,对极端情绪及非理性行为等敏感话题设置专门回应范式,构建未成年人专属安全回答模板库与风险拦截清单。

另一方面,创设未成年人数据贡献的回馈机制,实现数据利用与公共福祉的良性互动。应审慎保障未成年人数据贡献者的权利,构建回馈机制,而非简单赋权。尽管直接向未成年人支付数据对价不具有可操作

^① 参见刘权:《论互联网平台的主体责任》,载《华东政法大学学报》2022年第5期,第81页。

^② See Sophie Stalla-Bourdillon, *The Commission's Approach to Age Assurance: Do the DSA Guidelines on Protecting Minors Online Strike the Right Balance?*, DSA Observatory (31 July 2025), <https://dsa-observatory.eu/2025/07/31/do-the-dsa-guidelines-on-protecting-minors-online-strike-the-right-balance>.

^③ 参见汪全胜、宋琳璘:《我国未成年人网络安全风险及其防范措施的完善》,载《法学杂志》2021年第4期,第98页。

性,但利用未成年人数据获益的企业应承担相应公共责任。换言之,要求使用未成年人数据开展模型训练与商业化运营的企业和平台按照营收比例向国家未成年人网络保护基金缴纳专项费用,具有一定的合理性和可行性。与此同时,应当鼓励企业和平台开放优质教育资源与公益接口,参与未成年人保护技术研发;可将数据回馈履行情况纳入企业信用评价与社会责任认证,强调网络平台商事行为的公益属性,从而彰显最有利于未成年人原则的实质正义。

五、结语

随着生成式人工智能技术的爆发式发展,人类社会正经历着自互联网诞生以来最为深刻的数字技术革命。中国传统限制型未成年人网络保护模式在理念、机制与效能上的局限性日益凸显。从限制型保护向调控型保护的范式转型,强调打破单一主体治理格局,构建多元主体协同发力体系,推动技术伦理内嵌与生态治理。人工智能时代,未成年人网络保护模式的转型是一项系统性工程,这既是国家与社会的责任,也是人类文明在技术变革中守护自身未来的必然选择。唯有通过理念重构、治理革新与制度协同,方能构建既能有效防控风险、又能充分激发潜能的现代化保护体系,为未成年人在数字时代的健康成长筑牢根基,最终推动最有利于未成年人原则的体系化落地与动态化实现。

Improvement of China's Online Protection System for Minors in the Era of Artificial Intelligence

XU Jing

(Law School, Fuzhou University, Fuzhou 350108, China)

Abstract: China's regulatory system for the online protection of minors embodies a restrictive protection model, and the long-established path dependency on paternalism with rigid restrictions at its core presents certain inherent conflicts. In the era of artificial intelligence, the core of the transformation of the online protection model for minors lies in achieving a shift from a restrictive approach to a regulatory approach. This involves correcting the alienated governance logic centered on state parental authority in favor of the principle of maximizing the best interests of minors, transitioning from bureaucratic governance to collaborative governance involving multiple stakeholders, and adhering to a comprehensive rights perspective that balances the protection of minors' rights to survival and development. By establishing a system of guardian consent rules, this paper refines the identification rules for minors' consent capacity and behavioral capacity. This paper clarifies the standards for unified rights and responsibilities across multiple online platforms, conducts risk-based responsibility classification, and sets strict proportionality constraints. This paper emphasizes the legal obligation of generative artificial intelligence to align with minors' values, promoting positive interactions between data utilization and public welfare, thereby achieving institutional convergence between specialized artificial intelligence regulations and regulations on the protection of minors.

Key words: artificial intelligence; minors; online protection; platform liability

(责任编辑:盛 怡)