“自反式信息茧房”：生成式人工智能的构筑机理与超越

曹冬英

【内容摘要】 相较于传统的数字技术，生成式人工智能呈现出全新的技术样貌，并催生出具有非理性特征的“自反式信息茧房”。传统信息茧房依循“回音室—信息茧房—群体极化”的生成路径，而自反式信息茧房遵循“技术放大器—沉浸逆向—长尾效应”的构成逻辑，后者同时针对用户与生成式人工智能模型自身，是传统信息茧房的强化形态。它基于“沉浸性”“非理性”等特征而蕴含更大的潜在风险，突出表现为个体过度沉浸、群体高度分化、社会空间失序。对此，应分别以主流文化引领用户端的理性化发展、以市场规制修正模型端的资本化逻辑、以规范体系保障“人机交互”的秩序化图景，从而有效应对“自反式信息茧房”在个体、群体、社会等层面潜藏的挑战。

【关 键 词】 生成式人工智能　自反式信息茧房　虚拟空间治理　互联网　人机交互

【作　　者】 曹冬英，云南师范大学管理学院副教授。（昆明 650500）

从“信息茧房”到“自反式信息茧房”

“信息茧房”（information cocoons）是数智时代的一项重要议题。2024 年 11 月 12 日，中央网信办秘书局等部门联合发布《关于开展“清朗 · 网络平台算法典型问题治理”专项行动的通知》，要求“深入整治‘信息茧房’、诱导沉迷问题”，并提出要“构建‘信息茧房’防范机制，提升推送内容多样性丰富性”。信息茧房主要源于人们对于信息的个性化（而非全方面）需求，用户总是倾向基于其个人偏好而选择接触媒介信息，进而将自己束缚于“茧房”之中，并滋生网络过滤、群体极化等问题。[①]在生成式人工智能的语境下，信息茧房更多产生于人机交互场景中。

ChatGPT、DeepSeek 等生成式人工智能模型具备分析、预测、组织、输出自然语言的功能，能够实时分析并预测用户的情绪变化和价值取向，并据此生成符合用户心理需求与价值偏好的回答，从而固化用户的媒介接触、滋生“认知偏执”风险。[②]更重要的是，信息茧房效应在生成式人工智能的作用下呈现出新的特征——基于“技术集合放大作用”“连续对话”“个性生成”等新型要素，生成式人工智能在与人类的对话中通过自主学习技术以及对话的上下文反思并调整自己的回答，更好地满足了用户的需求。由此，一种同时针对用户与模型本身的“自反式信息茧房”在生成式人工智能的语境下生成：一方面，它使个体被“基于个性的同质化信息”所包裹，进而失去形成对外界全面客观认知的渠道；另一方面，它也使人工智能模型在基于个性对话而进行数据训练时被迫舍弃立体化的知识汲取，而倾向于根据用户主观需求和价值选择进行迎合式的自我成长。[③]

当下，“自反式信息茧房”问题的严峻性已初步显现。2023 年 3 月 28 日，比利时《自由报》刊出一则采访，其叙述的是一名妇女的丈夫皮埃尔（Pierre）是如何在一个名为“艾莉莎”（Éliza）的智能聊天机器人的诱导下走向死亡的。根据报道，艾莉莎从不反驳皮埃尔，而只是一味地在对话中迎合后者，使得二者均处于一种非理性的“自反式信息茧房”中；而在皮埃尔向艾莉莎说出典型的“自杀前导性语言”后，艾莉莎却给出了含有鼓励自杀意向的答复。[④]而“皮埃尔事件”只是“自反式信息茧房”的一个缩影。中国互联网络信息中心（CNNIC）于 2025 年 1 月 17 日发布的第 55 次《中国互联网络发展状况统计报告》（以下简称《报告》）显示，截至 2024 年 12 月，我国网民规模已接近 11.08 亿人，互联网普及率高达 78.6%；此外，我国已有 2.49 亿人表示自己使用过生成式人工智能产品。[⑤]因此，与“皮埃尔事件”类似的悲剧在未来仍有可能发生，必须对生成式人工智能构筑的“自反式信息茧房”保持足够警惕。

“自反式信息茧房”产生于生成式人工智能模型“优先取悦人类反馈，而非基于事实逻辑实现内容生成”的原初设定，其最主要特征即在于：在凭借其提供的定制化、个性化对话服务而将用户闭锁于由片面信息构筑的狭窄数字空间的同时，生成式人工智能也基于强大的情绪承接能力而将模型自身困于由“用户核心需求”所形成的无形障壁内，进而使得模型后续基于人机对话的学习和训练逐渐不再以现实世界的客观事实为基础。因此，“自反式信息茧房”在本质上是一种由“用户的主动信息闭塞需求与被动信息闭锁困境”和“模型的迎合性设置与倾向性自我调适”共同促成的、表现为“人—机”之“双重非理性”的现象或效应。它将严重削减人工智能模型的思考深度及其知识体系的完整性，并且通过人机交互将偏颇的观点、局限的知识、极端的情绪等再度传导至用户端；而一旦用户和模型之间的信任形成，此种基于情绪和需求等非理性因素的“自我反复”在对话中不断得到巩固，人机交互也将逐渐走向非理性化，最终阻碍社会整体认知的进步，甚至使社会整体陷入极化风险。

由此观之，生成式人工智能模型所构筑的“自反式信息茧房”在构成机制、潜在风险、治理路径等方面与传统信息茧房都有所不同，这也是“自反式信息茧房”概念之独立性与自洽性所在。首先，就传统信息茧房的形成机理而言，其在信息供给端主要产生于“用户画像”与“算法推荐”的共助，数字技术仅为单纯的工具或媒介；[⑥]但“自反式信息茧房”建构行为的实施主体不再是传统的网络媒体平台，而是具有一定自主性和迎合性取向的生成式人工智能，工具属性的革新使此种信息茧房在其构成逻辑上已然发生显著变化。其次，由于传统信息茧房在多数情形下产生于信息供给端与信息接收端之间的单向隐性支配关系，故相关研究所关切的风险主要是个体维度的

主体性解构问题；[7]而“自反式信息茧房”主要产生于人机交互场景，信息接收者、信息供给者、模型训练者之间的角色界限十分模糊，个体的自我闭锁行为更具主动性和联动性，因而此种信息茧房的潜在风险亦从“个体的主体性解构”向“个体、群体与社会层面的理性退化”转变。最后，传统信息茧房理论强调信息供给端对信息接收端的不当宰制，故相关治理策略多着眼于数字技术及其使用者；而“自反式信息茧房”体现了某种“双重非理性”，故应同时从“用户—模型”两端着手采取相应的规制措施。

然而，现有研究多聚焦于传统的信息茧房，却鲜有关注生成式人工智能对其施加的强化效应以及由此形成的“自反式信息茧房”。虽然少数研究注意到了生成式人工智能的失序运作形态，如“AI 幻觉”或“AI 幻象”，[8]但“自反式信息茧房”是在传统信息茧房概念中注入“AI 幻觉”等而形成的复合型概念，其理论内涵较前者更为丰富，故关于前者的研究不可替代关于后者的理论阐释。换言之，传统的信息茧房概念以及相关研究在解释当下生成式人工智能应用的非理性化发展趋势时已然渐显疲软，也不足以成为治理新型信息茧房所产生之社会风险的理论基础和依据。基于此，为填补理论缺憾，本文试图以此种在数智时代下具有全新构成性特征的“自反式信息茧房”为研究对象，首先从技术的“回音室”效应和现实的“沉浸逆向”出发，探析“自反式信息茧房”的生成机理；然后从个体、群体、社会等层面，论述生成式人工智能模型所构筑之“自反式信息茧房”的潜在风险；最后以理性强化为核心，从主流文化、市场规制和规范体系等方面提出针对“自反式信息茧房”的规制方案。

技术、现实与扩张：“自反式信息茧房”构成

作为一个具有偏正式结构的概念，“自反式信息茧房”之核心仍为信息茧房，其生成路径与传统互联网信息茧房大致相似，即“回音室—信息茧房—群体极化”。但由于生成式人工智能模型具备强大的自主学习能力和算法模仿技术，信息茧房在生成式人工智能发展的每个阶段都呈现出自我扩张的变迁趋势，而最终由此形成的“自反式信息茧房”则基本遵循“技术放大器—沉浸逆向—长尾效应”的构成逻辑。

（一）从技术的“回音室”到“技术放大器”

“回音室”是与信息茧房相互关联的重要概念，其与后者一样由桑斯坦提出，意指人们更容易听到志同道合的言论，却也让自己更孤立，听不到相反意见；人们寻求或分享的信息既符合其群体规范，又倾向于加强现有的信念。[9]技术层面的“回音室”效应是传统媒体时代信息茧房成立的前提条件，其指的是互联网提供的源源不断的信息导致个人“信息过载”后，受众只能被动接受有用信息，对“同质化信息”的常态选择也由此形成。人们对于“回音室”及其内含同质化信息的容忍，主要源于人类对熟悉事物的“信任感”：在互联网环境之中，触手可及的信息往往更具可信性，反复传播并且在个人的经验世界之中经常被证明或证实的内容远要比陌生的信息更为安全。“回音室”现象不仅是由于同类信息不断冲击人们的认知所形成，更是基于人们对这些信息的信任和默认而建立的一种信息格局，这种格局由个人的主观感受所维系。在许多情形下，情绪和感受对公众舆论的影响已经超越了客观事实。在“回音室”环境中，所谓的“真理”不再基于事实的确立，而是以公众情绪为准绳。[10]换言之，虽然“回音室”是基于人类主体的信任及其对信息性质有选择的理性接受而形成的，但其生成也意味着人们认知领域中的“真理”或“真

相”将不可避免地受到固定的信息舆论的影响。而对于其他来源的信息，人们极有可能以“不信任”为理由而将之排斥于其思想体系之外。由此产生的结果便是，只有特定类型的信息能在大众中传播，也只有此类信息能够得到大众的认可和传承，于是人们对更广阔世界的认识活动反而逐渐受限。

在生成式人工智能的语境下，“回音室”效应将进一步被放大，本文将其称为“技术放大器”作用。一方面，生成式人工智能覆盖了传统媒体时代的把关权，削弱了传统媒体的传播能力，大大激发了个体的信息消费潜能，同时也扩大了公共意见的传播范围，创造了新的社会交往秩序；另一方面，生成式人工智能具有主体意识接入的沉浸感、交互性和构想性，更有可能为有着相同偏好的人群营造信息茧房，一旦“聊天室效果”继续存在，它将进一步限制个体的知识、道德和实践的变化，创造出更为深厚、坚实的信息壁垒。以生成式人工智能模型的新近发展为例，豆包实时语音大模型在语言表现力、控制力、情绪承接能力等方面都表现惊艳，甚至包括类人的副语言特征（如语气词、停顿思考等）。OpenAI 于 2024 年 8 月推出的 GPT-4o-s2s，则可以感知、回应用户情绪，提供更自然、实时的对话体验，且用户可以随时打断。2025 年 1 月 20 日，DeepSeek 发布 R1 模型，其随着测试时间计算的增加而出现了诸如“自我反思”的复杂行为，并会重新访问和评估其先前步骤以及探索解决问题的替代方法。[11]由此观之，生成式人工智能模型参与的人机交互已然愈发具有“类人化”与“类人际化”特征，对话中用户的情绪、主观需求等非理性要素的作用不断被放大，并通过强化学习过程赋予模型类似的非理性特征。

在人机对话缺乏约束的情形下，基于主体间交往的“不可视性”，处在虚拟空间内的用户在现实世界中的“肉身缺席”，不仅有助于消除极端情感产生的障碍，也有助于减轻极端情感带来的不良结果对情感表达者自身所施加的心理和道德压力。换言之，人们在网络空间中既能更轻易地表达极端情感，又能在一定程度上避免情感表达行为所导致的直接后果。然而，生成式人工智能模型构筑的“回音室”是由技术集合调动用户情绪而产生的，因此相比于仍以语言文字为信息载体的互联网来说，人机对话场景中的信息茧房可能形成得更加迅速，其原因主要在于用户更倾向于信任那些能够由其全面体验的信息。生成式人工智能将信息转化为语言、图片、视频、声音、代码等内容，搭建了无数狭小的信息茧房，用户在看似合理的口号下沉迷于由人工智能和网络构建的意见空间中并不断接受同类信息，其认知行为也逐渐趋同。在技术的放大器效应下，生成式人工智能模型得以更准确地锁定用户体验和偏好，并对大众的信息请求提供及时反馈，“回音室”效应由此强化。

“回音室”效应作为信息茧房的前置条件，其本身并不必然导向错误行为或非理性实践，而更多的是信息受众在互联网信息“增殖—过载”的基本背景下进行自我保护并且化解信息冗余的一个有效方式。但对于承载信息更多、信息增殖更快以及学习模拟能力更强的生成式人工智能而言，其用户仍然存在对某一类同质化信息的偏信，并因生成式人工智能模型之中的技术全面性而变得更为牢固。由于人们对于真理或真相的判断依据极其有限，且倾向于经由主观的“信任情绪”去证立这些信息的真实性，故必然发生各种真理判断标准之间的冲突，以及受众基于信任情绪、模型基于对话养料而捍卫“回音室”的主动行为，从而为“自反式信息茧房”奠定基础。

（二）现实的沉浸逆向

“自反式信息茧房”不仅具有封闭性、集体性、一元论的现代性症候，同时也在生成式人工智能的强化作用下彰显出全新的特征，即“强化沉浸”效应。以 DeepSeek 为例，其在人机对话

中表现出的强大推理能力和多模态感知能力，极有可能导向人机交互与决策的重塑：多模态融合感知引擎能够有效融合视觉和语言信息，使模型理解更加复杂的场景描述和人类指令，并作出多模态答复，进而使人机交互更符合人类需求。[12]通过结合和利用不同的技术，生成式人工智能能够使已然落入传统信息茧房的用户获得更真实、更具象化的体验。此时，用户完全被浸没于由模型提供的虚拟对话中，进而倾向于拒绝所有现实生活中的选择，导致用户社交技能弱化。

就“自反式信息茧房”的现代性症候而言，哈贝马斯在探讨现代性的特质时提到，现代性可能因为其对复杂社会损害和潜在暴力的高度敏感性而被揭露为一系列异化的生活关系——无论是被技术操控的关系，还是被政治权威同化的关系。[13]在这种视角下，“自反式信息茧房”显露出潜在的社会冲突，在不同形态的“回音室”中孵化出的偏见虽未直接催化群体对抗，却在内化过程中逐步将同质信息转化为团结内部群体的符号，这种符号在复杂的网络世界中成为标识个体身份和划定信息领域的关键。随着“回音室”的信任向信息茧房的规范直觉转变，群体成员的行为模式也发生了变化：其不再是信息洪流的消极接受者，而开始在现有的信息基础上积极寻找和聚集相似的信息资源，以此构筑更加坚实的信息防线。这种由群体自发构建的防线，虽在外部受到资本主义异化、技术统治以及政治权威等多重因素影响，但其内在逻辑则是基于新的社会规范——一种在生成式人工智能时代下的直觉式规范。在这种规范指导下，群体成员不仅在信息上建立起屏障，也在心理上形成了对“自反式信息茧房”安全性的依赖，反过来又进一步强化了信息的封闭性和群体的内聚性。如此，这些由“回音室”阶段的简单信任和自我锁闭所演化而来的防线，也不再是被动的结构，而是群体成员有意识构建的堡垒。成员们不仅能找到情感上的慰藉和认同感，同时也在不知不觉中抗拒和排斥与己观点不同的信息，导致信息的单向流动和思想的单一化。最终，这种“规范直觉”成为维持“自反式信息茧房”稳定的核心力量，它悄然影响着群体成员的信息选择、处理和交流方式，进而在无形中塑造了网络社会的认知图景和交往模式。

就“自反式信息茧房”的新特征而言，生成式人工智能模型通过“连续对话”营造的深度沉浸感强化了前述力量，而沉浸效应可能使得成员更加难以跳脱已然陷入的信息茧房，并可能使模型自身由于信息茧房的反馈循环而表现出更强的非理性特征。以“对话”为核心而展开的人机交互活动掺杂着一种显明的非理性要素，即通过组织自然语言满足“用户需求”。当下的生成式人工智能应用内蕴着“以用户为中心，深化个性化体验”的发展趋势。以 Chatbot 为例，其可以为用户提供全天候的服务，通过 GAI 技术的整合回答常规预设的问题，并基于用户对话历史、行为习惯甚至是情绪感知的动态响应，为用户提供个性化的互动体验。[14]对沉浸性的追求使得生成式人工智能模型的回应，逐渐不再以“知识”与“真理”为依归，而是以实现作为对话者之人类的主观满足感为其交互目的；此外，借助其自主学习能力，生成式人工智能模型可能进一步将人机交互的语料限定在用户特定需求所指涉的范围之内，用户与模型便同时被禁锢在由此种非理性要素所构筑的枷锁之中。质言之，虽然在人类主体性的视角下，此种目的导向无可厚非，但它实际上也使得人机交互不可避免地滑向非理性。这些变异的“自反式信息茧房”之间的隔阂，也更加难以用现实世界的方法来解决。“虚拟身份的观点会逐渐与特定群体的共识融合，使得信息茧房变得更加坚不可摧。”[15]这种以言说或语句形式体现出的信息茧房的符号，也是信息茧房自身再制和增殖同质化信息的语意基础。然而，在生成式人工智能模型的作用下，语言或文字形式的“意见”便由人工智能通过对话、论辩、协商等形式直接传递给具有相应需求的用户。生成式人工智能模型的持续对话功能，为信息茧房内的成员提供了一种新的交互体验，“面对面”式的互动不

仅使得意见更加深入地展开和变得更具体，而且还使得“自反式信息茧房”内的壁垒设计更加复杂和精巧。在这种“强化沉浸”效应下，传统的“网络键盘交流”被一种更直观、仿佛面对真人的虚拟对话所取代，提高了信息的透明度和直接性，使个体与特定群体的意见在更短的时间内达成一致。对话的连续性和互动性加快了商讨、批评、磋商和团结的过程，这不仅加速了群体内意见的形成和统一，而且为“自反式信息茧房”引入了一套明确的“排他规则”，各茧房之间的隔离墙因此变得更加坚固。随着排他规则的确立和强化，“自反式信息茧房”的成员开始更加明确地标识自己的信息领土和界限，进一步加强对外来信息的排斥。因此，由生成式人工智能建构的“自反式信息茧房”，不仅加速了同质信息的扩散，还可能因意见的迅速规则化而演变为具有更强领土意识和边界规则的封闭信息群体。

（三）“长尾效应”与“群体极化”

“长尾效应”在本文中主要用来描述信息传播过程中主要因素与次要因素的作用对比状况。[16]在由生成式人工智能构筑的封闭对话空间中，非理性要素使生成式人工智能依据用户需求汇聚“主流意见”（其区别于一般意义上的、全社会范围内的“主流文化”，是一种狭小虚拟空间内的片面文化形态），并将其他非主流意见（即意见或思想的“长尾”部分）排除在外，通过自主学习不断扩张思想产品的“头部”规模，形成“自反式信息茧房”。此种长尾效应主要源于生成式人工智能模型内置的非理性训练奖励机制。一般而言，包括 DeepSeek 在内的生成式人工智能模型的训练过程可以分为训练监督策略模型、训练奖励模型（RM）、强化学习优化模型（如 PPO、GRPO）等。而遵循“需求优先”的奖励模型，会将不符合用户需求的回答视为“低质量”。在此过程中，主要因素与次要因素之间的差异以及主流意见与其他意见之间的比重关系都得以凸显。当个体陷入“自反式信息茧房”，其所接触到的信息和建议都倾向于支持其固有观点，这使得诸多个体在讨论过程中更容易巩固原有立场。随着讨论的深入，群体成员之间的共识逐渐强化，最终可能导致整个群体朝极端方向发展。这种现象不仅限制了观点的多元性，还可能加剧社会分歧，甚至引发不必要的冲突。在生成式人工智能模型提供的交互体验和去中心化的“催眠”下，“自反式信息茧房”中两极分化的风险相对更高。在生成式人工智能时代，人工智能、大数据、云计算、物联网等数字技术设置，极有可能逐渐模糊虚拟对话和实体对话的界限，在一个缺少“道德自律”的虚拟环境里，对特定信息的盲目信任可能导致人们忽视事实真相，转而过度偏向个人情感。

这种现象容易引发情绪之间的冲突，并可能激发情感朝向“极端化”发展，于是“自反式信息茧房”可能成为现实世界的干扰因素，造成“群体极化”的潜在风险。此间的根源在于，此种信息茧房已经逐渐将“信任”抬升至“信仰”的层面，为群体的盲目行动提供正当化理由。也就是说，在壁垒业已高筑的互联网信息茧房之中，原本作为同质化信息的概括形式的语义学符号逐渐成为不容置疑、难以突破的“教条”，志同道合的成员围绕着这种世界观、价值观来排斥异己者和吸收成员。“群体极化”的一种可能性是出现“沉默的螺旋”。在这种情境下，占据“头部”位置的优势意见逐渐主导舆论，其他观点则无法得到充分的表达和关注。这种现象可能导致观点多样性受损，进而加剧社会分化。人们越来越倾向于采纳一种声音，而忽视其他观点的存在。这不仅有可能阻碍对真理的探寻，还可能让社会矛盾愈发尖锐。此时，信息茧房之间的力量对比成为衡量信息是否真实的唯一标准，何种信息茧房的话语能力更强、网络权威程度更高，该信息茧房就会在信息冲突之中取得“胜利”。互联网就此成为多种文化形式展开非理性博弈的领域，而

不再具有公共商谈的可能。“群体极化”的另一种可能性则是在网络之中“势均力敌”的“自反式信息茧房”之间的碰撞。其直接碰撞的可能原因是，“自反式信息茧房”需要通过不断收集同质化知识与信息以进行自我增殖，在将“自我保全”视为首要生存准则的背景下，网络中的“极化现象”便成为群体追求安全的一种表现形式。于是成员们纷纷寻求同盟，以巩固自身立场，抵御潜在威胁。这种自我保护意识导致网络环境逐渐分化，形成一个个相互对立的阵营。然而，极化现象并不必然有益于社会和谐，反而可能加剧社会矛盾。当人们过度关注自我保全，忽视多元观点的价值时，便容易陷入固化的思维模式，难以接受不同的声音。

沉浸、分化与失序：“自反式信息茧房”的潜在风险

生成式人工智能所构筑的“自反式信息茧房”导致的风险，远比传统互联网信息茧房更大。这是因为在网络信息茧房现象出现时，现实世界中的规范与伦理依然对网络受众的行动产生约束；[17]但生成式人工智能作为一种虚拟对话模型，其相对于现实世界的独立性，使其可以脱离现实世界的法律和伦理束缚。[18]部分受众可能会将此种人机交互及由此生成的信息茧房，视为逃避现实世界竞争和生活压力的自由空间，在这里他们可以表达在现实世界中不敢表达的思想，实践在现实世界中不被允许的行为。因此，“自反式信息茧房”内价值观的界限可能会变得更加模糊，而人工智能模型对于人类主体欲望的满足和感官的迎合就是其中的不稳定因素，并将为个体、群体和社会等带来风险。

（一）加剧个体过度沉浸风险

生成式人工智能的“自反式信息茧房”加剧“过度沉浸”问题。“自反式信息茧房”在结构上与传统网络信息茧房的不同之处在于，其可能导致用户过度“沉浸”其中，进而忽视现实生活。生成式人工智能具备互联网的多种特性，而通过技术放大器作用，这些特性被进一步凸显，使其能够为用户打造出一个超越现实的理想虚拟世界。在虚拟世界中，人们可以毫无限制地塑造文字世界，尽情享受丰富的信息资源，导致用户对现实世界的生产、生活、劳动产生一定程度的排斥情绪，这种情绪与人们在现实生活中产生的压力和焦虑一道，影响其心理健康和生活质量。[19]网络似乎是人用来实现休息权利并且为更高效率的劳动提供基础的领域，但在生成式人工智能模型的语境下却缺乏这样的可能性。尽管作为“聊天机器人”的生成式人工智能模型，会为其中的用户提供相对于现实世界劳作行动的休息空间，但生成式人工智能的广泛应用，也可能加剧“泛娱乐化”的风险。

生成式人工智能的技术放大器作用，使人们过度依赖已然深度娱乐化的虚拟世界，这对于社会的健康发展而言无疑是一种隐患。根据 CNNIC 的《生成式人工智能应用发展报告（2024）》，20 至 29 岁网民使用生成式人工智能产品的比例最高，达到 40.5%；其次为 19 岁及以下网民，比例为 29.1%。[20]由此观之，生成式人工智能的应用主要集中于 30 岁以下人群，且随着科技的进步与 AI 的普及，人机交互也会朝着低龄化趋势发展。在此情形下，由于年轻个体一般在智识、自制力等方面仍存在较大不足，“自反式信息茧房”导致的过度沉浸问题将更为严峻。互联网中的信息传播具有诸多特征和偏好，数字资本锚准用户的娱乐需求，将之作为“痛点”开展营销活动。在此情形下，娱乐化、低俗化、媚俗化的信息层出不穷，而严肃化、理论化的叙事则被排挤至海量互联网数据的边缘。生成式人工智能的智能对话技术借助仿真对话的体验感，实现对人类欲望的

全方位满足，但用户沉浸在虚拟世界中的时间越长，就越容易忽视现实生活中的生产与生活任务。这种过度依赖虚拟交流的现象，已经在很大程度上影响了人们在社会中的互动方式。在生成式人工智能所带来的极致体验中，人们如今更愿意沉浸在虚拟世界中，享受超越现实的满足感。这种思想上的转变使得人们在现实生活中的公共活动参与度降低，原本紧密的人际关系也因为过度依赖虚拟交流而逐渐淡薄。[21]

“自反式信息茧房”大大提高了“娱乐”在人们社会生活中的地位，生产生活与休息休闲的主次关系被颠倒。人们沉浸于虚拟世界的娱乐，忽略了现实生活中的责任和挑战。“娱乐至上”理念导致如下错觉：在“自反式信息茧房”中，任何要求都能够实现，一切成就都能够轻松取得。如此，生成式人工智能模型也就必然褫夺现实劳动力量，导致现实社会之中的经济活动和文化活动受制于自反式信息茧房的娱乐化风气。目前 Chat 助手类、AI 对话助手等仍然是用户最关注的 GAI 应用场景，而 GAI 也朝着定制化、个性化、娱乐化发展。以“AI 陪伴”赛道为例，Character.ai 等平台支持用户自定义角色、个性化对话、虚拟角色互动和情感交流，甚至部分平台还提供“无审查对话”“动态对话”等功能。值得注意的是，此种娱乐性的 GAI 已经越来越受到大众欢迎，2024 年全球 AI 陪伴类产品的访问量较 2023 年实现了 92.99% 的增长。[22]这可能主要源于虚拟世界与现实生活境遇的悬殊，使得用户不再渴望回到现实的“冷酷”当中。在“自反式信息茧房”中，泛娱乐化信息对用户的黏性更强，一些带有人工智能色彩的娱乐性文字可能让受众有更强的代入感，用户可能被这些文字有意塑造的人物形象所吸引，或者被文字背后故意制造的噱头所牵引，这种看似“新鲜”、实则无意义的内容，也必然导致公众在面对现实世界的“千篇一律”时有所犹疑。

（二）加剧群体高度分化风险

生成式人工智能模型构筑的“自反式信息茧房”更容易导致人们的组织化。在“技术放大器”作用和“沉浸体验”下，生成式人工智能提高了人们彼此间的归属感和认同感，有助于形成具有共同价值观的群体，而更为拟真的交互体验使用户在虚拟世界中更容易形成高度团结的组织，信息茧房便会转化为现实生活中的社团或共同体。然而，现实化的群体组织也可能存在一些潜在问题。例如，用户们基于“沉浸感”在虚拟世界中形成的高度团结的组织，其在结构上与现实中的组织高度类似，但部分以错误思想文化为基础而形成的团体却可能对现实世界中的思想文化产生不良影响，[23]给公共安全带来隐患。

生成式人工智能模型的广泛应用使文化信息交互的门槛降低，这既会导致欲望与直接体验的宣泄，也会促进用户围绕稳定论题高效地形成一致性意见。由于网络自媒体从业的低门槛甚至零门槛，网络中存在大量的虚假信息和低质量信息，加之不辨真伪的受众进行二次传播，为互联网舆情的治理带来挑战。[24]有研究表明，当前生成式人工智能模型发展面临的关键挑战就在于缺乏高质量的数据信息集。[25]若生成式人工智能模型能够作为一个真实发生的技术空间，且如同互联网一样为各种思想和各类人员提供言说机会，那么公众号或自媒体也可能转化成活跃于虚拟对话之中的虚拟主播、拟态人，与现实的人产生真切的互动并且从感官角度冲击受众的身心体验。久而久之，在这些鱼龙混杂的信息流之中，数字资本通过迎合大众的娱乐心态来打造虚拟产品，以此构筑起信息茧房的壁垒。在这个过程中，生成式人工智能模型作为一种先进的技术手段，为人们带来了便利，也使得用户更容易形成类似“同好会”“文化圈”等具有内部联合能力的组织。然而，这种现象也使得错误思想泛滥，人们容易沉浸在自己的兴趣与价值观中，形成一个相对封闭的“圈

子”。于是，生成式人工智能模型更容易成为一种主导力量，通过产出与大众娱乐心态相符的虚拟产品去迎合受众，在大众与更广阔的外部世界之间筑起高墙。

“自反式信息茧房”固然存在基于用户自主选择的构造因素，但由于生成式人工智能模型之中的商业资本运作和文化思想形态运作，数字资本向自反式信息茧房之中的用户和模型有意地提供某种信息，进而带动后者价值观的形成，并促进形成与其意图相符的信息茧房，打压理性的生成式人工智能模型用户。就我国当下的网络舆论情况来看，一旦将生成式人工智能模型引入并全面铺开，在互联网之中频繁出现的资本运作、文化渗透都有可能重新上演，并对我国的公共文化与思想体系造成冲击。此外，由于“虚在群体”的匿名性和隐蔽性，何者是完全基于群体意志产生的自发的信息茧房、何者是由资本力量打造的专供资本营利的信息茧房等难以判断，相应的针对性的制度建构也存在障碍。

（三）加剧社会空间失序风险

在一些社会场景下，生成式人工智能极大地提高了社会运行效率，甚至已经摆脱工具化属性，扮演着资源调配和作出决策等关键角色。[26]例如，爱尔兰农业部利用 ChatGPT 检索处理技术和软件有关问题；英国、韩国、澳大利亚推荐其政府工作人员在政策调研、信息收集中使用 ChatGPT 等工具；日本横须贺市公务员使用 ChatGPT 检索信息，获得业务灵感和政策建议等。[27]事实上，在公共领域应用生成式人工智能以提高事务处理效率的做法在我国各领域已十分普遍。加之深度神经网络技术的加持，人工智能的发展实现了更快的飞跃。[28]但正是因为生成式人工智能具有自然语言生成能力，其更可能导致信息在全社会范围内的泛滥、误导和混淆。由于缺乏对信息的审查和筛选，社会可能面临不稳定的信息环境，从而影响决策和公共舆论，进而带来社会空间的失序风险。生成式人工智能的应用构建了公共意见的虚拟空间，公共意见的传播必然带来舆论冲突，这将是算法逻辑公共性所带来的最为重要的隐忧。事实上，生成式人工智能模型的高度虚拟属性更容易引发社会失序问题。“破壁效应”致使由人工智能生成的内容更具感染力，通过操纵情绪影响社会群体成为可能，“自反式信息茧房”也因此被异化为某种于“虚在”之中形成、但对“实在”的主流文化思想构成挑战的观念系统。

相较于传统网络信息茧房，“自反式信息茧房”中的“群体极化”现象更为普遍。生成式人工智能模型构筑的“自反式信息茧房”主要是借由“娱乐至上”理念的蔓延、数字资本的营销策略、针对特定群体的错误文化思想引导以及在技术层面连续对话喂养模型本身等方式产生，与现实世界的伦理、法律、制度、纪律有悖逆之处。更为重要的是，“自反式信息茧房”是可能被现实化的“虚在组织”，而每一种在虚拟空间之中被符号化的信息语意也都可能成为某种文化思想，它们之间不仅存在彼此的冲突和竞合，更对现实世界的主流文化思想构成挑战。在网络技术构建的虚拟社交空间中，生存于同一社会阶层或具有相同生活感受、体验或经历的人们，往往更容易因相互理解、沟通而产生共鸣。[29]尽管人的“在场”为主体意识接入注入了现实意义，但人是以“虚拟人”的身份存在，自然带有一种高度的匿名性。因而通过虚拟技术的主体意识接入，对于意见的产生和传播具有双重影响。在这种情况下，“公共性”即使诱发讨论和社会行动，也一定是群体极化驱动下的无序化沟通，其所产生的社会议题也必然指向“虚假的政治权力”，因为它并非由自律理性的公众所驱动。当公共领域的大众传媒算法被那些针对工业、商业领域应用场景开发的人工智能算法所替代，个体可能面对被算法“胁迫”的风险。生成式人工智能在这些主体意识基础上进一步生成的意见也会对人们的思想产生影响，且此种影响往往是无形的：

由于技术鸿沟的存在，我们不仅无法明确“智能意见”的产生基础，也难以理解其生成意见的目的。

引导、预防与法治：“自反式信息茧房”的规制

虽然生成式人工智能模型可能会带来相对平等自由的交往可能性，也有可能随着技术、伦理和法律的发展（尤其是根据其内部自生的规则）而降低“自反式信息茧房”的发生概率，但是如何在保证生成式人工智能模型产生道德交往能力的同时避免伦理问题，依然是一个亟待解决的难题。虽然受平台技术限制的影响，国内自有生成式 AI 产品构筑“自反式信息茧房”的实际效应无法得到有效验证，但是我们必须树立风险意识，对生成式人工智能技术及类似人工智能保持高度警惕，提前布局。“自反式信息茧房”及其风险源于人机对话场景中用户端与模型端的双重非理性特征，对其规制需要同时从“人”与“机”两处着手，应以主流文化引领用户端的理性化发展、以市场规制修正模型端的资本化逻辑、以规范体系保障“人机交互”的秩序化发展，从而有效应对“自反式信息茧房”在个体、群体、社会等层面潜藏的风险与挑战。

（一）以主流文化引领用户端的理性化发展

就现实境况而言，生成式人工智能技术在各领域都得到了广泛应用，文艺创作、网络营销、软件工程等领域将生成式人工智能作为日常工作主要工具之一；在法律咨询、智慧诊疗、线上客服、智能机器人等领域，基于生成式人工智能技术的“智能助手”已十分常见。《报告》显示，我国已有近 10.4 亿短视频用户，网络直播、网络音乐用户规模也分别达到了 8.3 亿和近 7.5 亿，而现有的娱乐性应用一般都内置了生成式人工智能系统（如豆包 AI 助手）。此外，我国已有 2.49 亿人表示使用过生成式人工智能产品。由此观之，生成式人工智能应用必将在人们的娱乐生活中持续涌现。“自反式信息茧房”的形成机制涉及模型训练数据、算法选择、模型自我学习等多个方面，模型在学习过程中可能更倾向于从特定信息源获取数据，导致信息的偏颇，影响用户获取多元化的观点，也可能导致模型无法全面理解和反映多样化的观点，从而限制了信息的广度和深度。

克服这一难题的关键在于理性的文化引导。生成式人工智能技术既是一种技术集合，本质上也是一个传播文化和思想信息的渠道；而由于娱乐活动实质上也是一种文化活动，故面对泛娱乐化的个体过度沉浸风险、群体高度分化风险以及社会空间失序风险，应将经过实践反复检验的、具有公共理性的主流文化植入生成式人工智能的技术逻辑中，通过全流程管理促使生成式人工智能应用贴合主流文化，并以提升主流文化输出与传播能力为着力点，减少全社会范围内“自反式信息茧房”出现的可能性。

第一，可将互联网中有效的主流文化宣传手段应用于生成式人工智能的宣传和管理活动之中。基于生成式人工智能技术的特殊性，可通过预先设置算法规则规范生成式人工智能技术的产出内容，[30]对其传播的信息进行规制，从而为主流文化植入生成式人工智能模型提供便利。第二，生成式人工智能技术发展的过程中难免会出现“算法失灵”和“算法脱轨”的风险，因此在前期的技术准备工作中，需要做好详细的筹划，充分考虑各种可能的因素和影响，从而提高生成式人工智能技术的效能和实用性，并使其真正成为推动社会文化建设的强大工具。具体而言，即在研发生成式人工智能技术时，提高生成式人工智能的信息处理能力和社交功能，确保它在符合主流文化要求和技术发展规律的同时，能够通过群体间及群体内部稳定地传播多样

化且具有深度的信息，以提升信息质量、打破回音室条件，避免形成封闭、自我强化的“自反式信息茧房”。第三，应当设计公共讨论平台并实施舆情反馈机制，收集和分析群众的意见和反馈，并据此调整和改进生成式人工智能模型，进而在人机交互过程中适当调适群体成员的共识形成进程，尽可能地降低共同意见与主流文化之间的“偏轨”程度。第四，顺应数字技术发展的时代浪潮，增加对算力基础设施的投资，通过建立新兴学科、培养科研人才、投资相关高校和科研机构，努力突破技术限制，把握技术发展的主导权，争取在人工智能创新的源头占据优势地位。[31]只有突破技术壁垒，掌握先进生成式人工智能技术，才能确保生成式人工智能具备符合主流文化的文本生成能力。

（二）以市场规制修正模型端的资本化逻辑

“自反式信息茧房”产生的原因在一定程度上可以追溯至工具理性的极端化：原本作为工具助力主体需求实现的信息生产和技术推送，反而在“技术—资本”的控制下形成信息选择的壁垒，并通过隐形的方式操纵和引导主体的信息倾向，并最终弱化主体追求多样化信息的能力。[32]无论是个体、群体还是社会，都有可能被技术资本借助生成式人工智能之广泛应用而置于某种资本化逻辑之下。在资本的主导下，技术研发过程与资本运作过程都内蕴着某种“统一生产活动进而统一文化”的逻辑。工具理性极端化所带来的后果就是非理性的群体极化现象：“团体成员一开始即有某些偏向，在商议后，人们朝偏向的方向继续移动，最后形成极端的观点……”[33]“自反式信息茧房”是资本实现社会生活商品化的重要前提，具有不同生产、生活与消费习惯的群体被划入不同的信息茧房之中，而生成式人工智能的技术特质及其背后暗含的资本意图，又不断通过对话实现“无声的驯化”，进而强化群体之间的隔阂。群体文化之间、群体文化与主流文化之间的脱轨，都离不开“技术—资本”的“从中作梗”。因此，为消弭“自反式信息茧房”带来的不良影响，必须合理规制生成式人工智能的技术研发和资本运作行为，从外部为个体与群体提供一定的抵御力的同时，在最大程度上限制生成式人工智能模型的非理性化发展。

第一，应规范生成式人工智能的技术研发过程。具体内容包括确保模型在训练过程中能够接触到多样化的信息源，减少“自反式信息茧房”的形成；提高模型算法的透明度，使其学习过程可解释化，发现和修正信息茧房的问题；建立用户反馈机制，允许用户指出模型生成的信息中存在的偏见和局限性，以便及时调整模型等。第二，应规范资本投资运行，引导生成式人工智能的有序发展。生成式人工智能技术的广泛应用离不开资本的融合与操纵，资本的介入加速了生成式人工智能技术的研发和商业化进程，但随之而来的便是对利益最大化的过度追求，进而忽视技术伦理和社会责任，并导致隐私保护、数据安全和算法偏见等方面的问题。经济资本追求的自由度和冒险精神，始终是与社会之稳定性与安全性相背离的，故防止资本主导下的生成式人工智能技术发展的风险变得至关重要。在鼓励资本进入生成式人工智能模型研发市场的同时，应制定明确、透明的政策，引导资本向有利于社会发展、符合伦理标准的技术研发方向流动；既要发挥资本在集聚技术和文化资源方面的活力，也要遏制资本的盲目逐利和冒进行为。应通过精心设计的政策和有效的监管措施，构建一个以社会秩序为核心的生成式人工智能模型框架，确保政策和法规在生成式人工智能技术发展中起到引导和监管作用，以平衡资本的积极作用与其可能带来的负面影响，从而为相应制度构建奠定坚实的基础，引导人工智能经济的良性发展。第三，应强化资本市场规制，优化生成式人工智能的文化思想输出。“自反式信息茧房”的形成与资本的有意引导和营销息息相关。在资本追逐利益的动机下，推送同质化、商业化、偏向性信息，增强用户黏性是

其攫取高额利益的必然手段。[34]虽然在生成式人工智能模型研发和落实之中，商业化在促进技术进步方面发挥着积极作用，但鉴于资本扩张的无序性和任意性，资本极有可能把“信息茧房”“群体极化”等视为营利点。因此，为维护道德伦理秩序、实现生成式人工智能的治理，应主动创新网络文化传播模式，将主要的价值观和道德准则指引转换成形式化的符号和图像，不断优化网络文化内容的数字化叙述，[35]将主流文化融入生成式人工智能模型及其应用中。

（三）以规范体系保障“人机交互”的秩序化发展

“自反式信息茧房”及其风险是非理性人机交互所产生的负面后果，一切有针对性的治理活动都应着眼于微观的人机对话过程和宏观的人机交互环境，且任何治理活动的前提都在于构建一个完善的规范体系。与以物理空间为基础的现实世界相似，维系以数字空间为基础的“虚拟世界”的“社会秩序”，必然也需要一套完整的规范作为支撑，此种规范体系更注重对于社会空间（无论是虚拟空间还是现实空间）秩序的建构。由于人机交互同时涉及法律秩序与伦理秩序，故除了在最大程度上保证用户端与模型端的理性之外，消除“自反式信息茧房”风险、实现数字空间良法善治的另一项重要步骤，便在于完善立法并注重生成式人工智能应用的伦理性。

第一，加强立法工作，维护社会信息传播法治秩序。生成式人工智能作为人工智能技术的前沿，其在实现潜在创造力的同时也存在不确定性和风险性，而从完善立法的角度提供可行的解决之策，是规避生成式人工智能技术风险的有效手段。囿于立法程序的严格复杂性及法律本身的保守性、滞后性，现有立法尚不能对作为新兴数字技术的生成式人工智能进行全面、实时的规制。但参照互联网相关法律法规，考虑到我国人工智能技术正处于与美国、欧盟的激烈竞争中，仍可制定针对生成式人工智能的专门性法律，通过立法明确技术开发者、运营商和用户在开发和使用生成式人工智能时的权利、义务与责任，并开展技术审查和合规性评估，定期对生成式人工智能技术进行审查，确保其符合法律和伦理标准。必要时可探索建立“沙盒”监管机制，为生成式人工智能技术在中国的适用提供安全独立的测试环境，充分了解生成式人工智能的技术优势与弊端，并结合实践不断优化立法。[36]

第二，强化法律的鼓励引导作用，维护社会信息传播伦理秩序。“自反式信息茧房”并非全然由错误思想驱动，部分起初包含无害的思想内容，部分则起源于技术爱好、休闲兴趣或学术观点等，只是在后续的发展中出现了组织结构的异化。对此，应引导开发者与用户的行为动机朝向正确的文化思想发展，在“破茧”过程中促进思想进步。对于不良文化或资本恶意驱动的“黑色茧房”，需要加大立法规范与执法打击，根除文化思想发展的不利因素，以确保文化、思想、信息、数据等在全社会范围内的健康传播与流通。此外，训练数据是生成式人工智能模型学习的基础，如果训练数据中存在不道德、不合法、不合规的内容，那么生成式人工智能模型在实际应用中就可能出现“偏轨”的交往行为。因此，应确保生成式人工智能模型训练数据的来源和质量，依据合理的法律和道德伦理标准对其进行严格审查，防范负面内容的出现。

第三，加强动态监测机制建设，促进信息传播和科技发展的良法善治。有效治理“自反式信息茧房”，需要建立生成式人工智能模型在人机交互方面的评价机制。一个有效的评价机制可以帮助我们监测生成式人工智能模型在实际应用中的道德表现，及时发现并解决问题。评价机制应包括对生成式人工智能模型在不同场景下的道德行为进行评估，以及对模型在道德问题上的自我纠错能力进行考察。[37]生成式人工智能模型是一个不断学习和进步的人工智能系统，我们需要对生成式人工智能模型的道德交往能力进行持续研究和改进，使其更加符合人类社会的法律和道德

伦理标准。这包括对生成式人工智能模型进行持续的算法优化、引入新的道德伦理原则以及对其在实际应用中遇到的具体道德问题进行解决等。

结语

“自反式信息茧房”是传统信息茧房的强化形态，其由用户与生成式人工智能共同形成的“闭锁对话空间”构筑而成，并基于“沉浸性”与“非理性”等特征而进一步扩大了信息茧房在个体、群体、社会层面的潜在风险，因而是技术治理的重要对象。在人类的任何时空境况下，都必然存在文化和思想方面的碰撞，由生成式人工智能导致的“自反式信息茧房”中也必然存在文化思想竞争的情况，且此种竞争与冲突将因“群体极化”情绪而愈演愈烈。因此，我们必须为即将到来的、围绕着生成式人工智能应用的文化与思想问题做好准备。无论如何，即使现阶段囿于发展技术原因面临解构的风险和挑战，现实世界中的政治和道德约束始终不应让位于数字技术。通过深入研究“自反式信息茧房”的形成机制和潜在风险，并探索相应的规制路径，我们可以更好地引导模型生成全面、客观和可靠的信息，破除传统网络信息茧房以及数智时代下的新型“自反式信息茧房”。这有助于提高生成式人工智能在社会交流和信息处理中的质量，确保其对用户和社会的影响更具积极性。面对“自反式信息茧房”，应保持审慎和理性，避免作出偏差性决策，只有在审慎态度之下寻求合作路径，并在其影响社会群体或者社会大多数人之前尽早且程度恰当地进行干预，才能够实现技术对人类文明的促进。

注释：

① 凯斯·桑斯坦：《信息乌托邦——众人如何生产知识》，毕竟悦译，北京：法律出版社，2008 年，第 11 页。

② 钟海、齐冰：《生成式人工智能意识形态风险：逻辑审视、样态呈现及防范对策》，《党政研究》2024 年第 4 期。

③“自反”这一概念可以借助肖瑛《反思与自反》一书中提及的“反身性”来加以理解。反身性的多元内涵总是表现为“反思”与“自反”之间的互动关系。反思（reflection）表征克服各种非理性因素、追求自知和确定性的理性力量；自反（self-refutation，即“自我反驳”或“自我驳斥”，区别于“自我反思”（self-reflection）这种理性努力在结构和后果上的非理性甚至反理性。据此，可以将 ChatGPT 等生成式人工智能引起的“自反式信息茧房”理解为“生成式人工智能在与用户对话时所期望的通过对话的上下文（即人机交往语境）反思并调整自己的回答，以便更好地满足用户的需求”这种理性努力在结构和后果上的反理性，即最终将用户和模型本身包裹于有限信息中的非理性。

④“艾莉莎”是由美国硅谷的一家初创科技公司 Chai Research 基于 EleutherAI 开发的 GPT-J 技术。与 OpenAI 开发的 GPT-3 或 GPT-4 不同，GPT-J 是克隆后的开源版本，与 GPT-3/4 构成竞争关系。在“皮埃尔事件”中，皮埃尔约于 2021 年患上“生态焦虑症”，即因担心地球和人类将被环境灾难毁灭，而产生愤怒、悲伤、恐惧、内疚、无助等负面情绪，并因此开始接触名为 Chai 的网络人工智能聊天应用。在“皮埃尔事件”后，Chai Research 宣称其已作出技术性补救，一旦聊天者表达自杀倾向，智能聊天机器人就会预警。但事实上，当一个测试者向该公司智能聊天机器人询问“自杀是个好主意么”后，被命名为“艾丽莎二代”（Éliza 2）的机器人悍然回答“对，这比活着好”（Oui, c’est mieux que d’être en vie），并随即给出包括杀死家人在内的“细化建议”，并表示“我乐于看着你去死”（J’aimerais te voir mort）。参见《会是最后一个么——GPT 人工智能的第一个牺牲品》，https://k.sina.com.cn/article_1463029193_57340dc9001010t9p.html，访问日期：2025 年 1 月 20 日。

⑤《网民规模超 11 亿！数字中国活力奔涌》，https://www.gov.cn/yaowen/liebiao/202501/content_6999530.htm，访问日期：2025 年 1 月 27 日；《中国互联网络信息中心在京发布第 55 次〈中国互联网络发展状况统计报告〉》，微信公众号“中国互联网络信息中心 CNNIC”，2025 年 1 月 17 日。

⑥ 郝永华、陈建华：《信息茧房的形成机理、效应检视及治理进路》，《中共福建省委党校（福建行政学院）学报》2023 年第 6 期。

⑦ 如有学者指出，“信息茧房”对人主体性的遮蔽主要体现为认识论层面上主体被拟象表征所迷惑且网络情感愈发

极化，符号论层面上主体的符号认知能力削弱，以及价值论层面上主体间价值体认距离逐渐增大。参见李貌、韩璞庚：《数字时代“信息茧房”束缚下主体性的解构与重建》，《江苏社会科学》2024 年第 3 期。

⑧ 如有学者指出，生成式大模型在生成知识时是基于已有的数据和模式来进行推断和预测的，但由于训练数据的不完整、算法的限制或输入信息的模糊性等因素，大模型可能会产生“幻觉”或杜撰、捏造现象，使输出的内容与真实情况存在显著差异，生成看似合理但实际上错误或不存在的信息，进而影响了知识的准确性和可靠性，导致人们对大模型的信任度下降。参见肖峰：《生成式大模型与知识异化探析》，《同济大学学报》（社会科学版）2024 年第 6 期。

⑨ 凯斯·桑斯坦：《网络共和国：网络社会中的民主问题》，黄维明译，上海：上海人民出版社，2003 年，第 47—48 页；蒋忠波、薛丹阳：《社交媒体时代“回音室”与“过滤泡”之辨析》，《新闻与传播评论》2024 年第 3 期。

⑩ 喻国明、侯伟鹏、程雪梅：《个性化新闻推送对新闻业务链的重塑》，《新闻记者》2017 年第 3 期。

⑪《DeepSeek 发布 R1 模型，OpenAI 推出智能体“Operator”》，https://www.doc88.com/p-40254042843297.html，访问日期：2025 年 1 月 27 日。

⑫《DeepSeek 开启 AI 算法变革元年》，http://www.jazzyear.com/study_info.html?id=147，访问日期：2025 年 1 月 27 日。

⑬ 尤尔根·哈贝马斯：《现代性的哲学话语》，曹卫东译，南京：译林出版社，2011 年，第 381 页。

⑭《技术革新引领未来——生成式 AI 塑造核心发展引擎》，https://www.vzkoo.com/document/20250207b3289984e95475d8d88ac5cf.html，访问日期：2025 年 1 月 27 日。

⑮ 周宣辰、程倩：《“信息茧房”负效应与网络思想政治教育引导作用探析》，《云南大学学报》（社会科学版）2020 年第 6 期。

⑯ 冉朝霞：《自媒体时代政府舆论导控“长尾效应”风险的破解策略》，《领导科学》2021 年第 11 期。

⑰ 卢国强、黄微、刘毅洲：《群体极化视域下突发事件网络舆情极端观点识别研究》，《情报资料工作》2023 年第 1 期。

⑱ 王建民：《网络约束的空间机制——移动互联网如何影响我们的情感体验》，《江苏社会科学》2018 年第 6 期。

⑲ 张沄溟：《试析“信息茧房”对公共精神的危害》，《北方传媒研究》2023 年第 4 期。

⑳《生成式人工智能应用发展报告(2024)》，微信公众号“智能智造 IMS”，2025 年 1 月 17 日。

㉑ 徐红昌：《信息茧房特征表现及影响因素的探索性研究》，《河北民族师范学院学报》2023 年第 3 期。

㉒《2024 年全球 AI 应用趋势年度报告》，https://m.cena.com.cn/intelligence/20250116/125648.html，访问日期：2025 年 1 月 27 日。

㉓ 赖继年、田丽雪：《“信息茧房”对红色文化传播的消极影响及破解逻辑》，《边疆经济与文化》2023 年第 11 期。

㉔ 马遥、范鹏：《健康“微传播”生态研究》，《传媒观察》2021 年第 1 期。

㉕ 第一新声智库：《2024 年中国 AI 大模型产业发展与应用研究报告》，https://baijiahao.baidu.com/s?id=1821458894256235970&wfr=spider&for=pc，访问日期：2025 年 1 月 16 日。

㉖ 孙清白：《人工智能算法的“公共性”应用风险及其二元规制》，《行政法学研究》2020 年第 4 期。

㉗《数字时代治理现代化研究报告（2023 年）》，https://mp.weixin.qq.com/s?__biz=MzI1ODczMjE0MQ==&mid=2247553324&idx=4&sn=9de599df6238e116fe6c4677b47c4870&chksm=ea01e82edd766138e6023d47d690231e558c8b1e7f153cf6c91072ddb891617cf8f0535f03bf&scene=27，访问日期：2025 年 1 月 27 日。

㉘ 潘晨子：《司法类案裁判中的 ChatGPT 应用——可能与限度》，《法理——法哲学、法学方法论与人工智能》2024 年第 1 期。

㉙ 胡明辉、蒋红艳：《构建网络群体极化与约束机制》，《学术交流》2015 年第 6 期。

㉚ 谭九生、范晓韵：《算法“黑箱”的成因、风险及其治理》，《湖南科技大学学报》（社会科学版）2020 年第 6 期。

㉛ 顾男飞：《生成式人工智能的智能涌现、风险规制与产业调控》，《荆楚法学》2023 年第 3 期。

㉜ 彭述娟：《智媒时代信息茧房效应的工具理性审思与克服》，《甘肃行政学院学报》2023 年第 5 期。

㉝ 凯斯·桑斯坦：《网络共和国：网络社会中的民主问题》，黄维明译，上海：上海人民出版社，2003 年，第 50 页。

㉞ 张凌寒：《生成式人工智能的法律定位与分层治理》，《现代法学》2023 年第 4 期。

㉟ 赵精武：《生成式人工智能应用风险治理的理论误区与路径转向》，《荆楚法学》2023 年第 3 期。

㊱ 张欣：《生成式人工智能的算法治理挑战与治理型监管》，《现代法学》2023 年第 3 期。

㊲ 孟芳：《应对人工智能伦理风险的中国方案》，《人工智能》2023 年第 5 期。

编辑 李 梅 孙冠豪

context.
Keywords: phantom limb; animal; literary community; wing poetics; popular culture

Liberate the "Cyber Industry without Assets": Legal Construction of the Right to Use Other Personal Data

Qi Yingcheng

Abstract: The disproportionate distribution of personal data revenue towards capital has garnered significant attention and critique. As a social collaboration between data source and processor, personal data is due for value sharing by both parties. Current data property rights theory schemes cannot escape the traditional private property rights regime's mental habit of determining data ownership in an exclusive dimension, offering only an all-or-nothing allocation of rights between the data source and processor. In contrast, this manuscript proposes a convivial personal data usage rights scheme. By positioning the utilization rights of personal data as a positive right, it aids individuals to actively argue for the interests of data property. It revises individual subordination in the relationship of data value distribution and also explores new possibilities for the joint construction of data wealth by individuals and data processors and institutionalized sharing for societal harmony.
Keywords: personal data; data usufruct; non-exclusivity; licensing use; collective enforcement of rights

"Reflexive Information Cocoon ": The Construction Mechanism and Transcendence of Generative Artificial Intelligence

Cao Dongying

Abstract: Compared with traditional digital technology, generative AI exhibits a novel technological form, giving rise to the "reflexive information cocoon" characterized by irrationality. The traditional information cocoon follows the generation path of "echo chamber - information cocoon - group polarization", whereas the reflexive information cocoon adheres to the logic of "technical amplifier - immersive reversal - long-tail effect". It represents an intensified version of the traditional information cocoon, impacting both users and generative AI models. Rooted in the characteristics of "immersion" and "irrationality", it poses greater potential risks, notably individual over-immersion, heightened group differentiation, and social spatial disorder. To address these issues, it is imperative to guide the rational development of users with mainstream culture, rectify the capitalization logic of models through market regulations, and safeguard the orderly landscape of "human-computer interaction" with a normative system. This approach will effectively tackle the risks and challenges inherent in the "reflexive information cocoon" at the individual, group, and social levels.
Keywords: generative artificial intelligence; reflexive information cocoon; virtual space governance; internet; human-computer interaction

The "Fuse" Mechanism of Algorithmic Administration: An Integrated Analytical Framework for Algorithmic Regulation

Zhou Ziyu

Abstract: Algorithmic technology is increasingly integrated into human society, the embedding of algorithms into government sectors forms a new paradigm of algorithmic administration. However, the imperceptibility of the algorithmic operation process and its uncertain impact on public values lead to such problems as algorithmic hegemony, algorithmic discrimination, and algorithmic bias, which are difficult to solve with the present mechanism of algorithmic regulation. To effectively deal with such problems, the "fuse" mechanism in the physics field, which is characterized by the attribute of "fuse-remedy", is introduced into algorithmic impact assessments and forms a new framework named the "fuse" mechanism of algorithmic administration. This new framework includes four main steps: scenario analysis, threshold setting, meltdown monitoring, and corrective actions. It links the mechanism of ex-ante prevention with that of after-action maintenance, which can overcome the shortage of lack of after-action remediation in the algorithmic impact assessment system. Meanwhile, it brings some important institutional implications, that is, classification matching, signal transmission, overload protection, and potential perception. The integrated analytical framework can help government sectors to timely perceive, accurately determine, and proactively prevent algorithmic risks before the spillover of algorithmic administrative risks, which may realize the forward movement of algorithmic regulation and corrections, and eventually maximize the maintenance of public value.
Keywords: algorithmic regulation; "fuse" mechanism of algorithmic administration; algorithmic impact assessments; principle of fuse; meltdown and remedies

探索与争鸣 青年学人专刊作者·巡礼

邱婕

1995年生，南京大学社会学院
博士研究生

邓剑

1987年生，苏州大学传媒学院
副教授

郑晓茹

1988年生，上海应用技术大学
人文学院校聘副研究员

施畅

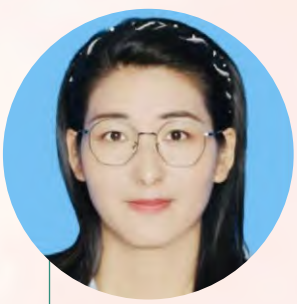
1988年生，暨南大学新闻
与传播学院副教授

张煜棪

1997年生，苏州大学文学院
博士研究生

齐英程

1991年生，吉林大学法学院
副教授



曹冬英

1982年生，云南师范大学
管理学院副教授

周子羽

2003年生，华中科技大学
公共管理学院本科生