法技并举,破除信息茧房
——推荐算法的治理之道

范　霖
中国人民警察大学　河北　廊坊　065000

摘　要：信息技术革命的日新月异，使人工智能与人类生活深度融合。而算法是人工智能时代的核心，用户可以在任何一款软件上，通过推荐算法轻松获取到自己感兴趣的内容。智能推荐算法无疑已经成为了当前时代信息传播的核心技术之一，它方便了人们的工作和生活，但网络内容服务提供商为了实现流量的资本转换，以捕捉用户的偏好来进行个性化传播作为行动逻辑，将算法推荐技术广泛应用于各类产品中，但是推荐算法的滥用也造成了价格歧视、信用歧视、就业歧视等乱象，同时也易使人们陷入“信息茧房”之中，算法治理因此成为国家网络空间治理的新领域。本文从算法系统开始，介绍了推荐算法的原理，并且通过实际例子来列举算法滥用的现状。然后通过相关的法律条例以及应用场景对国外对算法的治理情况进行了研究与概述，分析他们的优势与不足。接着对国内的算法研究现状进行法律层面与技术层面的分析，最后算法治理提出展望，在人工智能时代下，法律与技术多措并举、人与算法多重配合是当下以及未来时间内算法治理的最优解。

关键词：算法治理；网络安全；算法推荐；信息茧房；法技并举

引言

算法是解决问题的系统方法和策略机制，从特定输入生成输出。随着各种算法在我们生活中的广泛应用，它们已经对我们的决策、评估和分析产生了明显影响。随着互联网的迅速发展，人们对如何快速、准确地获取信息越来越关注。百度搜索引擎的成功就是一个典型例子，尽管它带来了便捷，但也存在固有缺陷，如缺乏个性化搜索、不能提供有效的购买建议和可能出现不相关的搜索结果。为了解决这些问题，推荐算法应运而生。然而，推荐算法的不当使用也导致了“信息茧房”现象，即用户在海量信息中只选择与自己兴趣相关的内容，忽视其他观点，这与“作茧自缚的蚕”相似。随着大数据和人工智能的普及，智能推荐算法成为了信息传播的核心技术，网络内容提供商利用这些算法进行个性化传播，从“人找信息”向“信息找人”转变。

但推荐算法也加大了算法歧视的风险，为了更好地适应人工智能时代，最大化用户和服务提供者的利益，以及减少推荐算法潜在风险，对算法进行有效治理变得至关重要。

1 算法推荐系统现状

1.1 基于协同过滤的算法推荐

协同过滤的核心思想在于借助其他用户的历史行为（群体智慧）来为当前用户给出推荐。基于协同过滤进行推荐的思想一般认为最早出现在GroupLens的新闻推荐系统中[1]，该工作也就是后来人们所说的基于用户的协同过滤方法，除此之外，该工作也第一次提出了用户物品评分矩阵的补全预测问题，并且这一问题在Herlocke中得到了进一步的形式化，并在Breese中得到了实验验证，影响了推荐系统今后十几年的发展方向；Sugiyama等将基于用户的协同过滤用于个性化搜索任务中并取得了不错的效果[2]。虽然该算法不需要消耗大量的时间，但是存在冷启动问题、所谓冷启动问题，就是指当新用户刚刚加入系统时，由于其只有很少甚至没有历史行为记录，使得协同过滤算法难以对其进行偏好建模。

1.2 基于内容过滤的算法推荐

首先，通过收集和标注特征信息，对用户和物品构建内容画像。基于内容的推荐算法利用特征匹配算法通过对特征信息的分析，实现个性化推荐。Debnath等研究了特征权重的选取方法及其对推荐效果的影响[3]；Martínez等将语言学模型运用到基于内容的推荐当中，从而允许用户以自然语言描述自身的兴趣爱好并获得个性化的推荐；Blanc和Gemmis等将语义网与基于内容的推荐相结合，利用语义网所蕴含的精确的特征关系为用户提供推荐。

2 算法乱象造成的歧视

2.1 价格歧视

价格歧视指的是商品提供者向条件相同的交易相对人提供相同等级、质量的商品时，却以不同价格进行交易的行为。企业实施价格歧视的目的在于最大化地获取消费者剩余。一般而言，成功实施价格歧视需要满足

三个条件：首先，企业必须在市场上占据主导地位；其次，企业需要深入了解不同消费者的支付意愿；最后，企业还必须能够有效防止低价购买者通过转售商品来获得套利机会。以往，企业难以获取消费者的最高支付意愿信息，但随着数据挖掘和分析工具等人工智能技术的发展，现在企业可以通过特定算法追踪和分析消费者行为，获取其偏好、习惯、支付能力等信息。这使得企业能够在大数据时代实施价格歧视，即使没有市场主导地位，也能有效实现个性化定价，这种现象被称为“大数据杀熟”[4]。

2.2 就业歧视

就业歧视是指在招聘、晋升、薪酬等就业环节中，基于种族、肤色、性别、宗教等因素对求职者或员工实施不公平待遇的行为。尽管支持者认为算法能提高人力资源管理效率，避免人类决策者的偏见，但怀疑者指出数据并非中立，算法可能加剧或产生新的偏见。比如，经过深入的数据分析，部分公司发现员工的家庭住址与工作地点之间的距离与其在职时间长短之间存在一定的关联性。这一发现为人力资源管理和员工留任策略提供了新的视角和思考方向，若人力分析系统依赖这一因素做出决策，可能对住址较远者造成不公平待遇[5]。此外，搜索引擎在就业方面也可能存在歧视，如搜索特定族裔姓名时更易弹出犯罪背景审查广告。这些现象表明，尽管算法不直接根据种族、性别等特征实施公开的歧视，但在雇主允许的情况下，歧视仍可能悄然发生。

2.3 信用歧视

信用歧视是指征信行业在评估个人信用时，基于种族、肤色、宗教信仰、性别、年龄等因素对个人信用进行不公正评价的现象。各国征信法律法规通常禁止此类歧视行为，但复杂的信用评分系统时常违反这些规定。信用歧视问题的根源在于信用评分算法的不透明性。这些算法通常受商业秘密保护，外界甚至政府监管机构难以了解其内部逻辑。征信行业声称保持算法秘密性是为了防止信用欺诈，但也有人认为即使算法公开，其复杂性也超出外界理解范围。此外，随着算法在个人信用评价中的新应用，如通过收集和分析社交媒体中的行为数据来评估财务信用，算法型信用歧视问题可能进一步加剧。这些算法以用户的社交媒体信息为基础，综合考量用户数量及其质量、网络行为所体现的人格特征、与联系人的互动模式，以及联系人的身份属性、个性特点和财务状况等因素，可能加剧基于个人特征的信用歧视现象。

3 国内外对算法在法制方面的研究

3.1 国外对算法在法治方面的研究现状

国外有关算法治理大致分为三个方向，即事前治理、事中监管式、事后救济式。事前治理是指事前禁止、备案等要求；事中监管是指在算法运行期间对其进行监督与管理；事后救济则是因算法使用不当造成不良影响后进行的救济式治理。

3.1.1 事前治理

2021年4月，欧盟委员会提出了《关于“欧洲议会和理事会条例：制定人工智能的统一规则（人工智能法案）并修订某些联盟立法》的提案，其中提出了基于风险等级区分的AI算法监管路径。该提案规定了AI算法系统在投入使用或市场投放前应确定其所属的风险等级。对于四类可能导致自然人身体或心理伤害的AI算法系统，被认定为风险不可接受，因而禁止使用和投放。此外，对于在关键基础设施、公民教育、公民就业、公共服务、执法等领域可能存在危害健康和安全或对基本权利造成不利影响的AI算法系统，被归类为高风险系统，并规定了注册义务。供应商注册时应提供的信息应向公众开放。

尽管这种治理方式有其必要性，但仍存在两大潜在问题亟待解决。其一，关于如何在事前准确判断AI算法是否“本身违法”，目前仍是一个具有挑战性的议题。欧盟《人工智能提案》第5条所采取的结果导向定义方式，在实际操作中对于预测可能产生的后果存在难度，特别是在算法决策过程缺乏透明度和可解释性的情况下，排除或确定潜在危害及其程度变得更为棘手。其二，此种监管方式的有效性可能受到AI算法结果不可预测性的制约。AI算法的输出不仅受其内部特性和设计思路的影响，更在很大程度上取决于输入数据的质量和偏差。因此，当AI算法投入实际运用后，其处理的数据可能产生与预期不符的输出结果，从而影响监管效果。

3.1.2 事中监管式

日本的《改善特定数字平台上的交易的透明度和公平性法》要求特定数字交易平台供应商在使用AI算法系统时持续向日本经济产业省部（METI minister）提交年度报告，说明其合规状况，并自我评估在法案下的义务履行情况。METI将审查这些报告，以确保交易的透明度和公平性，并将审查结果与每个供应商的报告大纲一并公布。

然而，这种事中监管式的方法存在一些不足之处：（1）AI算法供应商需要持续主动上报信息，但需要明确具体上报的信息类型以确保监管部门能够有效监测风险，并需要配套措施来保障信息的准确性、真实性和有效性。（2）主管机关持续监测和调查的核心问题在于如

何界定其权力范围。一方面，过度扩张主管机构的权力边界可能违反行政领域的比例原则，并给企业带来不必要的负担；另一方面，监管机关行使权力的条件存在较大不确定性，例如在何种情况下可以要求提供信息，可能导致权力寻租的问题。

3.1.3 事后救济式

在算法领域，事后救济规则与其他法律领域的规定相似，目的是防止危害进一步扩大并弥补受损利益相关方的损失。然而，国际社会目前更倾向于由国家权力机构行使或代为行使基于AI算法监管产生的诉权，而不是由利益受损方直接行使私人诉权。这一趋势可能与算法领域的性质有关，例如私人在行使请求权时面临的严重信息不对称现状等。

3.2 中国对算法在法治方面的研究现状

2021年9月18至20日，人民网连发三篇文章，三评算法推荐：首先强调不能完全依赖算法决定内容，而应保留编辑的角色和舆论空间；其次指出需要加强监管，避免算法困住用户在“信息茧房”中，鼓励走出去探索世界；最后警示智能信息平台可能破坏创新源动力，提出加强法规、企业承担社会责任的呼吁，强调智能平台应当注重道德与长远发展。

3.2.1 使用法律手段进行算法治理

我国的算法应用正在迅速发展，但随之而来的是“信息茧房”和大数据“杀熟”等问题。为此，我国开始加强对算法治理的关注。2021年8月，国家互联网信息办公室发布了《互联网信息服务算法推荐管理规定（征求意见稿）》，提出不建议互联网公司利用算法引导用户消费，并建议允许用户关闭算法推荐服务。相关机构表示将在三年内制定算法治理规则，并加强对个性化推荐和用户画像等算法的控制。

我国信息通信研究院积极参与算法治理监管政策的制定、技术认证和测试，并努力创建测试数据算法工具。尽管工作仍处于初期阶段，但有望为我国人工智能算法治理制度奠定基础，确保算法体系的健壮性、可靠性和可控性。我国科学技术部也宣布将制定道德标准，依靠互联网公司和算法研究人员自我监督，将算法治理原则应用于算法开发。为了消除算法歧视和偏见，国家互联网信息办公室等多个部门于2022年3月联合发布了《互联网信息服务算法推荐管理规定》。在此基础上，我国应进一步明确相应的算法保护和数据出境政策、法律法规，以确保全球性倡议和理念的安全保障，同时为我国企业的社会服务和全球商业模式提供支持。

3.2.2 使用技术手段进行算法治理

法律先行，技术辅之。技术在算法治理中扮演重要角色，以算法对抗算法是解决不良推荐算法问题的有效方式。纠正算法歧视及其潜在风险的关键，是推进算法的设计优化。利用技术手段对算法进行治理可从以下两个方面入手：（1）重新调整算法内容池的配比：需要重新配置内容池中的内容比例，并将价值导向正确性作为推送内容的独立标签。加大正能量内容的权重，使其更容易传播到每个用户。（2）改善算法的伦理和价值设计：解决偏差的策略之一是可以将公众的理解和认知融入算法决策中。算法设计的基本准则应当将公众对社会问题的普遍共识纳入其中，并将这种准则与算法技术规则同等对待，来促使算法决策不会违背社会公序良俗的基本判断准则，这是公众和推荐算法之间的良好契约。

结束语

在人工智能飞速发展的今天，推荐算法的应用日益广泛，但同时也带来了“信息茧房”等问题。为此，我们需要从法律和技术两方面着手，共同推动推荐算法的健康发展。法律层面，应制定和完善相关法规，规范算法的使用和监管，保障用户的权益和信息安全。技术层面，应不断创新和优化算法设计，提高算法的公正性、透明度和可解释性，减少算法的歧视和偏见。只有法律和技术的双重保障，才能更好地解决信息茧房问题，推动人工智能技术的健康发展。

参考文献

[1]李静辉.算法推荐意识形态属性的生成逻辑、风险及治理[J].理论导刊,2022(02):70-76.

[2]梅杰.算法传播批判——网络空间治理中的自由与秩序[J].理论导刊,2021(10):58-64.

[3]李成.人工智能歧视的法律治理[J].中国法学,2021(02):127-147.

[4]贾开.人工智能与算法治理研究[J].中国行政管理,2019(01):17-22.

[5]孙光浩,刘丹青,李梦云.个性化推荐算法综述[J].软件,2017,38(07):70-78.