

主流意识形态“自下而上”嵌入算法推荐的机制与路径研究

吴雨凡

(华南理工大学马克思主义学院, 广东省广州市, 510640; 962128939@qq.com)

摘要: 在当前国家高度重视思想政治工作、着力提升思政引领力的政策导向下, 主流意识形态的传播与建设面临算法推荐技术广泛应用带来的深刻机遇与挑战。算法推荐作为信息分发的重要渠道, 其内在逻辑与价值取向直接影响着主流意识形态的传播效能和阵地建设。当前算法推荐系统在主流意识形态传播中面临“价值空心化”“泛娱乐倾向”等挑战。针对该困局, 以“自下而上”的理论路径为切入点, 系统构建了一个主流意识形态嵌入推荐算法的机制框架, 本文从用户行为建模、内容语义标注、模型损失函数设计及平台治理机制四个维度展开, 提出具有操作性的价值嵌入路径。通过应用场景分析和实验指标设计, 探索该机制在主流媒体客户端、高校德育平台与短视频平台中的可行性与实效性。

关键词: 主流意识形态; 推荐算法; 价值嵌入; 算法伦理

引言

当前, 算法推荐机制本质上是基于数据和用户行为分析的个性化推荐, 虽然能够高效满足用户需求, 但在内容推荐中可能存在意识形态的偏颇或信息过载的问题, 甚至可能出现价值观失衡的现象。在这一背景下, 如何在算法推荐的设计与实现中融入思政引领力, 既是提升算法推荐社会责任的需求, 也是加快我国信息化进程、推动社会主义核心价值观传播的必要措施。

1. 研究背景与问题的提出

在数字技术加速渗透社会生活的当下, 算法推荐技术已成为信息传播与舆论引导的重要基础设施。特别是在人工智能与大数据技术持续演进的背景下, 以“千人千面”为特征的算法推荐系统正深刻重塑着网络内容生态、价值认知结构与社会舆论格局。算法不仅关乎信息的个性化传递, 更在潜移默化中对意识形态的传播路径、内容选择与价值导向施加影响。当前, 随着中国式现代化战略不断推进, 主流意识形态在新时代国家治理体系中承担着凝聚社会共识、维护政治安全、推动文化认同的重要功能。然而, 在平台逻辑、流量偏好与技术黑箱的合力作用下, 主流意识形态的传播效能在算法环境中正面临前所未有的挑战。一方面, 商业平台在逐利动因驱使下, 算法推荐系统容易陷入“点击至上”的单一逻辑, 导致信息茧房、娱乐至死、“把关人”失位等传播风险; 另一方面, 正能量内容因其传播结构的非强刺激性, 常常在算法排序中处于弱势地位, 难以获得足够的曝光和用户交互, 进一步削弱了主流意识形态的引领力与穿透力。

为了应对上述困境, 我国近年来出台了一系列算法治理政策, 如《互联网信息服务算法推荐管理规定》《清朗·网络平台算法典型问题治理》专项行动, 提出将主流价值观纳入算法设计, 实现算法推荐技术与社会主义核心价值观的耦合。这一“价值嵌入”导向代表了中国特色算法治理的独特路径, 即在尊重技术逻辑的基础上, 通过制度规范与文化主导, 推动技术与价值同频共振。然而, 现有的研究与政策实践多以“自上而下”的制度嵌入为核心, 强调政府监管、主流媒体主导和平台责任, 而对用户行为与算法交互在价值导向建构中的内生作用关注不足。实际上, 算法推荐系统本质上是以用户行为数据为训练依据的动态学习机制。每一次点击、点赞、评论、转发, 都构成了算法迭代演进的重要反馈信号。在长期的人机互动过程中, 算法逐渐“学习”用户的偏好特征, 塑造平台的信息流结构, 进而反哺内容生态。这一过程不仅反映出算法的技术性, 更彰显了其社会性与价值指向性。因此, 仅靠外部规训式的“强插值”治理模式, 难以实现算法价值塑造的深层嵌入。需要一种更加柔性、深层、动态的路径, 推动主流意识形态实现从“被动推送”向“用户驱动生成”的转变。

本研究拟从以下几个方面展开探讨：第一，分析主流意识形态在当前算法推荐生态中的传播困境及其生成逻辑；第二，梳理已有研究中的治理范式，指出其“自上而下”的局限性；第三，构建基于用户行为反馈的“自下而上”算法嵌入机制模型；第四，提出具有操作性的“价值标签-算法训练-分发推荐”三位一体路径；第五，以主流媒体平台、思想政治教育平台等为案例，探讨模型的实践适配与优化策略。通过上述分析，试图为构建中国特色的“价值算法融合模型”提供理论支撑与实践路径，以实现技术逻辑与国家治理逻辑的有机统一，推动算法推荐技术真正成为意识形态传播的积极动能。

2. 文献综述与理论基础

2.1. 国内外研究现状

2.1.1. 国内研究进展：从“技术嵌入”到“价值引导”

近年来，随着数字平台对公共舆论场的重构，国内学界对算法推荐与主流意识形态之间的关系展开了多维度探讨，研究议题主要集中于以下三个方面。一是对算法推荐技术本体机制的研究。学者普遍认为，算法推荐以用户画像为基础，在信息分发中通过行为数据驱动的自动化计算系统重塑信息流结构[1, 2]。这种以个性化为导向的推荐方式虽然提高了内容匹配效率，但也导致内容呈现的圈层化、价值导向的去中心化，进而弱化了主流价值的引导能力[3]。二是对算法技术对意识形态传播影响的评估。研究指出，在技术工具理性的主导下，算法系统常常将点击率、互动率作为唯一价值标准，从而放大低俗内容、削弱主流价值、抹平公共议题的严肃性[4]。与此同时，算法的“黑箱性”“资本嵌套性”也在一定程度上解构了主流意识形态的话语权与正当性。三是对算法治理与意识形态安全的应对路径提出策略建议。如朱萌提出在内容审核与特征标签体系中嵌入社会主义核心价值观标签，通过“清单化”方式构建一套价值识别参数体系，实现对“正能量”内容的标记、识别与优先推荐[5]。其他研究者则建议从法治保障、平台制度建设、算法透明机制等方面优化推荐系统的信息传递路径[6-8]。

总体来看，国内研究已初步形成“意识形态嵌入算法”的问题意识和方向性路径，但大多停留于制度治理与规则制定层面，普遍采纳的是“自上而下”的价值干预逻辑，尚未从用户行为驱动的视角探索价值如何在算法运行机制中“自然涌现”与“自适应反馈”的可能性。

2.1.2. 国外研究视野：从技术中立到伦理学习系统

在国外学界，推荐算法与道德责任、价值引导的关系已成为人工智能伦理研究的重要议题。其研究路径主要分为两个方向，一是对算法推荐系统“黑箱性”与社会偏见问题的揭示。Gillespie在《Custodians of the Internet》中指出，推荐算法作为一种“话语架构器”，具有结构性偏见，其背后植入的内容选择机制、排序逻辑和数据预设常常体现出特定政治立场和商业逻辑[9]。Milano等进一步提出，算法在筛选信息、建构世界图景的过程中，实际上是一种“规范的编码”，会在用户行为中反复强化主导性文化[10]。二是对伦理推荐系统与“自下而上”学习机制的探索。Silvia提出“道德机器”的概念，强调应从人类行为经验中提炼出价值规则，使人工智能系统具备“内生演化”的伦理判断能力。这种“自下而上”的路径强调通过机器学习持续学习行为数据中的偏好信号，构建出适应性更强、更具公共性共识的推荐模型，避免算法价值僵化与道德片面性[10]。此外，Zuboff从“监控资本主义”视角出发，揭示算法推荐在数字平台中不仅是信息推送工具，更是权力与意识形态的再生产机制，要求对推荐系统背后的政治经济结构予以重估。这些研究表明，推荐算法不应仅被视为中性工具，而应作为价值导向、意识形态嵌入的潜在机制加以理解与塑造。

2.2. 研究评述与问题归纳

纵览国内外已有研究，可将学界已有的研究内容归纳为以下三点。

第一，研究逻辑上倾向于“自上而下”的干预治理视角。国内研究多依赖制度引导、平台规约、政策驱动的治理模式，主张通过“清单管理”“算法透明”“平台责任”来实现价值导向的算法控制。虽然这些路径具有可操作性与现实支撑，但其面临的挑战在于：一方面平台内部运行逻辑仍以商业效率为主导，难以真正从技术系统层面内嵌主流价值；另一方面制度外部规训往往缺乏微观反馈机制，难以形成可持续的价值认同与用户响应。

第二，对用户—平台—算法三者互动关系的重视不足。当前的研究普遍忽视用户在算法演化中的“隐性参与”角色。事实上，每一次用户点击、转发、评论、停留时间，都是系统学习与偏好塑造的依据。用户的集体行为构成了平台内容池的排序依据，也对算法模型的训练产生反馈影响。如果无法将主流价值观转换为用户行为偏好特征，那么“技术与价值”便无法完成融合。

第三，缺乏可量化、可转译的“价值标签体系”与嵌入机制。如何将主流意识形态从抽象的文化认同转化为推荐算法中的具体标签参数，是实现“主流价值嵌入”的核心技术问题。目前尚缺乏一套可推广、可调整的价值识别指标体系，能够在算法训练模型中发挥有效作用。

2.3. 理论基础：自下而上的价值建构路径

为了弥补现有研究的不足，引入“自下而上”的算法价值嵌入路径作为理论基础，其核心理念可归结为从道德学习理论到推荐系统伦理、自下而上的社会学习路径和算法治理的协同三个方面。

2.3.1. 从道德学习理论到推荐系统伦理

Wallach & Allen [11] 提出的“自下而上”的道德学习模型强调，在不确定或快速演变的环境中，伦理系统不应完全依赖静态规则，而应通过对行为反馈的观察与归纳，提炼出具有普适性的价值规则。这一理念应用于推荐系统时，即强调通过用户行为数据挖掘个体或群体的价值偏好，从而构建推荐逻辑中价值约束机制。该路径不强调对算法结果的“事后管控”，而强调通过输入端的价值引导与行为建模，在训练过程中嵌入正向价值信号，使主流意识形态“自然进入”推荐逻辑之中。

2.3.2. 自下而上的社会学习路径

“自下而上”路径体现了从微观行为到宏观制度的建构逻辑。这一逻辑强调，制度化的价值秩序并非完全外部赋予，而是通过个体之间的交互、反复实践、偏好反馈逐步内生出来的系统性规则。在数字平台环境中，这种路径尤为重要：平台算法并不“制定规则”，而是通过学习用户行为不断“更新规则”。因此，通过识别用户在信息互动中的主流价值偏好（如对爱国、诚信、公正等内容的持续关注与高参与），就可以构建出一套“数据驱动的主流价值标签体系”，为推荐系统训练提供价值信号源。

2.3.3. 算法治理的协同模型

在“自下而上”嵌入路径中，算法不再被视为可以被单一主体彻底掌控的工具，而是作为一个不断演化、与社会行为动态耦合的复杂系统，其价值导向的塑造不能仅靠政府行政手段的单向输入，而必须依靠用户、平台和政府三类行为体的协同配合与交互塑造。每一主体都在价值嵌入过程中扮演关键角色，三者之间的动态协同构成了“价值治理共同体”。

第一，用户作为价值偏好信号的提供者，“自下而上”的核心机制是用户行为的可观察性与可转化性。在现代推荐算法体系中，用户的每一次点击、点赞、转发、评论、停留时间、互动频率等，都会被记录为算法系统的输入变量。这些行为不仅是兴趣偏好的显现，也隐含了用户的价值取向。例如，当用户频繁与含有“爱国主义”、“社会公正”、“公益行动”等内容的短视频进行深度互动时，算法系统会识别其偏好，并在下一轮推荐中提高相关内容的推送概率。因此，用户不仅是信息的接收者，更是价值传播路径的起点。算法在长期训练中，会逐步形成对用户行为与价值倾向之间的映射关系。若能构建识别用户“正面价值行为”的行为语义模型，即可在训练过程中给予主流意识形态更多“正反馈信号”，促进其内容在平台系统中的优先分发。进一步而言，用户行为的价值表达具有非命令性与非对抗性特点，是软性、间接的，能够自然嵌入算法学习机制，避免强制介入所带来的技术封锁或平台反弹。尤其在高校、主流媒体等思想政治教育场景中，通过用户真实行为的动态采集，可构建“行为—内容—价值”三维标签体系，实现对主流价值观传播的精细化支持与识别。

第二，平台作为价值表达机制的技术载体，平台在算法治理体系中处于“中介耦合”位置，既连接用户行为数据，又负责训练和部署算法系统，是推荐逻辑生成与传播规则运作的直接承载体。平台的核心作用包括三个方面：第一，标签设计与内容标注机制的开发者。平台拥有内容池的管理权限，可在推荐系统中设定“正能量标签”“主旋律内容优先池”等分类机制。如果政府或用户提供的价值标签可以通过自然语言处理模型与现有内容建立对应关系，平台便可在训练过程中嵌入“价值语义嵌套”。第二，模型训练与参数调优的参与者。推荐算法的核心是训练模型的参数设置。平台可以基于用户正向行为反馈数据，建立“价值权重因子”，将其嵌入排序机制之中。例如，将主流意识形态内容设置为具有更高冷启动权重，使其在冷启动阶段获得更多初始曝光机会，从而打破内容“边缘化”的困境。第三，反馈循环的管理与更新机制构建者。平台负责持续采集用户行为数据，并根据反馈更新推荐模型。当发现用户对某类正面内容持续产生正向反馈时，平台可对该类内容进行标签聚合与特征抽取，推动形成新的内容优先级权重，形成“自适应内容推荐规则”，实现从“值的标签”到“值的内容”的自然演化。因此，平台在整个治理模型中起到技术桥梁作用，是实现价值与算法耦合的关键一环。

第三，政府作为价值方向的制度引导者，政府在算法价值嵌入体系中既不应成为唯一控制者，也不应完全退出调控领域，而应扮演“方向设定者”“标准提供者”“治理保障者”的复合型角色。首先，政府作为价值共识的制定者，应提供可识别、可嵌入的主流价值指标体系。如朱萌所提出的“社会主义核心价值观十

二项指标”及其下设43个二级描述性指标，政府可以通过专家研判、学界协同，将其“清单化”为平台可使用的标签语料库、关键词向量集或语义词袋，用于内容语义识别与用户偏好分析[5]。其次，政府应推动平台设立“价值可信度调控机制”。通过设置“内容审核+用户偏好匹配+行为数据回馈”三重价值检索机制，在平台算法训练阶段引入“可信标签评估”指标，使正面内容在传播过程中获得算法持续识别。最后，政府需提供法律制度保障，设定推荐算法的伦理边界与监管底线。如《互联网信息服务算法推荐管理规定》《清朗行动》均体现出国家对算法价值导向的制度性回应。在实践层面，建议政府联合高校、主流媒体和技术平台，建立“算法治理协同实验室”，推动“数据可共享、模型可调试、价值可量化”的跨平台治理试点，强化伦理指导能力与算法可追责能力。综上，用户、平台与政府三方在“自下而上”的算法价值嵌入模型中承担着差异化、协同性强的功能角色。

表1 用户、平台与政府三方主体在算法价值嵌入模型中的功能角色表

主体	角色功能	嵌入机制	成效路径
用户	价值偏好信号提供者	行为数据、互动反馈	提供训练样本、塑造主流偏好模型
平台	技术载体与机制执行者	标签设计、排序控制、算法调优	实现标签嵌入与行为转译
政府	价值方向设定者与制度保障者	指标清单、政策引导、法制监管	提供语义标准与伦理边界

如表1所示，构建这一协同模型不仅有助于构建具有中国特色的算法伦理治理体系，也为主流意识形态的智能化传播提供了系统性框架和实现路径，推动形成“价值共建—数据驱动—技术反馈”的动态传播循环机制。

3. “自下而上”路径下的算法价值嵌入逻辑

在传统意识形态嵌入研究中，主流路径多以政府管控和平台责任为核心，试图通过规制手段引导信息流向。然而，在高度智能化的信息分发环境中，算法已具备高度自主的学习与进化能力，价值嵌入若脱离算法运行的“内在逻辑”，易流于形式或产生“逆反传播”效应。因此，主张从“自下而上”的视角出发，将用户行为反馈、平台内容机制与算法模型调试三者联动，构建出一种基于行为驱动的主流意识形态内嵌路径。

3.1. 用户行为的价值表达逻辑

用户在算法推荐系统中的行为并非孤立存在，而是嵌套于一定社会文化结构之中，具有明确的符号指向与价值表达。在“自下而上”的逻辑中，用户的微观行为数据被视为价值反馈信号，其通过行为语义的结构特征、主流价值内容的识别机制两个路径影响推荐系统的价值偏向。

3.1.1. 行为语义的结构特征

不同类型的用户行为所表达的价值语义存在差异。例如，“点击”行为更多代表兴趣初印象，“停留时间”反映内容吸引力，而“转发”和“评论”则多代表用户主动传播与立场表态。其中，“正向转发+积极评论”组合行为可作为高强度价值认同信号的重要指标。如表2所示，通过构建行为组合矩阵，可实现对“主流意识形态响应度”的初步建模。

表2 用户行为价值语义表

行为类型	所代表含义	可量化指标	价值权重建议
点击	内容吸引力初步确认	CTR (Click-Through Rate)	低
停留	信息内容匹配程度	Avg. View Time	中
评论	认知参与与态度表达	Comment Polarity	高
转发	主动传播意愿	Share Rate	高

3.1.2. 主流价值内容的识别机制

为使系统识别“主流意识形态响应行为”，平台需构建一个包含价值标签的内容语料库，如表3所示，以“社会主义核心价值观”为基础，构建涵盖爱国、敬业、诚信、友善等标签的内容分类系统。此标签系统可作为推荐算法中的“价值嵌入模板”，用于内容初筛、推荐权重设定与正向内容识别。可采用自然语言处理（NLP）技术对用户评论、转发语句进行情感与价值分析，生成用户的“价值偏好画像”，并据此对用户进行价值向度划分。

表3 用户行为划分表

用户类型	特征行为	推荐策略
高响应用户	对主流价值内容持续高互动	提高正向内容比例
潜在倾向用户	有局部正向行为，但偏好泛娱乐内容	适度引导，标签调整
弱响应用户	无明显价值偏好，行为随机	中性推荐、行为引导

3.2. 平台机制中的价值生成路径

在用户行为信号输入之后，平台系统承担起信息处理、模型训练、内容分发的中介功能。在“自下而上”的路径中，平台不再是价值的“灌输者”，而是连接行为、标签与内容之间的“耦合建构者”。

3.2.1. 构建行为-标签映射机制

平台需搭建一个“行为数据—价值标签”之间的映射机制，将高响应行为对应到预设的主流价值标签。例如，在用户大量转发“抗震救灾”“科技强国”等内容后，平台可通过行为聚类模型将其行为特征标记为“爱国”“社会责任”等标签，并反馈至用户画像中。这一过程可借助协同过滤和词向量模型实现，第一步，使用NLP技术对内容进行价值标签标注；第二步，将用户行为聚类，映射其偏好价值维度；第三步，更新推荐模型，嵌入价值标签偏好。

3.2.2. 主流价值内容池的建设机制

平台可设计“主流价值优先内容池”，并基于价值响应模型动态调整其内容入选权重。例如，当某一类正向视频（如“航天精神宣传片”）获得高停留与高分享反馈，其可被平台标记为“高反馈值内容”，并入驻内容池，以更高频次出现在推荐流中。内容池推荐算法可借助以下优化模型：一是通过内容质量函数 $Q(c) = \alpha \cdot \text{点击率} + \beta \cdot \text{停留时间} + \gamma \cdot \text{正面评论率} + \delta \cdot \text{价值匹配度}$ ；二是设定阈值函数 $F(Q) > \lambda$ 即可进入优先推荐池；三是定期由平台进行“价值反馈校验”，更新内容池构成。

3.2.3. 模型训练的动态调优机制

推荐模型的训练逻辑需设置“正向价值权重因子”，即对带有主流价值标签的内容，在模型初期阶段给予更高冷启动权重与推荐频率。此外，平台还应构建“负向价值惩罚机制”，对泛低俗、标题党、价值中立但高度煽动性的内容进行降权处理，提升主流内容在推荐排序中的相对优势。

3.3. 算法系统中的嵌入实现逻辑

推荐算法的实质是一个持续学习与反馈优化的过程。在“自下而上”的路径中，算法系统需具备识别、调整与反馈能力，实现从“用户偏好”到“系统推荐”的价值闭环。

3.3.1. 嵌入机制一：价值信号的提取与建模

该环节通过用户行为与标签内容交叉分析，建立“价值评分模型”：

- (1) 用户行为 (b) × 内容价值标签 (v) → 生成响应系数 $R(b,v)$
- (2) 对所有用户-内容对建立响应矩阵 $R = [r_{ij}]$
- (3) 将该矩阵输入推荐算法，更新偏好权重向量 θ

可使用矩阵分解与聚类识别技术优化该建模过程，使价值标签作为算法学习过程中的辅助向量特征。

3.3.2. 嵌入机制二：训练过程中的权重调整

训练算法时，在损失函数中加入价值导向调整项：

- (1) $L = L_{\text{recommend}} + \lambda \cdot L_{\text{value}}$
- (2) 其中 $L_{\text{value}} = \sum (\text{用户行为偏好} - \text{主流价值期望})^2$

该损失项能够在模型偏向非主流内容时产生“反向拉力”，引导其回归价值目标。

3.3.3. 嵌入机制三：分发排序策略的价值优先级

最终推荐输出阶段，可设定内容综合推荐分值：

$\text{Score}(c_i) = f(\text{行为匹配度}) + \alpha \cdot \text{价值优先值} + \beta \cdot \text{社会关注热度}$

此机制通过提高主流价值内容的排序起始位置（Top Rank Boosting），实现“算法自主下的价值引导”，而非强行干预式推送。

3.4. 典型案例初探：平台价值推荐实践路径

为验证上述嵌入机制的可行性，如表4所示，参考了以下平台的初步实践逻辑。这些平台虽处于不同发展阶段，但其共同趋势是：从“单向供给”转向“数据反馈—模型优化—价值自驱动”的推荐逻辑，逐步形成具有中国特色的主流意识形态算法嵌入机制雏形。

表4 平台差异化嵌入策略表

平台	机制设计	嵌入策略
人民日报客户端	用户标签细分+价值内容池	设置“主旋律加权”字段提升权重
学习强国	内容分层设计+红色积分机制	用户行为驱动内容分发频率
哔哩哔哩（B站）	青年价值传播工程	对“青春中国”等系列设置专项引导推荐位
抖音“强国号”	热门内容偏好再定向	通过点赞—转发反馈优化主流标签推荐算法

4. 主流意识形态嵌入路径设计

在“自下而上”算法嵌入模型的理论基础上，如何将主流意识形态转化为可识别、可量化、可操作的算法标签体系，关键在于路径设计的系统化与场景适配的灵活性。该路径设计不仅要求贯通用户—平台—政府之间的交互逻辑，更应兼顾数据训练的技术特性与价值表达的文化语义特征。为此，提出四种可实施的嵌入路径，分别对应内容层、用户层、模型层与平台层，形成“价值识别—行为映射—系统反馈—治理协同”的整体机制。

4.1. 价值内容标注机制：标签化转译的知识体系

价值识别的首要步骤是对内容进行结构化语义标注，即构建一套“主流意识形态内容标签语料库”，将社会主义核心价值观转化为机器可读的“内容参数”。如表5所示，可初步建立如下标签分类表，为实现算法可识别，需对标签进行向量化处理，可采用TF-IDF、Word2Vec、BERT等词嵌入技术构建“语义空间”。

表5 标签编码表

一级维度	子标签示例	对应语义关键词
爱国主义	抗疫精神、边疆建设、强军梦	国家、英雄、统一
社会公正	教育公平、环境正义、扶贫	公平、法治、责任
集体主义	志愿服务、社区共建	互助、团队、奉献

对平台内容（视频标题、简介、用户评论）进行自然语言处理分析，判断其与各价值标签的相关度。方法为基于BERT词向量模型的语义匹配评分，通过构建价值标签的“词袋集合”，计算内容文本与标签词集的余弦相似度。最后构建内容-标签相关性矩阵，确定主标签归属。首先，初期由内容审核专家团队建立高质量标注语料；后期训练分类器进行自动批量标注，提升效率；模型结果由专家进行抽检，形成“半自动-半监督”标注机制闭环。

4.2. 用户画像与偏好建构机制：价值导向行为识别路径

“自下而上”路径强调用户行为作为价值信号的输入通道，需建立用户在平台上的价值偏好标签画像系统，引导推荐模型更精准匹配主流内容。

4.2.1. 用户行为与价值标签的映射机制

在用户价值偏好建模过程中，首先可通过多维行为特征识别用户对主流意识形态内容的响应强度。具体而言，若用户在点击率、停留时间、转发频率及评论情感强度等维度上，长期表现出对特定价值类型内容高于平台平均水平的互动表现，则可将其标记为“高偏好用户”。进一步地，可借助协同过滤算法与聚类分析方法，将用户划分为若干具有相似价值取向的群组，从而构建更具针对性的用户画像体系。最终，通过提取各类用户群体的典型行为模式，并将其转化为推荐系统中的输入变量，有助于实现主流意识形态内容的精准推送与个性化匹配。

4.2.2. 用户标签系统设计示意

如表6所示，基于以上用户行为与价值标签的映射机制的标注，可以设计用户偏好及行为标签推荐策略表如下。

表6 用户偏好及行为标签推荐策略表

用户类型	偏好维度	行为特征	推荐策略
爱国倾向型	高匹配“国家、军队、科技强国”等内容	转发多、评论积极	提升爱国类内容初始权重
公共责任型	高匹配“志愿者、环保”等	点赞高、浏览时间长	优先推荐“社会参与”类内容
娱乐中立型	偏好搞笑、泛文化内容	高频点赞泛娱乐	引导式插入价值内容，避免突兀干预

4.2.3. 用户反馈与标签更新机制

在推荐系统运行过程中，用户的每一次新行为数据都将被系统实时采集，并用于动态更新其价值偏好向量，进而构建“标签动态进化模型”。该机制要求平台具备周期性回归分析能力，通常以7-14天为一个更新周期，对用户的价值倾向进行持续监测与建模修正。根据用户在主流意识形态内容上的响应行为，系统可相应调整其个性化推荐策略，增强内容匹配的精准度。若用户在特定阶段对主流价值内容表现出“认同增强”特征，即在点击、评论、转发等方面呈现显著增长趋势，则应同步提升该类内容在其推荐流中的比例，以实现价值导向与用户兴趣的动态耦合。

4.3. 模型训练调优机制：价值因子的技术建构

推荐算法的训练过程中嵌入价值因子，是整个机制能否持续运行的核心。提出基于损失函数调整与排序函数引导的双重机制。

4.3.1. 价值感知的加权损失函数构建

已在第三章中定义加权损失函数：

$$L = L_{rec} + \lambda L_{val} = \sum (r_{ij} - \widehat{r}_{ij})^2 + \lambda \sum |p_i - v_j|^{21}$$

4.3.2. 模型冷启动阶段的价值补偿策略

在推荐系统的冷启动阶段，面临用户行为数据和内容反馈信息严重不足的问题，尤其是在新用户首次接入或主流内容刚上线时，推荐算法难以准确判断其传播潜力与匹配度。针对这一挑战，可采用多种策略优化主流意识形态内容在冷启动阶段的可见度与分发效率。可通过设定“主流内容初始权重提升”机制，例如在“学习强国”等平台中，为红色题材内容预设较高的基础曝光权重（如 $\alpha=1.2$ ），提升其进入推荐流的初始排序位次。并引入“置信度插值补偿”策略，即将专家对主流内容的价值评估标签转化为推荐得分的一部分，以弥补用户行为数据缺失所带来的推荐盲区。最后，系统可设置“价值预设标签推荐位”，即在推荐结果中保留固定数量（Top-N）的主流意识形态内容插槽，确保其在算法启动初期具备基本曝光机会，从而实现价值传播的先发优势。

4.3.3. 推荐排序函数中的价值融合因子

最终推荐输出排序模型为：

$$Score(c) = f_{behavior} + \alpha \cdot f_{value} + \beta \cdot f_{hotness}$$

其中：

$$f_{behavior}$$

$$f_{value}$$

$$f_{hotness}$$

α 值可根据平台政治任务需求动态调整，如节庆期间提升“红色内容”排序优先权。

4.4. 平台协同治理机制：规则框架与价值反馈通道

作为连接技术与治理的中介场域，平台的制度保障与运营机制是实现嵌入路径落地的制度性保障。

4.4.1. 嵌入价值内容的审核机制

为保障主流意识形态在推荐系统中的表达质量与价值准确性，平台应建立“内容—价值双轨审核机制”。该机制融合人工审核与智能识别两种方式：一方面由人工审核团队对内容的思想导向与政治立场进行把关，确保其符合主流价值标准；另一方面借助自然语言处理等技术手段对内容的语义结构进行自动分析与标签标注，实现对大规模信息流的快速价值评估。在此基础上，平台可为每条内容设定“价值可信度”评分，分值范围为0至1，用以衡量其与主流意识形态标签的匹配程度。当某内容的可信度评分超过设定阈值时，系统可将其自动纳入“价值推荐池”，提高其在推荐排序中的初始权重。此外，为提升审核机制的弹性与容错能力，平台还应设立“价值纠错通道”，允许专家与普通用户对机器误判的内容标签提出申诉与修正建议，构建多元参与、动态校正的内容治理体系。

4.4.2. 平台治理指标设计建议

为实现系统化运营，如表7所示，平台可设置以下价值治理指标体系。这些指标既可用于系统自评，也可用于政府监管部门的政策反馈机制，构建“责任-价值-算法”之间的透明协商通道。

表7 平台治理指标设计表

指标名称	含义	数据来源
价值标签内容分发占比	推荐系统中主流内容占比	内容池统计
主流内容点击率提升幅度	与普通内容点击率比较	用户行为数据库
用户行为偏好倾向分布	主流偏好用户比例变化	用户画像系统
错误标记率/纠错率	标签错配与修正频次	内容审核系统
平台治理执行指数	标签同步更新频次	系统操作日志

4.4.3. 平台与政府协同机制建议

为进一步提升主流意识形态嵌入机制的实效性 with 制度化水平，建议构建平台治理、学界研究、政府监管三方协同治理体系，推动价值引导与算法设计的深度融合。在此机制中，平台方应设立专门的“算法伦理办公室”，负责主流价值内容的算法嵌入试验、推荐模型的伦理风险评估及技术路径的持续迭代；学术机构可作为中立智库，承担价值标签体系的理论更新、用户行为建模方法的优化，以及算法偏见治理的技术支持任务；监管部门则应从宏观治理层面入手，负责制定统一的价值标签标准与评价指标体系，完善相应政策工具与激励机制，引导平台建设正向传播生态。三方协作还可联合打造“价值调控开源平台”，共享模型接口、行为数据规范与治理案例库，以实现跨平台、跨行业的价值协同与技术共建，推动中国特色的算法伦理治理体系走向成熟。

本研究提出的主流意识形态嵌入路径设计，体现了从内容识别、用户建模、算法训练到平台治理的全链路协同逻辑，是“自下而上”理论与技术实现的有机统一。该路径不同于“行政推送”式的粗放治理，而是一种基于行为反馈、技术自组织与价值协同的精准治理路径，能更有效地在算法主导的传播生态中提升主流意识形态的影响力、亲和力与嵌入力。

5. 机制实证建议与应用场景

在完成“自下而上”路径下主流意识形态嵌入算法推荐机制的模型设计与技术路径构建后，关键的问题在于：如何将这些理论机制落地于具体平台场景之中？如何在真实系统中评估该机制的可行性、有效性与价值引导能力？为此，以当前三类典型数字平台为研究对象：主流媒体客户端（如人民日报、央视新闻）、高校德育平台（如学习强国、智慧树等）、短视频平台（如抖音、快手“强国号”），结合各自运行逻辑与用户行为特征，提出机制实证应用路径，构建可测量、可反馈、可迭代的主流意识形态嵌入实验系统。

5.1. 主流媒体客户端：价值标签驱动的分发机制测试

主流媒体平台在主流意识形态传播中天然具备政治属性与公共服务使命，但其在推荐系统中的主流内容分发能力却面临“冷启动难、标签弱化、同质化严重”等挑战。

5.1.1. 建立主流内容池标签语料库

为了构建具备高识别度和可计算性的主流意识形态内容基础，平台可由编辑团队与专家组协同制定覆盖政治、经济、文化、社会等维度的“主流意识形态标签体系”，作为内容语义识别的核心依据。在此基础上，结合自然语言处理（NLP）技术对历史内容进行批量标注，构建推荐系统训练所需的语料数据集。具体技术路径可采用多模组合策略：首先通过TF-IDF与LDA模型提取文本关键词与主题结构，初步完成内容的语义聚类；随后引入BERT向量模型对内容文本与价值标签之间的语义相似度进行深层计算，提高标签归属判断的精度与灵敏性；最终依据相似度计算结果为每条内容赋予0-1区间的“价值含量得分”，该得分既可作为模型冷启动阶段的初始排序依据，也可在后续推荐系统中作为价值排序因子嵌入。此种融合专家知识与语义模型的标签体系，不仅增强了主流内容的算法可识别性，也为价值导向内容的精准分发奠定了技术基础。

5.1.2. 用户行为反馈路径追踪实验设计

在主流意识形态嵌入机制的实证检验中，可采用A/B测试法对推荐策略进行对比实验。具体设计为：设立两组推荐系统样本，其中A组采用传统的算法逻辑，不加入任何价值权重调节；B组则在推荐排序函数中引入主流意识形态标签权重（即Score函数中的 $\alpha \neq 0$ ），以强化主流内容的分发优先级。

表8 指标体系建构表

指标	指标含义	评价目标
曝光率增长比	实验组主流内容曝光变化率	推荐排序优化效果
用户平均停留时间	用户在主流内容停留秒数	内容吸引力变化
主流内容点击率提升幅度	点击频率对比实验前	用户偏好调整趋势
主流内容转发率	内容传播范围变化	群体共鸣指数
用户正向行为比例变化	表态型评论、点赞等	价值认同外显趋势

实验过程中，平台应持续跟踪用户点击率、平均停留时长、转发频次等关键行为指标的变化，分析不同推荐策略对用户行为的影响差异，进而评估主流内容嵌入机制的传播效果与用户接受度。如表8所示，例如人民日报客户端可选取“科技强国”主题内容作为实验对象，对比加入标签权重前后该类内容的曝光量与用户正向互动水平的变化，从而验证主流标签嵌入在实际推荐场景中的有效性与优化潜力。

5.2. 高校德育平台：学生价值偏好建模与引导机制测试

高校德育平台如“学习强国”“智慧树”在学生思想政治教育中发挥着不可替代作用。传统模式侧重“教师推送+课程植入”，缺乏与学生兴趣行为数据的结合。“自下而上”路径可为该类平台构建一种“反馈式教学—兴趣驱动—价值内化”的传播路径。

5.2.1. 构建学生价值偏好画像系统

基于前述用户—价值映射模型，高校德育平台可通过系统化的数据采集，对学生在平台内的多维行为进行建模，以构建其个体化的主流意识形态价值偏好画像。具体行为数据包括：学生在德育类视频资源中的观看时长与观看频次，用以评估其内容接触深度与持续性；课程参与情况，如签到率、课中测验与互动答题频率，反映其对主流价值议题的学习主动性与参与度；评论与转发行为，则有助于识别其价值表达与再传播倾向；此外，还应重点跟踪学生在思政类平台（如“学习强国”）的“话题参与度”，即其在特定主流议题下的发言次数、观点态度与互动质量。上述行为特征共同构成学生价值偏好向量，为平台后续实现个性化、价值导向型内容推荐提供坚实的数据基础。

5.2.2. 价值标签驱动式内容精准推送实验

基于现代教学平台对德育功能的需求，如表9所示，研究构建了一套系统化的价值引领机制。首先，设计并建立了“价值内容精准推荐模块”，该模块通过数据分析和学生偏好挖掘，在平台首页自动生成符合主流价值观的内容池，实现个性化的价值内容推送，增强内容的针对性和有效性。其次，平台实现了课内与课外德育资源的深度联动，具体表现为将“课程学习进度”与“主题视频推荐”等资源有机结合，形成教学过程中的闭环反馈机制，从而促进学生在不同学习场景中的持续关注与内化。最后，平台引入激励机制，学生在

观看主流价值内容后可获得相应的“价值积分”，该积分可用于兑换多样化的学习资源，有效激发学生的学习动力和参与积极性，推动德育教育的深入开展。该机制的构建为提升德育教学的智能化水平和学生价值认同感提供了有力支撑。

表9 教学评估指标表

指标名称	说明	数据采集方式
价值行为响应度	学生对主流内容点击/评论比率	后台交互日志
价值内容偏好差异指数	各学生偏好维度离散度	用户画像矩阵
课堂教学参与度变化	试验前后教学活跃度变化	教学日志
主流内容学习时长比重	总学习时间中占比	后台记录
学生认同度回访调查	教育成效问卷	人工调研

5.3. 短视频平台：流量模型中主流价值引导机制的实实验证

以抖音、快手为代表的短视频平台具有海量流量与碎片化传播特征，主流价值内容往往被边缘化或淹没在泛娱乐流中。“自下而上”路径强调通过数据反馈、行为建模与标签引导，逐步建立“低干预、高共鸣”的内容推送机制。

5.3.1. 构建“价值共鸣热点推荐引擎”

平台通过对用户高频转发行为与评论情绪进行系统性分析，深入挖掘具有广泛共鸣效应的主流价值内容热点，典型案例包括“成都大运会精神”与“志愿者故事”等。这一分析过程不仅揭示了用户对特定价值观念的认同与反馈，也为内容优化提供了数据支持。基于此，平台构建了结构化的主流内容候选池，并在界面设计中设置专门的曝光引导位，旨在提升优质内容的可见度与传播广度，增强内容的影响力和覆盖面。同时，针对表现出明显价值共鸣行为特征的用户群体，平台实施精准推荐策略，确保推荐内容能够更有效地触达目标用户，提升推荐的匹配度与接受度。通过构建“内容—群体—平台”三方协同共建机制，不仅促进了优质主流价值内容的广泛传播，也推动了平台生态的持续优化与健康发展，为实现教育平台价值引领目标提供了坚实支撑。

5.3.2. 热门话题引导机制嵌入

如表10所示，平台通过每日自动识别主流价值内容中的高热度话题，依托#话题分析技术，实时捕捉社会热点与用户关注焦点，确保内容推送的时效性与针对性。推荐算法将这些识别出的高热度话题作为个性化的“推荐因子”注入内容排序机制，优化内容展示顺序，从而提升相关内容的曝光率及用户匹配度，增强用户的参与感与认同感。同时，用户的行为数据作为系统反馈，反哺推荐模型，通过动态训练“热点—价值—传播效率”函数关系，实现推荐机制的持续迭代与精准优化。以“五四青年节”期间“强国青年”系列短视频推荐热度提升实验为例，该实验定量分析了价值内容与特定时间节点及用户群体之间的强相关关系，验证了推荐机制在提升价值内容传播效果和精准触达方面的有效性。此类机制不仅提升了平台内容的传播效率，也为主流价值观的传播提供了数据驱动的技术支持。

表10 平台策略与风险控制表

策略名称	内容	预期成效
内容识别前置	对UGC内容加入主流价值语义扫描	减少内容偏离主线
标签调权后置	对部分主流价值内容加入分发优先因子	提升曝光率
用户行为加权	转发+正向评论 =主流喜好得分	精准识别偏好用户群
风险内容过滤机制	建立反价值标签池与异常行为检测模型	控制舆情风险扩散

5.4. 跨平台联动与持续优化路径建议

鉴于价值嵌入机制需依托长期反馈与多平台协同，实现价值引领的系统化和科学化，建议成立由政府牵头、平台参与、学术界支持的算法价值协同治理机制实验室。该实验室主要任务包括：首先，联合制定统一

的价值标签标准,确保跨平台价值内容的标识和分类具有一致性与规范性;其次,构建跨平台的推荐效果追踪体系和模型共享机制,实现多平台间数据与算法的互联互通,提升推荐系统的整体效能;再次,鼓励各平台间共享“主流内容高响应模型参数”,形成以价值传播为核心的协同网络,增强主流价值内容的广泛覆盖和影响力;最后,开展定期(季度级)评估与价值引导指数发布,科学监测主流价值内容的传播效果及舆论生态状况,推动舆论环境的持续优化与健康发展。通过上述举措,能够有效促进价值引领机制的多维协同与动态提升。

6. 结语

在全球数字化转型加速的背景下,算法技术正在深刻重塑人类社会的传播机制、思维范式与价值表达方式。推荐算法作为平台信息分发的核心机制,其价值中立性早已受到质疑——算法不是“黑箱”工具,而是编码社会规范与意识形态倾向的“技术制度”。在这一时代背景下,如何将主流意识形态科学而有效地嵌入算法分发机制,已成为维护国家意识形态安全、引导舆论生态正向发展的关键议题。我国的算法价值嵌入路径,正展现出有别于欧美“市场自由放任—事后纠偏”式治理的独特范式。在我国政治制度、舆论生态与社会组织能力的多维支撑下,推荐算法的伦理治理不应止步于“算法透明”或“技术可控”,而应走向更高层次的“价值共建”。“自下而上”的意识形态嵌入路径,提供了一种兼顾效率、公正与导向的可能方案。其基于真实用户行为形成反馈机制,以技术手段内嵌社会主义核心价值观,并通过制度性框架保证长期运行的稳定性与合法性。

参考文献

- [1] 陈昌凤,石泽.技术与价值的理性交往:人工智能时代信息传播—算法推荐中工具理性与价值理性的思考[J].新闻战线, 2017(17): 71.
- [2] 张志安,汤敏.论算法推荐对主流意识形态传播的影响[J].社会科学战线, 2018(10): 176.
- [3] 崔聪.人工智能时代思想政治教育的算法风险及其应对[J].思想理论教育, 2020(5): 76-81.
- [4] 谢俊,吴阳琴.算法推荐下网络主流意识形态面临的风险及防范策略[J].自然辩证法研究, 2023(10): 132-137.
- [5] 朱萌,张权,赵璐.中国式现代化视阈下主流意识形态嵌入算法的机制与路径研究[J].河南社会科学, 2025, 33(5): 40-46.
- [6] 王浩.算法技术塑造主流意识形态认同的路径研究[J].东南大学学报(哲学社会科学版), 2021(S1): 20-22.
- [7] 张林.算法推荐时代凝聚价值共识的现实难题与策略选择[J].思想理论教育, 2021(1): 86-92.
- [8] 赵爱霞,王岩.算法推荐与新时代主流意识形态建设[J].云南社会科学, 2021(3): 112-118. 第13页 共14页
- [9] TARLETON GILLESPIE, “The Relevance of Algorithms”, in Tarleton Gillespie, Pablo J.Boczkowski, Kirsten A.Foot, Media Technologies, Cambridge: The MIT Press, 2014: 167.
- [10] SILVIA MILANO; MARIAROSARIA TADDEO & LUCIANO FLORIDI.Recommender Systems and their Ethical Challenges [J].AI and Society.2020, Vol.35(No.4): 957-967.
- [11] WENDELL WALLACH and COLIN ALLEN: moral machines: teaching robots right from wrong Oxford University Press, 2009, 273 pp, ISBN: 978-0-19-537404-9