E-Commerce Letters 电子商务评论, 2026, 15(8), 182-189
Published Online August 2026 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2026.158863

Hans汉斯

AI驱动的电商个性化与伦理边界：算法透明度与数据隐私的平衡机制研究

左晶晶[1]，张晨瑜[2]

[1]上海理工大学管理学院，上海
[2]上海理工大学中英国际学院，上海

收稿日期：2026年6月16日；录用日期：2026年6月30日；发布日期：2026年8月10日

摘　要

随着人工智能技术在电子商务领域的深度渗透，个性化推荐系统已成为平台提升用户体验与商业转化率的核心工具。然而，算法透明度与数据隐私保护之间的结构性张力构成了当前数字商业生态面临的重大伦理挑战。本文基于隐私计算、可解释人工智能以及数据保护法律框架的理论视角，系统分析了个性化服务中透明度与隐私的博弈关系，探讨了差分隐私、联邦学习等隐私增强技术在推荐系统中的应用机制，并提出了“分层透明”与“动态同意”相结合的技术－制度协同治理框架。研究发现，算法透明度与数据隐私并非零和博弈，通过隐私保护下的可解释推荐设计，可以在保障用户隐私权益的同时实现适度的算法透明，从而构建可持续的人机信任关系。

关键词

人工智能，个性化推荐，算法透明度，数据隐私，伦理治理

The Balance Mechanism between Algorithmic Transparency and Data Privacy in AI-Driven E-Commerce Personalization: An Ethical Boundary Perspective

Jingjing Zuo[1], Chenyu Zhang[2]

[1]Business School, University of Shanghai for Science and Technology, Shanghai
[2]Sino-British College, University of Shanghai for Science and Technology, Shanghai

Received: June 16, 2026; accepted: June 30, 2026; published: August 10, 2026

文章引用: 左晶晶, 张晨瑜. AI 驱动的电商个性化与伦理边界: 算法透明度与数据隐私的平衡机制研究[J]. 电子商务评论, 2026, 15(8): 182-189. DOI: 10.12677/ecl.2026.158863

Abstract

With the deep penetration of artificial intelligence technology into the field of e-commerce, personalized recommendation systems have become a core tool for platforms to improve user experience and business conversion rates. However, the structural tension between algorithm transparency and data privacy protection has become a major ethical challenge facing the current digital business ecosystem. Based on the theoretical perspectives of privacy computing, explainable artificial intelligence, and data protection legal frameworks, this paper systematically analyzes the game relationship between transparency and privacy in personalized services, explores the application mechanisms of privacy-enhancing technologies such as differential privacy and federated learning in recommendation systems, and proposes a coordinated governance framework that combines "layered transparency" with "dynamic consent" in both technology and institutions. The study finds that algorithm transparency and data privacy are not a zero-sum game. Through the design of explainable recommendations under privacy protection, it is possible to achieve a proper level of algorithm transparency while protecting users' privacy rights, thereby building a sustainable relationship of human-machine trust.

Keywords

Artificial Intelligence, Personalized Recommendation, Algorithm Transparency, Data Privacy, Ethical Governance

1. 引言

1.1. 研究背景与问题提出

人工智能技术的迅猛发展正在重塑电子商务的竞争格局。超过百分之八十的电商平台已部署基于深度学习的个性化推荐系统，通过分析用户历史行为、偏好特征与社交网络数据，实现从“人找货”到“货找人”的商业模式转变。然而，个性化服务的精准化程度与对用户数据的依赖程度呈正相关，这种数据密集型技术架构引发了严峻的隐私风险：用户画像精细化构建使个人身份可被重新识别，行为预测深度分析可能导致隐私泄露，而算法“黑箱”特性更使用户无法知晓其数据如何被使用、决策如何被生成。这种技术逻辑与隐私权益之间的冲突，构成了“个性化–隐私悖论”。Awad & Krishnan (2006)系统阐述这一悖论，指出消费者在享受个性化服务的同时，对隐私泄露的担忧构成内在张力[1]。

与此同时，算法透明度问题日益凸显。深度学习模型虽在推荐准确性上表现卓越，但其复杂非线性结构使决策过程难以被人类理解。欧盟《通用数据保护条例》[1]第 22 条关于自动化决策的规定，以及第 13~14 条关于“有意义的逻辑信息”的披露要求，正是对这一问题的法律回应。然而，学术界对《通用数据保护条例》是否确立“解释权”存在激烈争议：汪庆华(2020)指出，算法透明不是单一概念，而是涵盖技术透明、程序透明与问责透明等多个层面[2]；而张凌寒(2018)认为，在商业自动化决策场景中，法律应当赋予相对人独立的算法解释权，以回应自动化决策下的信息不对称与缺乏救济的问题[3]。这种理论分

[1]https://secnotes.cn/grc/ai/gdpr/

歧反映了透明度与隐私之间的深层张力——过度解释可能暴露模型细节，从而增加隐私泄露风险。

在此背景下，核心问题浮现：如何在保障用户数据隐私的前提下实现算法的适度透明？现有文献多将透明度与隐私视为独立治理维度分别研究，缺乏对二者内在关联机制的系统性探讨。本文旨在填补这一理论空白，构建整合技术路径与法律框架的分析模型，为电子商务领域的负责任人工智能应用提供指导。

1.2. 研究意义与结构安排

本研究的理论意义在于突破传统的“隐私–透明”二元对立思维，提出二者可以协同实现的理论命题。Duralia 等(2025)通过对 1037 篇文献的计量分析，识别出“人工智能与人性化”、“算法与平台动态”、“隐私权衡的行为反应”等主题聚类，为本研究提供了重要文献基础[4]。

本文采用文献综述与理论建构相结合的研究方法。全文结构如下：第二部分阐述核心概念与理论基础；第三部分分析透明度与隐私的张力机制；第四部分探讨技术融合的平衡路径；第五部分提出制度设计的优化建议；第六部分总结全文。

2. 理论基础与概念框架

2.1. 算法透明度的多维内涵

算法透明度是一个多维度概念。在计算机科学语境下，透明度指模型的可解释性，即人类能够理解模型从输入到输出的映射逻辑。从治理视角看，算法透明度包含三个层面：技术透明关注模型内部决策机制；程序透明涉及算法设计与部署过程的公开性；治理透明延伸至组织层面的问责机制。

不同层面的透明度对隐私影响存在差异。技术透明往往需要披露模型内部参数或决策边界，可能泄露训练数据中的敏感信息；程序透明主要涉及流程公开，隐私风险相对较低；治理透明则与隐私保护呈正相关关系。因此，追求透明度需区分不同层次，采取差异化披露策略。

2.2. 数据隐私的保护维度

数据隐私在电子商务语境下具有技术–法律双重内涵。从技术视角看，隐私保护核心在于防止敏感属性的推断攻击与成员身份的重识别风险。推荐系统面临的隐私威胁主要包括：基于用户评分历史的敏感属性推断、通过模型参数反演原始数据的模型反演攻击，以及通过观察推荐输出判断特定用户是否存在于训练集中的成员推断攻击。

从法律视角看，《通用数据保护条例》确立了目的限制、数据最小化、存储限制等基本原则。汪庆华(2020)与张凌寒(2018)关于算法透明与解释权功能边界的不同理解，反映了透明度与隐私保护之间的深层张力：解释义务虽有助于缓解“算法黑箱”问题，但若解释范围过度扩张，也可能因披露过多决策逻辑与处理细节而增加隐私泄露风险[2] [3]。

2.3. 个性化–隐私悖论的理论阐释

个性化–隐私悖论描述了消费者对个性化服务的矛盾心理：一方面期望获得精准推荐以节省决策成本，另一方面担忧为获得精准性而必须让渡的隐私权益。Awad & Krishnan (2006)系统阐述这一悖论。该悖论的成因在于个性化价值与隐私风险感知的不对称性[1]：个性化收益即时可见，而隐私成本延迟隐性。破解这一悖论需要重构用户–平台关系，通过透明度提升降低信息不对称，通过隐私增强技术降低风险感知，最终建立基于信任的价值交换机制。

2.4. 隐私增强技术的理论框架

隐私增强技术为破解透明度与隐私的张力提供了技术路径。差分隐私作为隐私保护的黄金标准，通

过向数据或模型输出添加校准噪声，确保个体记录的存在与否不会显著改变查询结果。Dwork & Roth (2014)系统阐述了差分隐私的算法基础[5]。联邦学习通过“数据不动模型动”的范式，将模型训练下放到用户本地设备，仅交换模型参数而非原始数据。张芳等(2025)提出了基于差分隐私的自适应个性化联邦学习方法。同态加密允许在加密数据上直接进行计算，为“端到端隐私保护下的透明推荐”提供了理论基础[6]。

这些技术的共同特征在于：在不暴露原始数据的前提下实现计算与推理，从而为“隐私保护下的透明度”创造可能性。然而，差分隐私引入的噪声可能降低推荐准确性并加剧流行度偏见；联邦学习的分布式架构增加了系统复杂性；同态加密的计算开销目前仍难以支撑大规模实时推荐。

3. 透明度与隐私的张力机制分析

3.1. 张力来源：技术层面的内在冲突

算法透明度与数据隐私之间的张力首先源于技术层面的内在冲突。推荐系统的核心算法，尤其是基于矩阵分解与深度神经网络的模型，其预测准确性依赖于对大量用户行为数据的学习。当平台试图提升透明度，向用户解释“为何推荐 A 而非 B”时，可能需要披露模型的权重参数、特征重要性或邻居用户的偏好信息。然而，这些正是攻击者用以推断敏感属性或重构原始数据的关键线索。

冯晗等(2023)系统梳理了推荐系统隐私保护中的性能权衡问题，指出隐私注入机制在降低泄露风险的同时，往往会对推荐精度产生影响，并可能进一步引发推荐偏差问题[7]。为保护用户隐私而引入的随机噪声虽降低了隐私泄露风险，却同时损害了推荐准确性，并加剧了流行度偏见——即推荐结果更倾向于热门商品，而忽视长尾用户的独特偏好。更重要的是，当尝试通过可解释技术增强透明度时，模型对训练数据的“记忆”可能被暴露，从而增加成员推断攻击的成功率。这种技术冲突的本质在于：透明度要求信息的流动与共享，而隐私保护要求信息的限制与隐藏。

3.2. 张力的制度放大效应

法律与制度因素进一步放大了技术层面的张力。《通用数据保护条例》的合规要求对电商平台施加了双重压力：一方面，第 35 条要求的数据保护影响评估需要详细记录数据处理活动，这实际上要求内部透明度；另一方面，对外披露的逻辑信息可能帮助攻击者重构隐私保护机制。张恩典(2021)指出，算法影响评估不应停留于形式化的告知与备案，而应成为贯通风险识别、解释说明与责任分配的治理工具，若仅将透明度理解为单纯的信息披露义务，便难以真正回应算法应用中的隐私与问责风险[8]。

此外，不同司法辖区的监管差异造成了合规复杂性。欧盟强调严格的数据保护与算法问责，美国倾向于行业自律与事后救济，而中国则在数据安全与算法推荐管理规定中提出了备案与透明度要求。平台在全球化运营中面临“合规拼图”困境。

3.3. 张力的社会心理维度

从用户感知角度，透明度与隐私的张力呈现更为复杂的形态。研究表明，用户对算法透明度的需求并非越高越好，而是存在“适度透明”区间。过度详细的解释可能引发“算法厌恶”，当用户了解到推荐是基于对其行为的深度追踪时，可能产生被监视的不适感，从而降低对系统的信任。信任在这一过程中扮演关键的中介角色：当用户信任平台会负责任地使用其数据时，他们更愿意接受个性化服务，也更宽容推荐的不准确性。

4. 技术融合：隐私保护下的可解释推荐

4.1. 差分隐私与可解释性的协同

差分隐私与可解释人工智能的协同是当前研究的前沿领域。高广尚(2024)系统梳理了可解释推荐模

型中的主要可解释性方法，指出可解释性不仅关系到推荐结果的说服力与用户体验，也是提升推荐系统透明度与可信赖性的关键路径[9]。研究发现，隐私保护机制在保护用户数据的同时，可能削弱解释方法的准确性，因为添加的噪声干扰了特征重要性的精确计算。

一种解决思路是“隐私保护下的局部解释”。通过将差分隐私应用于局部近似模型(如 LIME 的扰动样本生成)，可以在保护训练数据隐私的同时提供对个体预测的解释。陈彩华等(2024)指出，可信模型的可解释性与隐私保护之间存在矛盾关系，即提升模型透明度需要披露更多决策逻辑，而这恰恰可能增加隐私泄露的风险[10]。其核心洞见在于：解释本身可能泄露隐私，因此需要对解释过程施加隐私约束。

另一种思路是“可解释的隐私预算分配”。Müllner 等(2024)的 ReuseKNN 方法通过识别“高重用性邻居”，仅对这部分高隐私风险用户应用差分隐私，而大部分“安全用户”无需噪声保护，从而在隐私保护与推荐准确性之间取得更好平衡[11]。

4.2. 联邦学习中的透明机制

联邦学习中的透明性问题更为复杂：由于训练过程分布在多个节点，全局模型的决策逻辑难以追溯至特定用户的数据贡献。卞欣雨等(2026)提出了面向异质性的自适应差分隐私联邦学习方法，通过将差分隐私机制嵌入联邦学习过程，在保护本地数据的同时实现了全局模型的协同优化[12]。

针对联邦学习的可解释性挑战，研究者提出了“联邦可解释性”概念。在横向联邦学习场景中，可以通过聚合本地模型的解释生成全局解释，同时利用安全多方计算保护个体贡献的隐私。Gu 等(2024)的研究表明，隐私、准确性与模型公平性之间存在复杂的权衡关系，差分隐私的应用可能改善群体公平性，但严格的隐私保护也可能加剧歧视[13]。解决这一矛盾需要引入“零知识证明”等密码学工具，使审计者能够验证模型属性的合规性，而无需访问原始数据。

4.3. 同态加密与透明计算

同态加密允许在加密数据上直接进行计算，计算结果解密后与明文计算结果一致。这一技术为“端到端隐私保护下的透明推荐”提供了理论基础。然而，全同态加密的计算开销目前仍难以支撑大规模实时推荐。冯晗等(2023)指出，同态加密虽能提供高强度的隐私保护，但其计算开销和通信复杂度仍是制约其在大规模推荐场景中部署的主要瓶颈[7]。

从透明度角度看，同态加密的引入改变了信任模型：用户不再需要信任平台不会滥用其数据，而是信任密码学协议的安全性。这种“技术信任”替代“机构信任”的范式转移，对电商平台的治理架构提出了新要求。

5. 制度设计：分层透明与动态同意

5.1. 分层透明框架的构建

基于前述技术分析，本文提出“分层透明”的治理框架，将算法透明度划分为数据层、模型层、解释层与治理层，针对不同层级实施差异化的透明度与隐私保护策略。

在数据层，透明度主要体现为数据收集范围与使用目的的告知。此处应遵循数据最小化原则，仅收集实现特定推荐功能所必需的数据，并通过隐私增强技术在数据源头进行保护。用户应能清晰了解哪些数据被收集、用于何种目的、保留多长时间，但无需接触原始数据的处理细节。

模型层涉及算法架构与训练逻辑的透明度。对于商业敏感的核心模型参数，平台可保留一定程度的保密性，但应披露模型的类型、训练数据的规模与来源、以及主要的性能指标。对于高风险场景，则应要求更高的模型透明度，甚至接受外部审计。

解释层是用户直接交互的界面，应提供个性化、情境化的解释。根据用户偏好与算法素养，提供不同深度的解释选项：对普通用户，采用“基于受益”的解释；对高素养用户，提供“基于逻辑”的解释。所有解释生成过程应纳入隐私保护约束，避免泄露其他用户数据。

治理层关注问责机制与救济途径。平台应建立算法影响评估制度，定期评估推荐系统对隐私、公平性、多样性的影响，并公开评估报告。当用户质疑推荐结果时，应提供人工复核渠道。

5.2. 动态同意机制的设计

传统的“一揽子同意”模式已难以适应个性化服务的动态性与隐私保护的精细化要求。本文倡导“动态同意”机制，使用户能够在服务使用过程中持续调整其隐私偏好与透明度需求。

动态同意的技术基础是“隐私仪表盘”，为用户提供数据使用的实时可视化，包括：数据被访问的频率、推荐算法利用的数据类型、个性化带来的收益。基于这些信息，用户可以动态调整隐私设置，如选择“高隐私–低个性化”、“平衡模式”或“低隐私–高个性化”等预设配置。

动态同意还应包含“解释选择权”：用户可以选择接收不同详细程度的推荐解释，并知晓每种解释级别对应的隐私影响。从法律合规角度，动态同意符合《通用数据保护条例》关于“易于理解的同意”要求，也响应了《中华人民共和国个人信息保护法》[2]中关于“撤回同意”的规定。

5.3. 技术–制度的协同治理

技术工具的有效应用离不开制度环境的支撑。本文提出以下协同治理建议：

首先，建立“隐私设计”的组织流程。电商平台应在推荐系统的全生命周期中嵌入隐私与透明度考量：在需求阶段进行隐私影响评估；在设计阶段选择隐私增强技术架构；在测试阶段验证隐私保护的有效性与透明度机制的可用性；在运维阶段持续监控隐私指标与用户反馈。

其次，推行“算法透明度报告”制度。平台应定期发布透明度报告，披露推荐系统的整体运行状况，包括：用户数据的收集与使用情况、算法性能指标、隐私保护措施的落实情况、用户投诉与救济处理统计。

再次，构建多方参与的治理生态。算法治理不应仅依赖平台自我规制，而应纳入用户代表、学术机构、监管部门、行业协会等多元主体。可建立“算法伦理委员会”，对重大算法更新进行伦理审查；开展“算法审计”试点，引入第三方机构对推荐系统的隐私保护与透明度进行独立评估。

最后，加强用户算法素养教育。平台应通过用户教育计划，帮助用户理解个性化推荐的工作原理、数据的价值与风险、以及可使用的隐私控制工具。

6. 结论与展望

6.1. 主要结论

本文围绕“人工智能驱动的个性化与伦理边界”这一主题，深入探讨了算法透明度与数据隐私的平衡机制。主要结论如下：

第一，算法透明度与数据隐私之间存在结构性张力，但这种张力并非不可调和。技术层面的冲突可以通过隐私增强技术与可解释人工智能的融合创新得到缓解。差分隐私、联邦学习、同态加密等技术为“隐私保护下的透明”提供了技术可能性。

第二，透明度与隐私的平衡需要分层、分类的治理策略。不同层面的透明度对隐私的影响不同，应实施差异化的披露要求。对于高风险场景，应提高透明度与隐私保护的双重标准；对于一般商业推荐，

[2]http://www.npc.gov.cn/npc/c2/c30834/202108/t20210820_313088.html

可在保障基本隐私权益的前提下适度放宽透明度要求。

第三，用户信任是连接透明度与隐私的关键中介。通过“适度透明”与“动态同意”机制，可以在用户控制感与平台商业利益之间找到平衡点，构建可持续的人机信任关系。

第四，技术解决方案需要制度环境的支撑。分层透明框架与动态同意机制的有效运行，依赖于“隐私设计”的组织流程、透明度报告制度、多方参与的治理生态以及用户算法素养的提升。

6.2. 理论贡献与实践启示

本研究的理论贡献主要体现在三个方面：首先，突破了传统的“隐私–透明”二元对立思维，提出了二者协同实现的理论命题。其次，构建了整合隐私计算、可解释人工智能与法律框架的分析模型。最后，提出了“分层透明”与“动态同意”的治理框架。

对于电子商务平台的实践者，本研究提出以下建议：在技术架构层面，积极探索联邦学习、差分隐私等隐私增强技术的应用；在用户界面层面，设计分层、个性化的解释机制；在组织管理层面，建立算法伦理审查流程；在生态构建层面，参与行业标准的制定，推动负责任人工智能的集体行动。

此外，对管理教育的启示同样不容忽视。算法透明度与数据隐私的平衡能力，应成为未来电商管理者的核心素养之一。建议在管理类课程中嵌入算法伦理模块，通过情境化案例(如隐私泄露事件复盘)、可解释推荐系统的模拟操作等教学设计，使学生直观理解隐私增强技术的价值与局限。这种教育实践既回应了数字时代对“科技向善”人才的迫切需求，也为本研究所提出的治理框架培育了认同基础与执行土壤。

6.3. 研究局限与未来展望

首先，本研究主要基于文献综述与理论建构，未纳入实证数据检验所提出框架的有效性。未来研究可通过跨文化问卷调查、实验研究或案例研究，验证分层透明与动态同意机制对用户信任、满意度与行为意向的影响。

其次，本文对“透明度”的内部类型划分尚不够细致，“分层透明”框架与既有研究相比的创新边界仍有进一步阐明的空间。未来研究应进一步区分技术透明、程序透明、解释透明与治理透明等不同类型，并通过与现有文献的系统比较，更准确地揭示“分层透明”框架的理论贡献。

再次，随着生成式人工智能与大语言模型在推荐系统中的应用，透明度与隐私问题将呈现新的特征——大模型的“涌现能力”使得解释更为困难，而其对海量数据的记忆能力则加剧了隐私风险。如何应对这些新挑战，将是未来研究的重要方向。另一个值得关注的方向是跨文化比较研究：不同文化背景下的用户对透明度与隐私的偏好存在差异，电商平台在全球化运营中需要考虑这些文化维度，设计本土化的伦理治理策略。

最后，本研究框架的实践有效性，还可通过教育干预实验加以验证——例如，在管理类课程中引入“分层透明”与“动态同意”的教学模块，测量学生在伦理决策能力上的前后变化，这既能为治理理论提供实证支撑，也能推动算法伦理教育学的学科交叉发展。

基金项目

“上海理工大学本科课程思政示范课程建设项目”专项资助。

“上海理工大学本研一体化课程建设项目”专项资助。

参考文献

[1] Awad, N.F. and Krishnan, M.S. (2006) The Personalization Privacy Paradox: An Empirical Evaluation of Information Transparency and the Willingness to Be Profiled Online for Personalization1. *MIS Quarterly*, **30**, 13-28.

https://doi.org/10.2307/25148715 .

[2] 汪庆华. 算法透明的多重维度和算法问责[J]. 比较法研究, 2020(6): 163-173.

[3] 张凌寒. 商业自动化决策的算法解释权研究[J]. 法律科学(西北政法大学学报), 2018, 36(3): 65-74.

[4] Duralia, O., Ogrean, C., Ţichindelean, M. and Ţichindelean, M. (2025) Decoding the Personalization-Privacy Paradox: From Thematic Scholarly Clusters to Practical Insights. *Studies in Business and Economics*, **20**, 70-97. https://doi.org/10.2478/sbe-2025-0025

[5] Dwork, C. and Roth, A. (2014) The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, **9**, 211-487. https://doi.org/10.1561/0400000042

[6] 张芳, 龙士工, 郭晟南, 刘光源, 张珺铭. 基于差分隐私的自适应个性化联邦学习方法[J]. 计算机工程与设计, 2025, 46(9): 2525-2532.

[7] 冯晗, 伊华伟, 李晓会, 等. 推荐系统的隐私保护研究综述[J]. 计算机科学与探索, 2023, 17(8): 1814-1832.

[8] 张恩典. 算法影响评估制度的反思与建构[J]. 电子政务, 2021(11): 57-68.

[9] 高广尚. 可解释推荐模型中的可解释性方法研究综述[J]. 数据分析与知识发现, 2024, 8(Z1): 6-19.

[10] 陈彩华, 余程熙, 王庆阳. 可信机器学习综述[J]. 工业工程, 2024, 27(2): 14-26.

[11] Müllner, P., Lex, E., Schedl, M. and Kowald, D. (2024) The Impact Of differential Privacy On recommendation Accuracy And popularity Bias. In: Goharian, N., *et al*., Eds., *Lecture Notes in Computer Science*, Springer, 466-482. https://doi.org/10.1007/978-3-031-56066-8_33

[12] 卞欣雨, 胡小明, 王安保, 等. 面向异质性的自适应差分隐私联邦学习方法[J]. 计算机系统应用, 2026, 35(5): 74-83.

[13] Gu, X., Zhu, T., Li, J., Zhang, T., Ren, W. and Choo, K.K.R. (2024) Privacy, Accuracy, and Model Fairness Trade-Offs in Federated Learning. *Computers & Security*, **122**, Article 102907.