arXiv:2303.15035v1 [cs.DM] 27 Mar 2023

几行代码能改变社会吗？超越事实核查与内容审核：推荐系统如何毒化社交网站

David Chavalarias[1, a, b, *]、Paul Bouchaud[a, b, *]和 Maziyar Panahi[a]

摘要 近年来，网络敌意和极化现象均有所加剧，因此我们需要超越社交网络服务（SNS）上的事实核查本能或对改善内容审核的赞誉，研究这些平台对社会结构和社会凝聚力的影响。尤其是，Facebook 或 Twitter 等数字平台大规模部署的推荐系统所发挥的作用一直被忽视。本文借鉴认知科学、数字媒体和意见动力学领域的文献，提出一种忠实的模型，用以复现 Twitter 等社交网络服务中推荐系统、意见动力学与用户认知偏见之间的相互交织；该模型以政治活动人士推文的大规模纵向数据库为基础进行校准。借助这一模型，我们得以比较不同推荐算法对社会结构的影响，并量化它们与若干主要认知偏见之间的相互作用。特别是，我们证明，旨在单纯最大化用户互动的推荐系统必然会导致用户过度接触负面内容（某些系统中高达 300%），这一现象被称为算法性负面偏见；还会导致意见格局极化，并使社会权力集中于最有害用户手中。在最有影响力的 1% 用户中，后者的人数是总体人口中相应比例的两倍多。总的来说，我们的研究结果凸显出迫切需要识别推荐系统对个人和社会有害的实施方式，以便更好地规范其在系统性社交网络平台上的部署。

重要性声明 2021年1月6日对民主制度造成了冲击。一切迹象都表明，这并非昙花一现，而社交网络也发挥了作用。然而，要遏制全球范围内公共舆论持续不断的极化，我们需要超越“事实核查”和对有害内容的内容审核。本文研究了自学习推荐系统在 Facebook 或 Twitter 等系统性平台中的作用，以及它们与用户认知偏差之间的相互作用。我们表明，这些系统目前最可能的实现方式必然 会给个人和社会带来有害后果。除非大型科技公司能够证明情况并非如此，否则这不是用户行为问题，而是技术问题。这意味着，系统性数字平台目前对社会凝聚力构成系统性风险。本文提出了循证监管的关键要点。

[a] 法国国家科学研究中心，巴黎法兰西岛复杂系统研究所（ISC-PIF），法国巴黎75013区国民街113号[b] 法国高等社会科学研究院，社会分析与数学中心（CAMS），法国巴黎75006区

* 这些作者对本研究作出了同等贡献[1] 通讯作者：
David Chavalarias

1 研究缘由

2018年1月，Facebook 宣布改变其 新闻推送，这是一种推荐系统，也是其22亿用户的主要信息源。其目标是优先推荐能够产生最多互动的内容：分享、评论、点赞等。遗憾的是，对于公共讨论而言，心理学研究表明，这类内容平均而言最为负面，这种现象被称为 负面偏见 （Rozin 和 Royzman，2001）。这一变化的影响很快便显现出来。根据泄露的 Facebook 内部文件（Hagey 和 Horwitz，2021；Zubrow，2021），此后用户之间的交流变得更加对立，错误信息也更加广泛。与此同时，由于该平台的影响，用户的政治极化程度有所加剧（Allcott 等，2020年）。这些变化如此激进且深刻，以至于记者和政党都感到必须“在 Facebook 上的传播中采取更负面的倾向，其后续影响是使他们采取更加极端的政策立场”。

在线讨论中的极化和敌意有所加剧，这一现象在其他平台上也已出现。例如，在 Twitter 上，自 2016 年起，用户的主页时间线默认由推荐系统驱动，法国政治信息中负面推文的比例从 2012 年的 31% 上升至 2022 年的 50% 以上（Mestre，2022）。研究还证明（Vosoughi 等人，2018），在该平台上，虚假信息的传播“显著更远、更快、更深入、更广泛”，同时具有最强的回音室效应（Gaumont 等人，2018），因而也导致更强的极化效应。

只需改动全球推荐系统中的几行代码，就能从根本上改变人际关系乃至整个社会吗？社交媒体推荐系统在多大程度上改变了全球在线公共辩论的结构和社会群体形成过程？在极化已知会从线上环境蔓延至线下的当下（Doherty 等人，2016），这些问题关乎民主制度能否健全运行，至关重要。此外，欧盟等国家和地区开始监管数字服务行业之际[1]，要制定循证政策，也需要对这些问题作出科学解答。

以往研究探讨了在线社交网站（SNSs）对社会的影响，例如推荐系统对在线社会群体形成的影响（Ramaciotti Morales& Cointet，2021；Santos 等，2021）、对在线社交网络极化的影响（Tokita 等，2021），或网络拓扑结构对意见动力学的影响（Baumann 等，2020年）。然而，文献鲜少探讨推荐系统对意见动力学与社会网络形成之间耦合关系的影响。

本文填补了这一空白，提出了一个方法论框架，兼顾个性化推荐系统、人类认知偏差、意见动力学和社交网络演化之间的相互交织。该框架使人们能够探索推荐系统的不同设计对社会结构造成的后果，并量化这些设计与若干主要认知偏差之间的相互作用。

作为案例研究，我们将这一框架应用于一个类似 Twitter 的社交网络模型。我们构建了最先进的意见动态模型，并利用一个包含 5 亿条政治推文的政治推文数据库，对该框架的不同组成部分进行经验校准和经验验证；这些推文发布于 2016 年至 2022 年期间。接下来，我们将阐述

1. 《数字服务法》（DSA）https ://ec.europa.eu/commission/presscorner/detail/en/ip_22_2545

推荐系统对社会的影响：比较四种设计各异的推荐系统，这些系统基于使用我们的 Twitter 数据校准的行为模型、意见模型和网络模型。最后，我们展望将这一方法拓展至其他类型的社交网络。

这一案例研究凸显了人类认知偏差以及新型数字环境的特征，在分裂观点空间并在很大程度上扭曲互联网用户对现实的认知的自我强化过程中所发挥的作用。

具体而言，我们证明，只要用户存在轻微的负面偏见，旨在仅仅最大化用户互动的推荐系统就会导致用户过度接触负面信息，这种现象被称为 算法性负面偏见 （Chavalarias，2022），并且与信息以中性方式传播的情形相比，会导致在线舆论空间出现更严重的意识形态碎片化。

框架描述

让我们对在线社交网络的通用属性进行建模，以研究与其某些关键特征相关的风格化现象。社交网络服务的具体特征因平台而异，但它们都具有一些共同的核心特征：

1. **[发布]** 在任何时候 t，用户 i 都可以发布消息 m_i^t；2. **[建立网络]** 每个用户 j 都可以订阅 i的信息传播网络 $\mathcal{N}_i^d$（t）（在某些社交网站上，订阅是开放的；在其他网站上，则需要得到 i的同意）。3. **[信息]** 每个用户 i 都可以阅读其所订阅账户集合 $\mathcal{N}_i^r$（t）发布的消息，并最终将其分享给自己的订阅者$\mathcal{N}_i^d$（t）。

社交网站用户之间的订阅网络可以表示为一个不断演化的有向网络 $\mathcal{N}$（t）$= \cup_i \{\mathcal{N}_i^d$（$t$）$\cup \mathcal{N}_i^r$（$t$）$\} = \{s_{ij}\}_{i,\ j}$当用户 j 订阅了 i的账户时，就存在一条边 s_{ij}（信息从 i 流向 j）。 $\mathcal{N}$是此类平台上的信息流通骨干网络。其演化通常会受到社交推荐系统的影响，该系统会向用户推荐新的“好友”。

由于每位用户的订阅数量通常相当高（例如 > 300 在 Facebook 上、> 700 在 Twitter 上），大多数社交网站都实现了内容推荐系统，帮助任何用户 i 在其社交邻域产生的消息中找到“最相关”的消息 $\mathcal{N}_i^r$（t）。在 Facebook、Twitter、YouTube、领英、Instagram 或 TikTok 等平台上，这些内容推荐系统以个性化新闻推送 $\mathcal{F}_i$ 的形式呈现，将“相关”消息汇总成信息流。它们构成了这些平台用户获取信息的主要来源（例如，在 YouTube 上，推荐器负责带来 70% 的观看时长[2]）。社交推荐和内容系统通过对信息总体流动以及社会关系形成过程施加的约束，塑造用户的意见。尽管其中大多数都是黑箱，但我们知道，这些推荐系统会根据用户的行为，按照某种非常通用的目标函数进行学习。

话虽如此，为了建模人类认知、推荐系统、用户观点与社交网络演化之间的互动，我们必须对以下三件事进行建模：

2. https ://www.cnet.com/tech/services-and-software/youtube-ces-2018-neal-mohan

— **[推荐系统]** 内容推荐系统对新闻推送中的消息进行排序的过程；— **[用户注意力、认知偏差与意见动力学]** 用户阅读和分享消息的动机、其潜在的认知偏差、用户意见，以及用户接触消息后意见的演化；— **[网络演化]** 用户决定订阅或取消订阅其他账户的方式，以及社交推荐系统在这一过程中的作用。

我们将把所有这些要素纳入一个离散时间模型，其中每个时间步对应用户之间一天的互动。这些要素中的每一个本身都是一个独立的研究领域，因此这里并不打算分别在每个维度上提出进展。相反，我们将针对每个要素采用现有的最先进模型，以便根据实证数据对其进行校准，并研究它们之间的相互作用。

内容推荐系统

内容推荐系统可以访问一组用户特征，例如每个用户的订阅者数量、他们订阅的账户列表、每条消息的分享数量等，以及一组消息特征，例如消息的分享数量或情感倾向，并在每个时间步 t 为每个用户 i 生成一份有序列表，列出账户从 $\mathcal{N}_i^r(t)$ 产生的所有消息，以供阅读。

用户

用户被描述为具有内部状态（其“意见”）、与环境进行某种交互的接口（例如，阅读消息），以及对环境采取的一系列行动（发布消息、分享消息、订阅某个用户的账户等）的实体。为简化起见，我们假设意见动力学完全由用户之间的互动驱动。我们将这种简化的用户表示称为“主体”。

用户的意见

我们基于非贝叶斯意见动力学建模的相关文献（综述见（Noorazar等，2020年）），并为每个智能体 i 在 t 赋予一个意见 o_i^t，该意见位于度量空间$\mathcal{O}$中，并赋予一个意见更新函数 μ_i： $\mathcal{O}^2 \to \mathcal{O}$ 该函数定义了智能体在阅读传达意见的消息后 o_i 改变其意见的倾向，所传达的意见属于 o_j 智能体 j。 $\mathcal{O}$，且 μ 将通过实证进行估计。

智能体规则 1（意见更新）：在时间 t 分享主体 j的消息后，主体 i 的意见根据 $o_i^{t+1} \leftarrow \mu_i\left(o_i^t, o_j^t\right)$ 更新。

用户的在线活动

在每个时间步 t，每个主体 i 发布 $n_i^p(t)$ 条新消息（假定这些消息完全反映其观点），并分享 $n_i^s(t)$ 条由其他主体撰写的已读消息。 $n_i^p(t)$ 和$n_i^s(t)$ 将通过实证进行估计。

主体规则 2（阅读消息）：在每个时间步，主体 i 会在其信息流中“滚动浏览”，并随机停下来仔细阅读一些消息。

阅读后，用户可能会与消息互动：

主体规则 3（与消息互动）：主体 i 从主体 j 处分享消息的概率（即在下一个时间步原样重新发布该消息）取决于观点差异 $|o_i^t - o_j^t|$。

在相关文献中，人们提出了不同的互动概率函数形式，因此，应根据所建模的观点空间类型，通过实证估计其确切形式。

用户认知偏差

从分析推荐系统影响的角度来看，许多认知偏差都值得研究。作为示例，我们将重点关注心理学中两个著名的偏差：前文提到的确认偏误和负面偏见（Epstein，2018；Knobloch-Westerwick 等，2017；Rozin & Royzman，2001）——即倾向于更加重视负面信息。这个示例旨在评估算法性负面偏见（Chavalarias，2022）的强度：即算法机制导致负面内容被大规模过度曝光。

为了量化算法性负面偏见的影响，我们为智能体发布的消息赋予效价，这些消息可以是“负面”或“正面/中性”。因此，我们为每个智能体分配 i 一个比例 ν_i^t，表示其发布的负面消息占比 t，以及一种倾向 Bn_i，表示其以优先方式与负面消息互动的倾向（负面偏见）。ν_i^t 和 Bn_i 将通过经验方法进行估计。随后，我们通过对规则2 进行变更，实现智能体的负面偏见：

主体规则 4（阅读具有某种效价的消息）：在每个时间步，主体 i 会在其信息流中“滚动浏览”，并随机停下来仔细阅读一些消息。停下来阅读负面消息的概率是阅读非负面消息的 Bn_i 倍。

上述定义的规则集使我们能够研究前述认知偏差与学习型推荐系统之间的反馈回路。一方面，推荐系统力求最大化用户参与度；另一方面，用户更有可能参与与其既有信念一致和/或具有负面性质的内容。因此，我们可以预期，随着推荐系统不断学习用户随时间产生的偏差，其自身也会变得越来越有偏。

网络演化

意见与交互网络在反馈回路中共同演化。人类互动的同质性表明，用户倾向于与和自己相似的人互动并建立关系（McPherson 等，2001），并与那些恰好分享与自己观点不一致内容的人断绝社会关系。此外，社交网络服务通常会通过社交推荐系统向用户推荐新的联系人，而这些系统大多基于结构相似性（例如，共同好友）（Tokita 等，2021）。

我们将在一个简约而现实的链接形成与剪枝模型中纳入这些因素。我们模拟初始化时的网络规格（连通性、智能体类型等）将通过实证确定。

链接抑制（重连规则 1）

智能体会为自己的订阅评分，以监测自己对保留这些订阅的兴趣。对于 s_{ji} 对 j的每项 i 订阅，i 对通过 s_{ji} 0 δ_{ij} $t+$ 接收到的内容的意见分歧 $\delta_{ij}(t) \overset{t}{\geq} 0$ 初始值为 1，并在每个时间步按照 $(\,) = \gamma \times \left(\delta_{ij}(\,) + n_{ij}^t |o_i^t - o_j^t|\right)$更新，其中 $\gamma < 1$ 为每日折扣因子，而 n_{ij}^t 是 i 在时间步 t 期间阅读的、由 j撰写或转发的消息数量。如果 $s_{ji}(t) = 1$ 且意见分歧 $\delta_{ij}(t) > \tau$，i 将取消对 j的订阅，即 $s_{ji}(t+1) = 0$。

$1/\gamma$ 是智能体演化的特征时间，难以通过经验进行估算。该值将任意设为一个合理值。τ 也将如此，其确定取决于对 γ的了解。我们已验证，结果并不取决于对这两个参数的精确了解。

链接形成（重连规则 2）

为保持通过实证测量得到的连通性，我们假设，当主体与一名立场不一致的用户断开一条边时，它会开始关注随机选择的一名二阶邻居（这是一种在社交网络服务中经常观察到的重连机制（Tokita等，2021））。

推荐系统的实例化：Twitter 示例

为了理解对特定推荐系统的选择与其对意见动力学和社交网络演化所产生的系统性影响之间的复杂关系，随后我们将上述框架应用于 Twitter 上政治意见动力学的建模。顺带一提，我们找到了可用于模拟其他社交网络服务推荐系统影响的现实参数值。

在本研究开展时，Twitter 的数据可用性、其广泛的使用程度——全球月活跃用户超过 3 亿——以及其在政治传播中的主导地位，构成了我们选择将其作为实验场来检验所提出框架的依据。此外，根据实证测量，Twitter 也是一种数字媒体，在其中负面内容比其他内容更具病毒式传播性（见图S18），且用户自身也倾向于生产负面内容（见图 S11）。这就提出了一个对公共讨论和用户福祉都至关重要的问题：这种负面性泛滥在多大程度上是由 Twitter 的算法架构造成的。

简而言之，Twitter 是一个于 2006 年推出的在线社交网络，用户可以在此公开交流 280 字符的消息，并将消息广播给他们的“关注者”，也就是订阅了作者账户的用户。内容会根据个性化推荐显示在用户的信息流中，这一信息流称为 主页时间线 。消息由机器学习算法排序，该算法会预测用户与推文互动的可能性。下文将聚焦于 Twitter 上互动的两种主要形式（Twitter，2020年）：（1）仔细阅读推文，通常需要点击才能展开

内容——（2）转发，即原样重新发布消息，并注明作者，不添加任何评论，也不作任何修改。

尽管社交影响远不止于转发，但实证研究发现，至少在政治语境下，与 Twitter 提及等方式相比，转发更能反映人们的意见并追踪其变化（Conover 等，2011；Garimella 等，2018）。事实上，仅凭转推数据，就能以很高的准确率预测政治活动人士的政治倾向（Gaumont 等，2018）。下文对我们框架的实证应用将聚焦于转发网络。

1.1 推荐器的选择

多起泄露事件以及官方公告表明，许多社交网站将用户互动最大化作为其推荐系统的目标函数。接下来，我们将分析此类目标函数对社会结构的影响。

推荐器规则 1：在每个时间步，推荐器都会根据其预测的消息被分享的概率，为每个主体 i 从 $\mathcal{N}_i^r(t-1)$ 中筛选并展示一部分消息。

由于其灵活性和效率，我们通过 XGBoost 算法实现了这一优化（Chen 和 Guestrin，2016）。

推荐系统通过依赖特定输入来实现其目标。因此，对这类算法进行建模时，应明确它能够获取哪类数据。商业推荐系统使用的输入数据种类属于商业机密，因此人们对实际使用哪些输入数据知之甚少。本文将选取两大类商业推荐器可能使用的数据（参见（Xu 和 Yang，2012）、（Huszár等，2022）：

— 情感分析：推文的负面或中性性质，以及用户过去转发的负面内容所占比例。— 受欢迎程度评估：推文作者的受欢迎程度，即其消息的平均转发次数、消息被转发的次数，以及用户转发该作者推文的频率。

为了探究不同输入特征带来的影响，我们将比较推荐系统的三种不同实现：

— *Neg*：仅使用情感分析的输入数据，— *Pop*：仅使用受欢迎度评分的输入数据，— *PopNeg*：使用*Neg* 和 *Pop* 算法的组合特征。

为了评估这三种推荐系统实现对社会结构的影响，我们将它们与一种中性推荐系统进行比较，即按时间倒序呈现内容，以下简称*Chrono*。由于简单易用，*Chrono* 常被称为非算法推荐。Twitter 曾短暂采用这种方式，直到埃隆·马斯克收购 Twitter；此后便无法再关闭推荐算法。

经验校准

本节中，我们使用有关法国政治的实证数据对模型进行全面校准。这些数据于 2021 年秋季在 Twitter 上收集，来自 *Politoscope* 项目（Gaumont 等，2018），该项目是一个用于研究 Twitter 上集体动力学的社会宏观镜。自 2016 年以来，*Politoscope*一直持续收集有关法国政治的政治推文，并支持在任意给定时段内筛选最活跃用户的子集。

用户互动网络

虽然可以通过 Twitter API 获取 Twitter 的关注者—被关注者关系图，但如果将这样的图用于我们的模型，就会产生误导。实际上，Twitter 上实际使用的内容推荐器已经经过充分训练，被关注者发布的内容可能永远不会展示给用户，从而扭曲我们的模拟。为规避这一限制，~我们转而考虑由转推和引用推文相结合形成的经验网络 $\mathcal{N}$。这样的网络确实似乎是平台实际向用户展示内容的合理代理。正如下文所述，考虑引用推文而不仅仅是转推，可以纳入意识形态上不一致的内容。我们的每次模拟都基于所选时间段内的~用户互动 $\mathcal{N}$ 经验网络进行初始化。

意见空间校准

下文中，我们将“意见”一词理解为政治舞台中的意识形态定位，并排除事实上属于政治不可知者的人士。由于并非所有候选人都采用相同的数字传播策略，以下内容仅纳入在所考察时期内于 Twitter 上具有显著影响力的领导人。

利用社交网络服务数据重建观点空间，过去数年来一直是一个十分活跃的研究领域；相关重建涵盖一维空间（Barberá，2015；Briatte &Gallic，2015）、二维空间（Chomel 等人，2022；Gaumont 等人，2018），甚至维度可变的空间（Reyero 等人，2021）。至于转推网络，研究已证明，反复转发某人的内容可以作为意识形态一致性的指标（Conover 等人，2011；Garimella 等人，2018；Gaumont等人，2018）。

通过对政治转发图进行聚类分析，Gaumont 等人（Gaumont 等人，2018）在意见分类上达到了95% 的准确率，验证了使用转发图开展此类研究的可行性[3]。2021年秋季 Politoscope 转发图的空间化呈现出一个多极圆形政治舞台（见图1），政治领导人的相对位置与公众所呈现的政治格局相符。正如补充信息中所述，我们利用这一空间化结果，将观点空间建模为一个圆形的一维度量空间 $\mathcal{O}$，其中 $o_i \in]-1,\ +1]$，从而能够用实证估计值初始化~智能体的意见 $\mathcal{N}$，计算推荐系统的建议对用户意见的影响，并确定不同推荐系统对 $\mathcal{O}$用户意见分布的整体影响。

3. 我们在自己的数据集上进行了同样的验证，并得到了相似的结果。

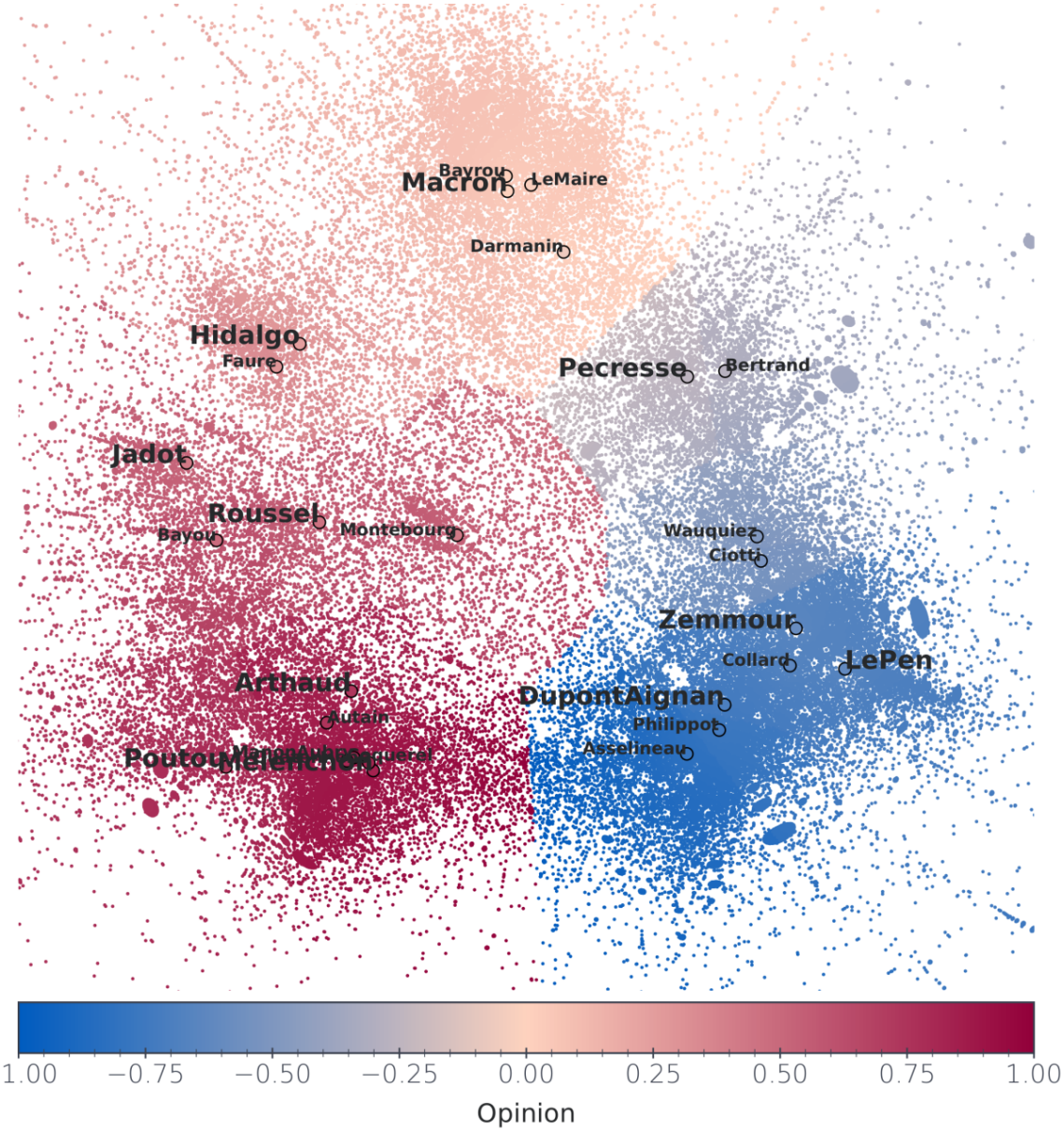

图1——2021年9月期间计算得到的法国选举前政治推特圈多极图。每个节点对应一名用户，并根据所述方法赋予的意见进行着色。图中突出显示了政治领导人，尤其是2022年法国总统选举的候选人。

智能体意见更新的校准

有了观点空间的度量空间，我们便可借鉴大量意见动力学研究（Deffuant等人，2000年；Jager与Amblard，2005；Noorazar 等人，2020年）。利用 Twitter 数据集中完整的用户互动历史，并为简化起见，假设 μ（意见更新函数）的函数形式对所有智能体都相同，我们通过符号回归确定了最可能的 μ。回归采用遗传算法（Fortin 等人，2012），在算术函数、三角函数以及对差异的实现构成的函数集合上进行

周期性意见空间。

这种经验校准使我们能够确定一个线性的意见更新函数$o_i^{t+1} \leftarrow o_i^t + \lambda_i (o_j^t - o_j^t)$，其中 $\lambda_i \in \mathbb{R}$。需要注意的是，这一函数早已广泛应用于意见动力学文献中（Deffuant等，2000年；Jager与Amblard，2005）。

由于我们缺乏有关推文曝光量的信息，并且考虑到我们的意见归因方法，我们决定简化模型，假设智能体只有在转发一条消息时才会改变其意见，即意识形态定位。

随后，我们为每个智能体拟合了意见更新参数 λ_i，其绝对值反映了该智能体的易受影响性，即在转发他人的消息时，其意见会在多大程度上发生改变；所使用的数据是用户实际转发的每日消息列表（见 SI**??**）。

这种拟合可带来相对较高的准确率：经过 30 次迭代（对应于 10 月底，参见图 **??**）后，我们最终拟合得到的意见中，有超过 75% 的意见与实际值相差小于 0.05。这一数值小于社区内意见多样性。我们应当强调，本研究的目标并非准确预测在线社交媒体用户的意见，而只是忠实地仿真在线用户的行为，以研究算法推荐所产生的后果。具体而言，用户的意见在仿真中用于确定某条内容的转推概率，因此意见相差 0.05 并不会改变仿真的行为。意见发生的唯一显著变化是较大的变化（$\Delta_{op} > 0.05$）；对于这些变化，拟合得到的更新规则在超过 60% 的预测中相对误差小于 25%，而对于特别大的位移，准确率甚至更高 $\Delta_{op} \in [0.5, 1]$（参见图 **??**）。为验证所用方法的合理性，我们考察了其他时间段和其他图空间化设置，并预测了拟合后一个月（11 月）的意见，所得准确率相近，详见补充文本。

代理人活动校准

由于缺乏明确说明哪些消息会显示在用户屏幕上的信息，我们假设用户会持续阅读消息，直到达到其每日转推数量，或读完来自 $\mathcal{N}_i^r\ (t-1)$ 的所有消息。我们确定了2021年秋季期间最活跃的 $110k$ 名用户，获取了他们发布的政治推文，并估计了其发布行为。每日发布的推文数量（$n_i^p\ (t)$，原创发布）和~转推（$n_i^s\ (t)$，共享发布）在个体层面上均服从指数分布（正如（Baumann等，2020年；Perra等，2012年）中所观察到的）。在群体~ ~层面，不同用户的经验指数尺度 θ_i^p 和 θ_i^s 按照图S1所示的分布进行分布。基于这些经验观察结果~，我们将主体 i 在 $\mathcal{N}$ 中的推文和转推数量设定为分别从具有经验确定速率 θ_i^p 和 θ_i^s 的两个~ ~指数分布中独立抽取。

接受度范围

一旦分配好意见，我们就确定了意见差异分布Δ_{op} 在用户与转发消息作者之间。为了消除平台呈现方式带来的偏差（Huszár等人，2022），同时考虑到

社群的规模和位置各不相同，因此我们必须结合引用推文的发布模式，对从转推中观察到的意见差异分布进行重新归一化（参见 S1.4）。

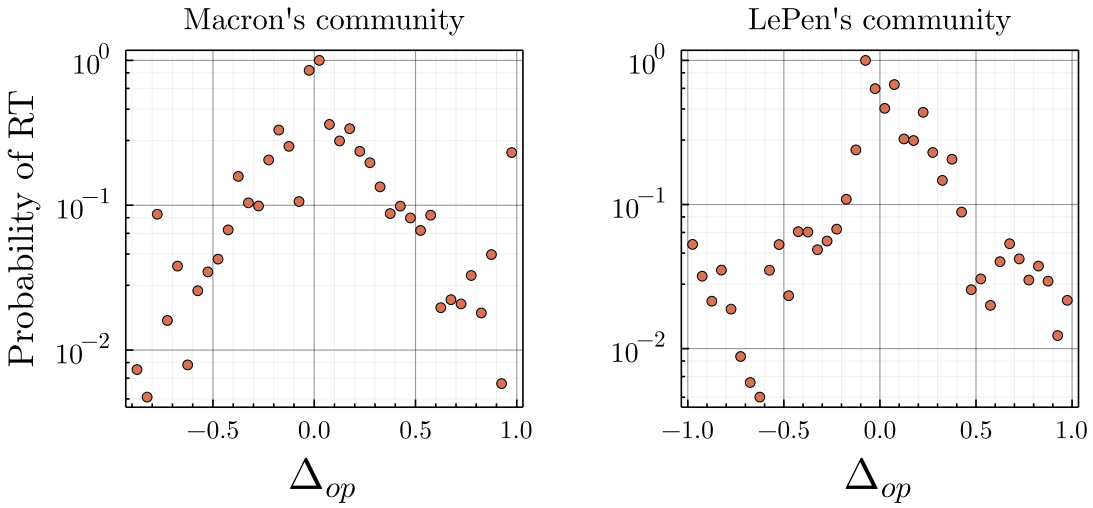

图 2——对于处于埃马纽埃尔·马克龙或玛丽娜·勒庞意识形态邻域中的用户，其转发一条已读消息的估计概率随该用户与消息作者之间的观点差异而变化，$\Delta_{op} = o_{reader} - o_{sender}$，其中考虑了周期性边界条件。我们进行了重新归一化，使得与用户观点完全一致的消息必然会被转发。

从图 2中可以看出，随着观点差异增大，转发一条消息的概率大致呈指数衰减，其中一些细微之处揭示了政治策略。分布的不对称性表明，埃马纽埃尔·马克龙的社区（中心）倾向于转发比自身立场更偏右的内容，而不是更偏左的内容；对于左翼领袖让-吕克·梅朗雄的社区，则观察到相反的效应（参见补充信息 S3）。在确定了覆盖全部意见范围的此类分布后，我们将与模拟主体初始意见相对应的分布分配给这些主体。

负面性

为校准模型中与负面性相关的属性，我们对 $190k$ 2021年10月期间由500 名不同用户发布的法国政治推文进行了情感倾向分析。该分析使用法语版的双向变换器编码器 CamemBERT（Martin等，2020年），并在法语推文上进行了微调。随后，我们根据智能体的初始意见，从经验分布中抽取并为其分配内在负面性 ν_i，即其发布的负面内容所占的比例；同时还为其分配了负面偏见，具体讨论见补充材料。

推荐器效果评估

为刻画我们基于主体的模型的行为，我们在此引入以下特别关注的指标：

算法性负面偏见 Γ： 这是推荐系统所造成的负面内容过度曝光，定义为感知环境（即时间线内容）中的负面性与“真实环境”中的负面性之比即一个人的入邻域 $\mathcal{N}_i^r$。

为进一步探究该模型，我们使用 Leiden 算法（Traag 等人，2019）对生成的转发图进行社区检测。该算法是对常用 Louvain 方法的改进，可确保社群连通。完成后，我们考察：

Newman 模块度 $\mathcal{Q}$（Leicht 和 Newman，2008）：评估社区内部连接的密度。

社区内多样性 σ_X^{intra}：某个聚类中某项可观测量（例如意见、内在负面性或感知负面性）的标准差，除以总体人口中该可观测量的标准差，再对各聚类取平均值。

社区间多样性：σ_X^{inter}：各聚类可观测量（例如平均意见、内在负面性或感知负面性）的标准差，除以整个群体的多样性，用以衡量不同社群相对于总体人口的多样性。

结果

使用经验网络评估推荐器对负面性和观点极化的影响

该模型使用真实数据进行初始化，除两个参数外，所有参数均经过经验校准；我们已知这两个参数不会影响结果。我们模拟了一个月的互动，以估计真实数据集中每个账户的活跃度和意见演化，并分析了上述指标。关于账户活跃度预测，由于我们的框架具有随机性，四种推荐系统均无法预测低强度互动（$\leq$ 10 次转发/月），相对于真实分布，它们高估了较小的权重（见图S1）。然而，对于较大的权重（$>$ 10次转发/月），*PopNeg* 能够忠实地匹配经验分布，而*Pop* 则高估了较大的权重，这正如预期；*Chrono* 则低估了这些权重。

如图3所示，对负面性的过度暴露 Γ 在时间顺序模式下并不存在，这符合预期；而三种算法推荐系统会使用户过度暴露于负面内容。仅基于负面性的*Neg* 推荐器导致最高程度的过度暴露：用户平均看到的负面内容比在中性Chrono模式下多*26%*。

在群体中，对负面性的过度暴露存在极大差异，如图 4所示；一些用户受到算法性负面偏见的影响，过度暴露程度超过 300%。即使是邻域规模较大的用户，以及没有任何负面偏见的用户，也会出现这种情况。对于好友数量较少（少于 10）的用户，我们发现入邻居数量与消极内容过度暴露之间存在微弱（$r = 0.02$）但显著（$p < 10^{-7}$）的相关性。事实上，随着好友数量增加，推荐系统可供挑选的消息池也会扩大

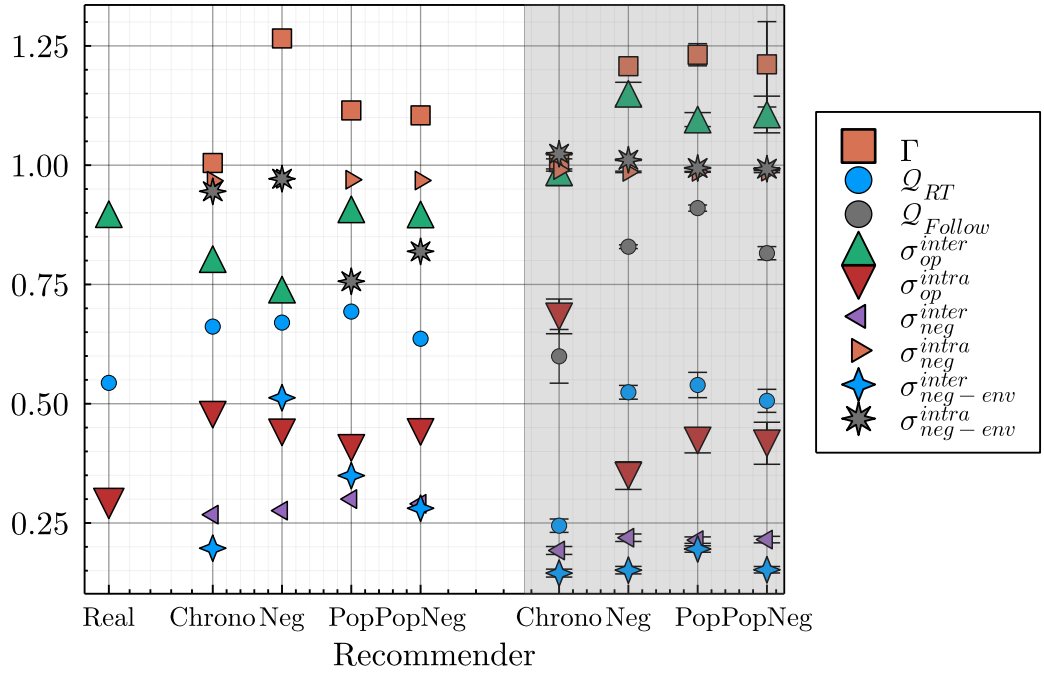

图3——四种推荐器与经验观测数据之间的指标比较。仿真分别在真实图（白色区域）和合成图（灰色区域）上初始化；后一种情况下，误差条表示从相同合成输入开始进行10次重复仿真所得结果的标准差

从而能够选出最能引发互动的消息（而这些消息大多数时候也是最负面的），进而导致更严重的消极内容过度暴露。这些结果直接源于人类消极偏见与推荐系统中硬编码的参与度最大化目标之间的反馈回路。

个体层面负面内容过度暴露程度的巨大差异值得注意，因为从个体角度来看，这可能会使用户陷入与现实脱节的有毒环境，并可能对其心理健康和社会关系造成不利影响；例如，新冠肺炎疫情期间就有相关记录（Levinsson等人，2021）。

意见多样性取决于推荐系统，这揭示了对网络社会健康的另一种有害影响。事实上，虽然*Chrono* 和*Neg* 导致相同的$\sigma_{op}^{inter/intra}$，但两种社会模式，即*Pop* 和*PopNeg*，会使社会结构更加碎片化。聚类内部的平均意见多样性 σ_{op}^{intra}较低——但没有经验观测到的那么低——而不同聚类围绕不同意见形成中心—— σ_{op}^{inter}更高，接近经验观测值。

相比之下，转发图的模块度 $\mathcal{Q}_{RT}$ 被发现与推荐系统无关，聚类$\sigma_{neg}^{inter/intra}$内部和彼此之间的负面性多样性也是如此；图结构在初始化时受到强约束

通过考察推荐器特征的重要性，我们注意到，用户与作者之间的过往互动频率是迄今为止信息量最大的特征，这再次说明了在人类确认偏误基础上、并受到推荐器强化的确认偏误。同样，不同聚

对合成网络中智能体负面偏见的敏感性分析还表明，算法性负面偏见现象（$\Gamma > 1$）会在智能体具有一定负面偏见后立即出现；其强度几乎不受智能体自身负面偏见强弱的影响（见图S16）。只要用户偏好负面内容而非中性内容，基于互动的推荐系统就必然会导致对负面性的过度暴露。

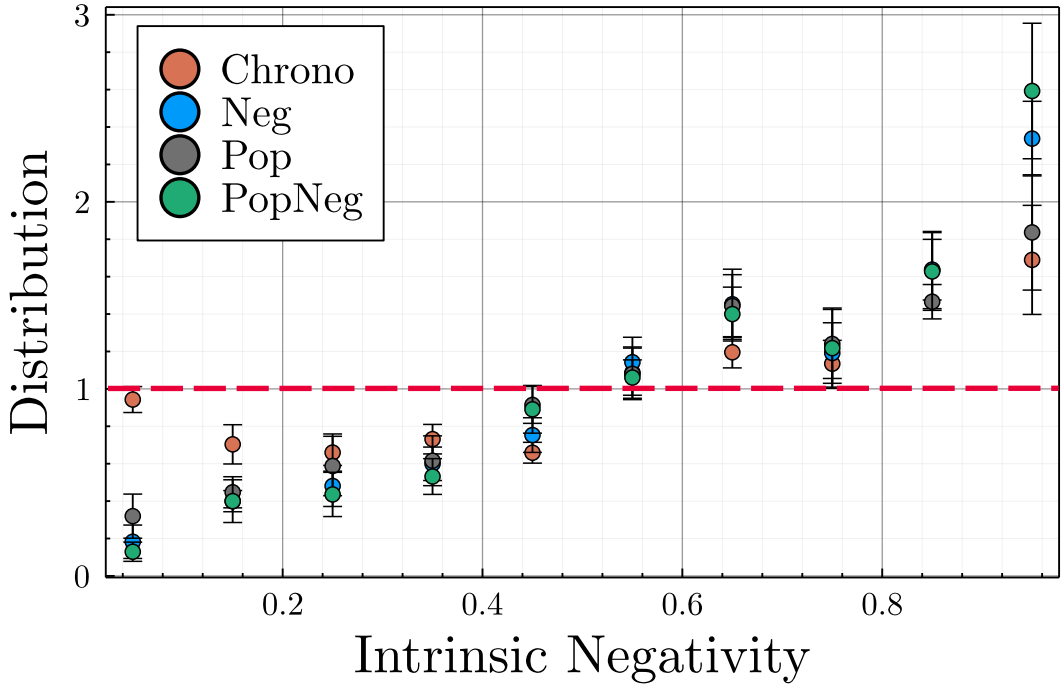

图5——与总体人口相比，在最受欢迎的1%智能体中，负面主体的代表性过高。分析在两个月的仿真后进行，误差条表示从相同的合成输入开始重复10次所得的标准差。

从一个不受约束的随机网络开始，$\mathcal{N}$ 可以充分展现推荐系统的作用，并得以检验所提出的网络演化模型是否符合经验观察（示例见图S12）。如图3所示，经过两个月的仿真演化，与中性内容呈现相比，算法推荐显著提高了转发网络和关注网络的模块度，这种情况在*Chrono*中亦然；意识形态碎片化或对负面性的过度暴露也有所增加。

算法性负面偏见不仅会使主体的信息环境变得更加有害，还会影响群体中社会权力的结构；社会权力定义为主体影响公共讨论的能力（Jia等，2015）。图S13展示了总体人口与按推文获得最多转发的前1% 主体之间内在负面性的失衡，而图S14则显示了在 Neg算法下，负面主体在最受欢迎分位群体中的比例。该分析清楚地表明，以互动为优化目标的推荐算法会放大个体的负面偏见，使在线社会权力集中到最负面的用户手中。

例如，发布负面内容占一半的主体在最受欢迎群体中的比例与其在总体人口中的比例相符；而在三种算法推荐系统下，完全不发布负面内容的用户都不会出现在最受欢迎群体中。令人担忧的是，持续发布负面内容的主体在最受欢迎群体中的数量是其在总体人口中数量的两倍多；其中，考虑消息负面性的两种推荐器，即 *Neg* 和 *PopNeg*，造成的过度代表最为显著。

同样值得注意的是，尽管 *Chrono* 在选择内容时保持中立，但由于个体的负面偏见，最终仍会导致明显的不公平呈现。即使处于中立模式，用户也更有可能阅读负面内容，并因此转发这些内容，从而提升其作者的受欢迎程度。

讨论

2021年1月6日，一群坚信选举被窃取的人群冲入了华盛顿哥伦比亚特区的国会大厦。无论这一事件在多大程度上可归因于关于选举过程的错误信息，有一点很明确：这并非一时风潮：1月6日一年后，“52%的特朗普选民以及41%的乔·拜登选民，在某种程度上同意或强烈同意，现在是时候将这个国家一分为二了”（《政治》，2021），而2020年末的一项调查得出结论：“美国人很少像今天这样两极分化”（Di-mock & Wike，2020年）。为了改善这种公众舆论极度极化的状况——这种状况往往也会在巴西、英国或意大利等其他国家重现——我们必须超越“事实核查”的惯性，以及对在线社交网络中更好地审核有害内容的推崇。

正如其他采用与我们互补的方法开展研究的学者所指出的（Ceylan等，2023；Törnberg，2022），我们必须承认社交网络服务对社会结构的影响，尤其是其在群体之间放大极化和敌意方面的作用。这不仅是影响公众的一种现象，整个信息生态系统也处于风险之中。Facebook 于 2018 年更改算法，以优先促进“有意义的社交互动”后，“公司研究人员发现，内容发布者和政党正在调整其帖子，使其更倾向于愤怒情绪和耸人听闻的内容”；一份内部备忘录还提到，“在转发内容中，错误信息、毒性内容和暴力内容的出现比例异常之高”（Hagey &Horwitz，2021）。

在各国政府都在考虑监管大型社交网站（社交网络服务）之际，建立模型来量化其对社会的影响比以往任何时候都更加必要。在本文中，借助将社交网络建模为复杂系统，并使用大数据对模型进行校准，我们可以一窥社交网站幕后究竟发生了什么。

通过使用一个涵盖政治活动人士推文的大规模纵向数据库（Gaumont等，2018），我们构建并校准了一个基于主体的模型，用于模拟用户在 Twitter 上的行为、他们的一些认知偏差，以及推荐系统影响下其政治观点的演变。此外，我们还通过实证估算了许多意见动力学模型共有的参数，而这些参数此前都是任意设定的——例如广泛使用的意见更新规则「主体规则」1。我们还超越了通常采用的假设，例如参与互动所需的意识形态分歧固定阈值，转而考虑互动发生的概率并估算其分布规律。

借助这一校准模型，我们得以比较各种推荐算法对社会结构的影响，并量化它们与某些

主要认知偏差的相互作用。我们尤其证明，旨在单纯最大化用户互动的推荐系统必然会使用户过度接触负面内容，这种现象被称为算法性负面偏见(Chavalarias，2022），并导致在线舆论格局出现意识形态碎片化和极化。

重要的是，推荐系统当前的实现方式会带来这些损害个人和社会的后果，但这些后果未必是有意为之，只是人类认知缺陷与社交网络服务的经济目标之间正反馈的结果。由于大多数此类平台规模庞大，已具有系统性，它们不受监管地追求利润，给用户和民主制度的健全性都带来了系统性社会风险。

政策制定者日益意识到社交网络服务当前的实现方式对民主制度构成的威胁，却缺乏监管这一领域的关键手段。采用跨学科方法，对社交网络服务及其对个人和社会群体的影响进行建模，可以提供其中一些关键手段。

例如，我们已经证明，当使用自学习算法推荐内容时，若以用户或内容受欢迎度的衡量指标作为输入，就会提高平台对个人及其所组成群体的整体毒性。这还会导致社会权力集中在最有毒的账户手中。这意味着，监管机构应关注输入推荐系统的数据类型，并考虑禁止使用其中某些类型的数据。这类发现有助于确定，在监管平台的过程中，何时何地必须放宽商业机密限制——平台被要求与公共机构合作时，往往会以商业机密为由拒绝配合。

另一方面，平台可以采取措施，在用户层面减轻负面偏见，防止其演变为算法性负面偏见并扩散至集体层面。

推荐算法影响研究是学术界一个新兴的领域，相关主管部门应予以支持，以便在完全独立的情况下确定适当的监管杠杆，并实施循证政策。毋庸置疑，这需要社交网络服务向学术界更加开放数据[5]。本研究在 Twitter 上进行的一些经验校准，例如意见更新规则或转发行为，可用于对 Facebook 等其他平台进行建模，但没有什么能比得上基于某个平台原生数据进行的经验校准

总之，仅仅将矛头指向生产有害内容的恶意用户并呼吁加强内容审核是不够的。我们需要进一步研究大型科技公司大规模部署推荐器系统在个人和集体层面的影响，并评估其潜在危害。《科学》应通过建模这些平台对社会结构的影响，为循证政策制定作出贡献。民主正处于危险之中。

致谢

我们感谢 Victor Chomel 富有成效的讨论，也感谢 Jeanne Bruneau–Bongard 仔细校对本文初稿。本研究获得了“Fondation CFM pour la Recherche”的资助，以及

5. 请注意，Twitter 一直是少数开放部分数据的大型社交网络服务之一，而在撰写本文时，Twitter 正考虑结束这种数据开放性。

巴黎法兰西岛复杂系统研究所和法兰西岛大区的支持。

参考文献

Allcott、Hunt、Braghieri、Luca、Eichmeyer、Sarah 和 Gentzkow、Matthew。2020年。社交媒体的福利效应。《美国经济评论》，**110**(3)，629–76。

Barberá，Pablo。2015年。物以类聚，人以群分：同类相聚推文相随：使用 Twitter数据进行贝叶斯理想点估计。《政治分析》，**23**(1)，76–91。

Bastian，Mathieu、Heymann、Sebastien 和 Jacomy、Mathieu。2009年。*Gephi*：用于探索和操作网络的开源软件。

Baumann, Fabian, Lorenz-Spreen, Philipp, Sokolov, Igor M., & Starnini, Michele. 2020年. 社交网络中的回音室与极化动力学建模。《物理评论快报》，124(4), 048301. arXiv：1906.12325 [物理学]。

Bouchaud, Paul, Chavalarias, David, & Maziyar,Panahi. 2023. 关于：*Chavalarias*、*Bouchaud*、*Panahi*（*2023*），《几行代码能改变社会吗？超越事实核查与内容审核：推荐系统如何毒化社交网站》的复现数据。

Briatte, François, & Gallic, Ewen. 2015（5月）。从 Twitter 数据还原法国政党空间。载于：*Science Po Quanti*。00000。

Ceylan, Gizem, Anderson, Ian A., & Wood, Wendy. 2023. 错误信息的分享是一种习惯行为，而不仅仅是懒惰或偏见。《美国国家科学院院刊》，120(4), e2216614120。出版方：《美国国家科学院院刊》。

Chavalarias,David. 2022.《有毒数据》——社交网络如何操纵我们的意见. 弗拉马里翁出版社版。

Chen, Tianqi, & Guestrin, Carlos. 2016. XGBoost：一种可扩展的树提升系统。载于：第 *22* 届 *ACM SIGKDD*国际知识发现与数据挖掘会议. ACM。

Chomel, Victor, Cuvelle-Magar, Nathanaël, Panahi, Maziyar, & Chavalarias, David. 2022（11 月）。利用时序图上的表示学习，以欧拉描述识别多个时间尺度上的极化。

Conover, Michael D., Goncalves, Bruno, Ratkiewicz, Jacob, Flammini, Alessandro, & Menczer, Filippo. 2011. 预测 Twitter 用户的政治立场。载于：*2011* 年 *IEEE* 第三届隐私、安全、风险与信任国际会议和 *2011* 年 *IEEE* 第三届社会计算国际会议。IEEE。

Deffuant, Guillaume, Neau, David, Amblard, Frederic, & Weisbuch, Gérard. 2000年. 相互作用的主体之间的信念混合。《*Adv* 复杂系统》，**03**(01n04)，87–98。

Dimock, Michael, & Wike,Richard. 2020（11月）。美国的政治分裂具有特殊性质。

Doherty, Carroll, Kiley, Jocelyn, & Jameson, Bridget. 2016（6月）。*2016*年的党派性与政治敌意。技术报告。皮尤研究中心。Epstein, Robert. 2018（4月）。搜索建议效应（SSE）：搜索建议如何被用于显著改变意见和投票偏好，且人们对此毫无认知。载于：西部心理学协会第*98*届年度会议。Fortin, Félix-Antoine, De Rainville, François-Michel, Gardner, Marc-André, Parizeau, Marc, & Gagné, Christian. 2012。DEAP：让进化算法变得简单。机器学习研究杂志，**13**（7月），2171–2175。Garimella, Kiran, Morales, Gianmarco De Francisci, Gionis, Aristides, & Mathioudakis, Michael. 2018。量化社交媒体上的争议。*ACM*社会计算汇刊，**1**（1），1–27。Gaumont, Noé, Panahi, Maziyar, & Chavalarias, David. 2018。政治活动人士推特网络的社会语义动态重建——应用于2017年法国总统选举的方法。*PLOS ONE*，**13**（9），e0201879。Hagey, Keach, & Horwitz, Jeff. 2021。Facebook曾试图让其平台变得更健康，结果却让平台更加愤怒。《华尔街日报》，9月。Huszár, Ferenc, Ktena, Sofia Ira, O'Brien, Conor, Belli, Luca, Schlaikjer, Andrew, & Hardt, Moritz. 2022。Twitter上的政治算法放大。《美国国家科学院院刊》，**119**（1），e2025334119。Jacomy, Mathieu, Venturini, Tommaso, Heymann, Sebastien, & Bastian, Mathieu. 2014。ForceAtlas2：一种为Gephi软件设计的、便于网络可视化的连续图布局算法。*PLoS ONEONE*，**9**（6），e98679。Jager, Wander, & Amblard, Frédéric. 2005。利用基于主体的仿真模型，将一致性、两极分化和多元化捕捉为一般化风格行为。计算数学组织理论，1月。Jia, Peng, MirTabatabaei, Anahita, Friedkin, Noah E., & Bullo, Francesco. 2015。意见动力学与影响力网络中社会权力的演化。*SIAM*评论，**57**（3），367–397。出版者：工业与应用数学学会。Knobloch-Westerwick, Silvia, Mothes, Cornelia, & Polavin, Nick. 2017。政治信息选择性接触中的确认偏误、内群体偏误和负面偏见。传播研究，**47**（1），104–124。Leicht, E. A., & Newman, M. E. J. 2008。有向网络中的社群结构。《物理评论快报》，**100**（11）。Levinsson, Anna, Miconi, Diana, Li, Zhiyin, Frounfelker, Rochelle L., & Rousseau, Cé-cile. 2021。新冠肺炎疫情期间青年人的阴谋论、心理困扰与对暴力激进化的认同：一项横断面研究。《国际环境研究与公共卫生杂志》，**18**（15），7846。编号：15。出版者：多学科数字出版研究所。

Martin、Louis、Muller、Benjamin、Suárez、Pedro Javier Ortiz、Dupont、Yoann、Romary、Laurent、de la Clergerie、Éric、Seddah、Djamé 和 Sagot、Benoît。2020。CamemBERT：一种美味的法语语言模型。载于：第*58*届计算语言学协会年会论文集。计算语言学协会。

McPherson、Miller、Smith-Lovin、Lynn 和 Cook、James M. 2001。物以类聚：社交网络中的同质性。《社会学年度评论》，**27**(1)，415–444。

Mestre、Abel。2022。埃里克·泽穆尔，法西斯圈的新总统？《世界报》，3月。

Noorazar、Hossein、Vixie、Kevin R.、Talebanpour、Arghavan 和 Hu、Yunfeng。2020。从经典到现代的意见动力学。《现代物理学国际期刊 *C*》，**31**(07)，2050101。出版商：世界科学出版公司。

Perra, Nicola、Gonçalves, Bruno、Pastor-Satorras, Romualdo 和 Vespignani, Alessandro。2012年。时变网络的活动驱动建模。《科学报告》，**2**(1)，1–7。出版商：自然出版集团。

弗吉尼亚大学政治中心。2021年（9月）。新倡议探讨拜登与特朗普选民之间深刻且持久的分歧——萨巴托的水晶球。

Ramaciotti Morales, Pedro 和 Cointet,Jean-Philippe。2021年（9月）。审查社交网络推荐对几何意识形态空间中极化的影响。载于：*RecSys '21*：第*15*届 *ACM* 推荐系统会议。

Reyero, Tomás Mussi、Beiró, Mariano G.、Alvarez-Hamelin, J. Ignacio、Hernández, Laura 和 Kotzinos, Dimitris。2021年。选举时期政治意见格局的演变。《*EPJ* 数据科学》，**10**(1)，31。编号：1 出版商：施普林格柏林海德堡。

Rozin, Paul、& Royzman, Edward B. 2001。负面偏见、负面主导效应与传染效应。《人格与社会心理学评论》，**5**(4)，296–320。出版商：SAGE 出版公司。

Santos, Fernando P.、Lelkes, Yphtach、& Levin, Simon A. 2021。链接推荐算法与在线社交网络中的极化动态。《美国国家科学院院刊》，**118**(50)，e2102141118。出版商：《美国国家科学院院刊》。

Tokita, Christopher K.、Guess, Andrew M.、& Tarnita, Corina E. 2021。极化的信息生态系统可以通过信息级联重组社交网络。《美国国家科学院院刊》，**118**(50)，e2102147118。公司：美国国家科学院 分销商：美国国家科学院 ISBN：9782102147111 机构：美国国家科学院 标签：美国国家科学院 出版商：《美国国家科学院院刊》。

Traag, V. A.、Waltman, L. 和 van Eck, N. J. 2019。从Louvain算法到Leiden算法：确保形成连接良好的社区。《科学报告》，**9**（1）。

Twitter。2020年。使用 *Twitter*。

Törnberg, Petter。2022。数字媒体如何通过党派排序推动情感极化。《美国国家科学院院刊》，*119*（42），e2207159119。Vosoughi, Soroush、Roy, Deb、和 Aral, Sinan。2018。网络上真实新闻与虚假新闻的传播。《科学》，*359*（6380），1146–1151。Xu, Zhiheng 和 Yang, Qing。2012。分析 Twitter 上的用户转发行为。载于：*2012* 年 *IEEE/ACM*社会网络分析与挖掘国际会议。IEEE。Zubrow, Keith。2021。Facebook 举报人称，公司会激励“愤怒、极化、分裂性内容”。*CBS*新闻，10 月。