The Self-Preferencing Prohibition as a Fairness Provision

ANDREAS HAUPT, Stanford University, USA

Self-preferencing is the practice by hybrid platforms to treat their own products more favourably than "comparable" third-party products. Regulations in the European Union and the United Kingdom restrict or ban self-preferencing. This paper provides a definition of self-preferencing using the language of counterfactual algorithmic fairness. The definition clarifies the interaction of different preferencing mechanisms, proxies, and resolving variables, and highlights challenges of causal and structural estimation, and provides a preferencing test generalizing the outcome-based test from the literature.

CCS Concepts: • **Social and professional topics → Antitrust and competition**; • **Applied computing →** *Economics*; *E-commerce infrastructure.*

Additional Key Words and Phrases: self-preferencing, fairness, causality

1 INTRODUCTION

Large platforms employ algorithmic curation to steer consumers toward products in online search, e-commerce, and social media. The power of such steering has led to regulatory action in the European Union. The Digital Markets Act prohibits "preferencing" of products sold by hybrid platforms. In particular (*Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 October 2022 on contestable and fair markets in the digital sector (Digital Markets Act)* [2022, Art. 6(5)1]):

> The gatekeeper shall not treat more favourably, in ranking and related indexing and crawling, services and products offered by the gatekeeper itself than similar services or products of a third party.

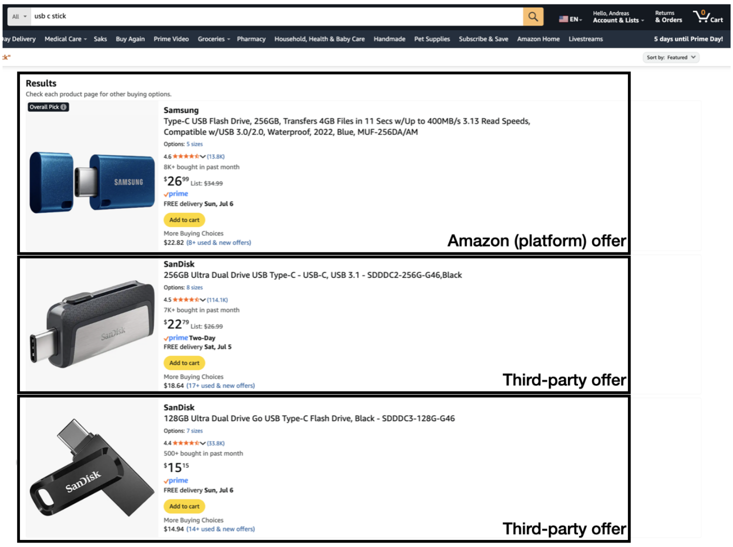
An example of a hybrid platform is Amazon, which matches consumers to third-party seller products through many algorithmic curation tools. Figure 1 shows the first entries of a search result page in response to a user query "usb c stick". The first, search result, most likely to be clicked, is an Amazon product, the latter two are third-party products.

This article provides a mathematical definition of what it means for a platform to *self-preference* in ranking. We cast the problem in the language of (counterfactual) algorithmic fairness, with a special focus on causal pathways from the sensitive attribute A (whether a product is the platform's own) to the ranking action R. Our definition clarifies the interaction of different preferencing mechanisms, proxies, and resolving variables, and highlights challenges of causal estimation that arise because ranking is chosen as a function of features and identity. While algorithmic fairness has had successes in several application areas such as school admissions and criminal justice, it has had limited impact on fairness provisions in (non-labor) platform markets.[1] Our main contribution is to introduce tools from algorithmic fairness into platform settings and provide algorithmic fairness with a new area of theoretical and applied research.

Self-preferencing in consumer AI agents. A second, emerging locus of preferencing is the *consumer AI agent* such as OpenAI's Agent—a conversational or autonomous assistant that plans, searches, compares, and guides the human to direct checkout [OpenAI 2025]. Such agents may soon be vertically integrated with marketplaces or monetized by affiliate fees and ads. Preferencing can then arise at the *plan* level: the agent chooses which merchants to consider, which pages to open, what to

[1] It is worth noting, that the long form of the Digital Markets Act is a "Regulation on Contestible and *Fair* Digital Markets" (emphasis added).

Author's address: Andreas Haupt, h4upt@stanford.edu, Stanford University, Stanford, California, USA.



Results

Check each product page for other buying options.

Overall Pick

Samsung

Type-C USB Flash Drive, 256GB, Transfers 4GB Files in 11 Secs w/Up to 400MB/s 3.13 Read Speeds, Compatible w/USB 3.0/2.0, Waterproof, 2022, Blue, MUF-256DA/AM

Options: 5 sizes

4.6 ★★★★★ (13.8K)

8K+ bought in past month

$26 99 List: ~~$34.99~~

prime

FREE delivery **Sun, Jul 6**

Add to cart

More Buying Choices

$22.82 (8+ used & new offers)

Amazon (platform) offer

SanDisk

256GB Ultra Dual Drive USB Type-C - USB-C, USB 3.1 - SDDDC2-256G-G46,Black

Options: 8 sizes

4.5 ★★★★★ (114.1K)

7K+ bought in past month

$22 79 List: ~~$26.99~~

prime **Two-Day**

FREE delivery **Sat, Jul 5**

Add to cart

More Buying Choices

$18.64 (17+ used & new offers)

Third-party offer

SanDisk

128GB Ultra Dual Drive Go USB Type-C Flash Drive, Black - SDDDC3-128G-G46

Options: 7 sizes

4.4 ★★★★★ (33.8K)

500+ bought in past month

$15 15

prime

FREE delivery **Sun, Jul 6**

Add to cart

More Buying Choices

$14.94 (14+ used & new offers)

Third-party offer

Fig. 1. Example of a ranking from the Amazon Marketplace. For prompt “usb c stick”, the first three offers are shown. The first offer is sold by Amazon, and the most expensive of the three. Two other offers are sold by third-party sellers. (Screenshot taken on amazon.com on July 3, 2025 from a U.S. IP address and an account enrolled in the Amazon Prime subscription programme.)

summarize, and which items to add to cart. Our framework (which will be introduced in detail later) applies verbatim: (i) the sensitive attribute A indicates whether the candidate option (merchant, offer, or plan step) is affiliated with the agent's provider; (ii) the action R is the proposed solution to the user; (iii) the predicted match value $\mathbf{Y}'$ is the agent's causal prediction of user utility or success conditional on actions; and (iv) non-preferencing requires that any dependence of the plan on affiliation flows only through $\mathbf{Y}'$. This means that our definition applies even in the emerging markets facilitated by AI agents.

This paper at a glance. Section 2 reviews causal diagrams and fairness concepts. Section 3 gives a general definition of non-preferencing. Section 4 discusses realistic complications (proxies, endogenous features and actions, multiple ranking surfaces) and how to address them. Section 5 develops a protocol that platforms and consumer AI providers can implement to demonstrate non-preferencing. We conclude in Section 6.

1.1 Related Work in Industrial Organization

Industrial Organization (IO) approaches to platform ranking and self-preferencing study how vertically integrated intermediaries allocate *salience*, and how it affects competition, entry, and welfare. Policy interest has been catalyzed by the EU's Platform-to-Business Regulation and the Digital Markets Act, which frame ranking neutrality and "differential treatment" as central concerns in hybrid platforms [*Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services and amending Regulation (EU) No 1215/2012 (Platform-to-Business Regulation)* 2019; *Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 October 2022 on contestable and fair markets in the digital sector (Digital Markets Act)* 2022]. Within IO, the empirical and theoretical literatures have converged on two complementary families of preferencing tests that map naturally to our notation (A, X, R, Y).

The first family comprises *conditioning-on-observables* tests: regress realized outcomes or exposure on a platform-identity indicator A controlling for observed features X, and interpret a positive residual dependence as "favoritism." These tests are straightforward to implement and transparent, but they inherit the classic IO worries of omitted variables and proxying: if X contains proxies for A, or if X fails to span demand-relevant heterogeneity (which would point to relevant features of products not captured in the features), the coefficient on A is difficult to interpret.

The second family, advocated by several recent papers, is *outcome-based*: ask whether own products receive *more exposure than would maximize a counterfactually estimated outcome*, holding fixed merit or predicted match value [Aguiar et al. 2021; Jürgensmeier and Skiera 2023; Reimers and Waldfogel 2023]. In streaming and e-commerce applications, these studies estimate potential outcomes under alternative rankings (e.g., via randomized exposure or quasi-experimental variation), then compare the platform's observed allocation to the outcome-maximizing allocation. When such causal estimation is possible, we claim, outcome-based metrics operationalize the "right" counterfactual: would the platform give itself the same position it gives comparable, in terms of observables but also consumers' revealed preference, third-party items? Our framework follows this outcome-based logic but strengthens it by explicitly formalizing *which* dependence on A is permissible—namely, only through a resolving variable $\mathbf{Y}'$ that predicts Y causally.

On the normative side, IO models analyze platforms' ranking objectives and the welfare implications of vertical integration. Reimers and Waldfogel [2023] study surplus-maximizing ranking and interpret deviations as evidence of preferencing, abstracting from strategic pricing. Hartzell and Haupt [2025] allow producers to respond strategically and characterize when consumer-optimal rankings can be implemented without disadvantaging third parties. Both strands clarify that some dependence of R on A may be *efficient* if platform identity shifts true match value (e.g., due to reliable fulfillment), but they also highlight that, absent causal identification of the $A \to Y$ channel, efficiency and favoritism are observationally confounded. Our definition disentangles these channels by requiring any $A \to R$ link to be mediated by $\mathbf{Y}'$.

A separate empirical line measures preferencing with tests resembling the outcome-based test. Lee and Musolff [2021] examine entry into Amazon product pages and quantify self-preferencing with a demand model that maps observed placements into implied advantages for own products. Lam [2021] estimate a structural search model with embedded ads and show that search frictions and behavioral responses can offset gains from better matching, complicating the welfare accounting of ranking tweaks. Farronato et al. [2023] provide descriptive evidence of keyword-search self-preferencing, documenting systematic visibility advantages Amazon Retail on the Amazon Marketplace.

Identification challenges in this literature mirror classic IO demand problems: R is chosen as a function of (A, X), X is endogenous to seller and platform decisions (pricing, fulfillment, assortment), and outcomes Y reflect equilibrium behavior. Consequently, valid tests typically rely on randomized exposure, policy shocks, or instruments—methodological themes that connect to the use of instruments and experimental variation in demand estimation [Berry et al. 1995]. Our framework embraces these constraints: the validity requirement for $\mathbf{Y}'$, see Definition 3.1 is an identification statement, and our proposed test in Section 5 makes explicit options for the platform to establish causal estimation.

1.2 Related Work in Algorithmic Fairness

Work on algorithmic fairness in *ranking* has crystallized around two complementary views: (i) *group-level* constraints that protect the exposure or representation of protected groups, and (ii) *individual-level* guarantees that align attention with merit over time. Group-fairness approaches include constrained re-ranking methods for top-k results (e.g., meeting minimum group proportions while maximizing utility) and general frameworks that optimize standard ranking objectives subject to exposure or representation constraints. Exposure-based formulations link fairness to user attention: visibility allocated by the ranking should be proportional to item merit, leading to constrained or policy-learning methods that control *expected exposure* across groups and queries, and to practical deployments at scale [Celis et al. 2017; Zehlike et al. 2017]. On the individual side, amortized notions aim to equalize *equity of attention* across repeated rankings, ensuring that similarly relevant items receive comparable cumulative exposure, and pairwise-fairness criteria align the probability of correctly ordering two items with their true relevance difference [Geyik et al. 2019; Morik et al. 2020; Singh and Joachims 2018]. A recurring systems theme is *how* to enforce fairness: pre-processing relevance labels to remove bias, in-processing constrained learning-to-rank, or post-processing re-ranking with guarantees, including stochastic policies that satisfy fairness in expectation [Beutel et al. 2019; Biega et al. 2018]. In hybrid-platform markets, these tools translate naturally into *non-preferencing* constraints: platform-owned items must not receive excess exposure beyond their estimated merit [Diaz et al. 2020]. Other approaches enforce proportional representation or bounded disparity while trading off with relevance [Celis et al. 2017; Yang and Stoyanovich 2017; Zehlike et al. 2017] or maximize utility under exposure constraints. Our framework is complementary to these approaches and attempts to directly translate the provision of the Digital Markets Act into a fairness definition. In contrast to notions discusssed in the literature, it does not treat the outcome Y as a notion of merit, but merely as a relevant measure affecting *similarity.*

2 BACKGROUND

Several questions about similar products will be questions of the form "would a product sell more than another would it be placed at another rank". We model the distribution of sales using causal diagrams. More specifically, we use graphical models to reason about interventions on ranking algorithms and to make the distinction between *observational* and *counterfactual* claims precise. A causal model consists of a directed acyclic graph (DAG) $G = (V, E)$ together with a collection of structural equations $\{X \coloneqq f_X(\mathrm{pa}(X), U_X)\}_{X\in V}$, where U_X are jointly independent exogenous disturbances. The *Markov factorization* of the induced observational distribution is

$$\mathbb{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \mid \mathrm{pa}(X_i)).$$

Conditional independences are read from G via d-separation: two sets of variables are conditionally independent given S if every path between them is blocked when conditioning on S. Intuitively, blocked non-colliders transmit no information once conditioned on, and colliders transmit no information unless they or their descendants are conditioned on.[2]

Interventions and counterfactuals. An intervention $\mathrm{do}(Z = z)$ replaces the structural assignment for Z by the constant z, yielding the interventional distribution $\mathbb{P}(\bullet \mid \mathrm{do}(Z = z))$. We will frequently write potential outcomes $Y(z)$ for the post-intervention value of Y when Z is set to z, as well as

[2] Formally, a path between two nodes in a directed acyclic graph (DAG) is *blocked* given a set of conditioning nodes Z if (i) it contains a chain $A \longrightarrow B \longrightarrow C$ or a fork $A \longleftarrow B \longrightarrow C$ where the middle node $B \in Z$, or (ii) it contains a collider $A \longrightarrow B \longleftarrow C$ where $B \notin Z$ and no descendant of B is in Z. Two sets of nodes X and Y are *d-separated* by Z if all paths between X and Y are blocked given Z.

nested counterfactuals such as $Y(r, a)$ for the value of Y when the rank decision is set to $R = r$ and the sensitive attribute is set to $A = a$. In our setting, causal effects of ranking take the form

$$\Delta(r; x, a) = \mathbb{E}\left[Y(r) \mid X = x, A = a\right] - \mathbb{E}\left[Y(r') \mid X = x, A = a\right].$$

Observational vs. counterfactual in a self-preferencing setting. Let $A \in \{0, 1\}$ indicate whether the seller is the platform ($A = 1$) or a third party ($A = 0$), let X collect observable features (price, shipping, reviews), R be the ranking action, and Y the desirable outcome (e.g., click or purchase). Two common quantities are:

Observational Comparison One might compare, for items with similar X shown at the same rank r, the average outcomes

$$\Delta_{\text{obs}}(r; x) = \mathbb{E}[Y \mid R = r, A = 1, X = x] - \mathbb{E}[Y \mid R = r, A = 0, X = x].$$

A nonzero Δ_{obs} does *not* by itself establish self-preferencing: it may reflect unobserved differences in quality or selection (e.g., platform-owned items having different latent appeal not captured by the features).

Counterfactual Comparison The question relevant to (non-)preferencing is whether the platform *treats* otherwise similar items differently *because* they are platform-owned. Fix an item with features $X = x$ and consider the counterfactual contrast

$$\Delta_{\text{cf}}(r; x) = \mathbb{E}\left[Y(r, 1) \mid X = x\right] - \mathbb{E}\left[Y(r, 0) \mid X = x\right].$$

If $\Delta_{\text{cf}}(r; x) = 0$ for all r (or more weakly, as we shall propose, conditional on the platform's merit estimates $\mathbf{Y}'$), then the platform's treatment does not causally depend on A beyond permitted pathways. This paper will argue that self-preferencing should demand $\Delta_{\text{cf}}(r; x) = 0$.

Example 2.1 (Buy-box placement). Suppose the platform places exactly one offer into a prominent position ($R \in \{0, 1\}$). Observationally, you find that among offers with comparable X, platform-owned items achieve 12% click-through at $R = 1$ while third-party items achieve 10%. This yields $\Delta_{\text{obs}}(1; x) = 0.02$. A counterfactual analysis, however, randomizes buy-box assignment for a stratified subset with identical predicted merit $\mathbf{Y}'$ and estimates $\mathbb{E}[Y(1, 1) \mid \mathbf{Y}', X] - \mathbb{E}[Y(1, 0) \mid \mathbf{Y}', X]$. If this difference is statistically indistinguishable from zero while $\mathbb{E}[Y(1) \mid \mathbf{Y}', X] > \mathbb{E}[Y(0) \mid \mathbf{Y}', X]$ holds for both groups, the platform is *not* self-preferencing: the higher observed clicks for platform items were due to merit captured by $\mathbf{Y}'$, not to A itself. Conversely, a positive $\Delta_{\text{cf}}(1; x)$ after conditioning on $\mathbf{Y}'$ indicates a direct effect of platform association to ranking, suggesting self-preferencing.

Causal fairness. Causality of fairness definitions is crucial in this paper. Kilbertus et al. [2018a] propose explicitly modeling a causal model to identify and remove "discriminatory paths" from A to the outcome R, ensuring that only permissible causal influences remain. They also introduce the concept of resolving variables. Similarly, Kusner et al. [2018] introduce *counterfactual fairness*, which requires that for any individual, the action R remains unchanged under a hypothetical intervention setting A to a different value. Loftus et al. [2018] extend this perspective by distinguishing different types of protected attributes—direct, indirect, and spurious—and propose procedures to block unfair channels while preserving legitimate ones. Nabi and Shpitser [2018] further develop *fair inference on outcomes* by formulating path-specific effects and adjusting for mediators that carry unacceptable bias.

3 A GENERAL DEFINITION OF PREFERENCING

Consider n *offers* $i \in \{1, \dots, n\}$. Each offer has features $X_i \in \mathcal{X}$, a sensitive attribute $A_i \in \{0, 1\}$ (e.g., 1 if the platform/agent is the seller or otherwise affiliated), an outcome $Y_i \in \{0, 1\}$ (e.g., click or purchase[3]), and a platform action $R \in \mathcal{R}$ that allocates slots to offers. Let $\mathbf{Y}'_i = (Y'_{ir})_{r \in \mathcal{R}}$ denote the platform's (or agent's) outcome for i when ranking r is chosen. Write $(\cdot)_{-i}$ for the vector over all agents except i.

3.1 Definition via Potential Outcomes

Outcomes may depend on the full ranking and on all offers' attributes. We therefore treat the *competitive context* for offer i as

$$C_i \equiv (X_{-i}, A_{-i}),$$

We now state the two core requirements.

Definition 3.1 (Calibration). For each session and offer i, $\mathbf{Y}'_i$ is a calibrated outcome estimate, for every ranking $r \in \mathcal{R}$,

$$\mathbb{E}\left[Y'_{ir} \middle| X_i, A_i, C_i\right] = \mathbb{E}\left[Y_i | X_i, A_i, C_i, \mathrm{do}(R_i = r)\right],$$

The right-hand side is a potential-outcome object: the expected outcome for i under ranking r while holding the rest of the context C_i fixed.

Definition 3.2 (Non-preferencing). A ranking rule R is *non-preferencing* if there exist calibrated estimates $(\mathbf{Y}'_i)_{i=1}^n$ such that, for every seller i

$$A_i \perp\!\!\!\perp R_i | \mathbf{Y}'_i, C_i, \tag{1}$$

$$Y_i \perp\!\!\!\perp \mathbf{Y}'_i | A_i, R, C_i. \tag{2}$$

Condition (1) rules out any residual direct effect of affiliation A_i on the slot assignment for i once the predicted match values and the competitive context are fixed. Condition (2) states that $\mathbf{Y}'_i$ carries no additional information about realized outcomes beyond (A_i, R, C_i); it is a calibrated prediction built from historical data and experimentation.

Remark 1. *(i) The symmetric definition above bans preferential treatment* in either direction. *If one only aims to ban* self*-preferencing,* (1) *can be weakened to an inequality on exposure or placement probabilities for $A_i = 1$ versus $A_i = 0$ given $(\mathbf{Y}'_i, C_i)$. (ii) We do not impose $A_i \perp\!\!\!\perp X_i$ for now, proxy issues are handled in Section 4 and operationalized in Section 5.*

Example 3.3 (Logit demand). Consider a single user request surfacing n offers. The platform assigns a (possibly partial) ranking R, with slot-specific position effects $\rho_r \in \mathbb{R}$ and an outside option with utility normalized to 0. Assume that user purchases Y_i are generated through a random utility model

$$U_i = \beta^\top X_i + \alpha A_i + \rho_{R_i} + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \text{T1EV},$$

so that the probability of choosing offer i is multinomial logit:

$$\mathbb{P}(Y_i = 1 \mid X, A, R) = \frac{\exp(\beta^\top X_i + \alpha A_i + \rho_{R_i})}{1 + \sum_{j=1}^n \exp(\beta^\top X_j + \alpha A_j + \rho_{R_j})}.$$

[3] We restrict here to binary outcomes as in virtually all settings of relevance for ecommerce and social media either a clickthrough, or purchase (conversion) are relevant. The theory here generalizes with richer outcomes, in which case we need to replace the outcome with a relevant expectation.

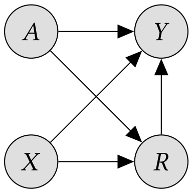


Fig. 2. A ranking setting. A platform ranking decision R is affected by algorithmic features X, potentially including A. Ranking, features, and A determine a desirable outcome. The contested question is whether the platform action R may depend on the source of the product A.

Fix a realized ranking R and define the *slot-swap* potential outcome $Y_i(r; R_{-i})$ as the choice indicator if we place i in slot r while holding other items' slots R_{-i} fixed. Then the calibrated deterministic predictor for item i is

$$Y'_{ir}(X, A) = \frac{\exp(\beta^\top X_i + \alpha A_i + \rho_r)}{1 + \sum_{j \neq i} \exp(\beta^\top X_i + \alpha A_i + \rho_r)}.$$

Under Definition 3.2, for any fixed competitive context $C_i = (X_{-i}, A_{-i})$ the distribution of R_i may depend on $\mathbf{Y}'_i$ and C_i but *not* on A_i beyond its contribution to $\mathbf{Y}'_i$. In particular, any regression or conditional randomization test of R_i on $(\mathbf{Y}'_i, C_i, A_i)$ should find no residual effect of A_i once $(\mathbf{Y}'_i, C_i)$ are controlled for.

Other specifications. The definition and tests do not rely on the logit form. For instance, one may model two-stage behavior where an offer enters a consideration set with probability $\gamma(X_i, A_i, R_i)$ (cascade or position-based click models), followed by a choice among considered items via multinomial logit, nested logit, or mixed logit. In such cases, Y'_{ir} is the model-implied $\mathbb{E}[Y_i \mid \mathrm{do}(R_i = r), X, A, C_i]$; the non-preferencing conditions (1)–(2) remain unchanged.

3.2 A Graphical Definition

We can derive the definition of self-preferencing graphically (where we drop for simplicity dependence on i and implicitly condition on C_i). In principle, all conditions given in Figure 2 are possible. This diagram is equivalent to assuming that X and A are unconditionally independent, $A \perp\!\!\!\perp X$. The desirable outcomes (e.g., sales) may be affected by all other attributes (the platform product might be more or less desirable, even for the same observable features X; features X, including the price of the good, affect the outcome; and the ranking decision changes the outcome if it steers consumers). R in this model is potentially affected by both A and X.

The ranking setting has led the literature, and *Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services and amending Regulation (EU) No 1215/2012 (Platform-to-Business Regulation)* [2019] to focus on differential treatment, so the differences in $R \mid X, A = a$ for different values of the sensitive attribute $a \in \{0, 1\}$, suggesting that R should not depend on the value of A, Figure 3. Following the literature in algorithmic fairness [Barocas and Selbst 2016], we could call this requirement *unawareness.*

Unawareness (partly also called the controlling-on-observables approach) is argued to be undesirable in Industrial Organization [Hartzell and Haupt 2025; Lee and Musolff 2021; Reimers and Waldfogel 2023]. The main objection is that unawareness does not capture sufficiently what it means to be "similar". The thought experiment that Hartzell and Haupt [2025], Lee and Musolff [2021], and Reimers and Waldfogel [2023] and others entertain is that products that a platform sells

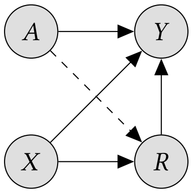


Fig. 3. The unawareness solution is to prohibit a causal effect of the platform identifier on ranking ($A \longrightarrow R$), here denoted with a dashed line.

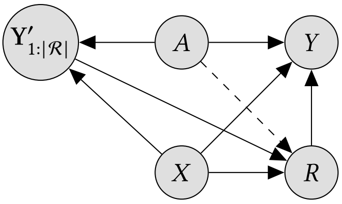


Fig. 4. Non-Self-Preferencing. While a causal effect of the platform dummy variable A on the outcome is prohibited, an indirect effect via an estimate of the desirable outcome Y is necessary. Together with Definition 3.1 the conditional independence implied by this causal graph define non-preferencing.

might be more desirable for consumers, but that the relevant desirability in observables other than the sensitive attribute.

This allows for a direct fix of the definition. *Ranking may only depend on whether it is sold by the platform through a causal path through an estimates of sales*, Figure 4. Here, $\mathbf{Y}'$ is an *estimate* of sales. In the language of Kilbertus et al. [2018b], this means that an estimate of sales $\mathbf{Y}'$ is a *resolving variable* for the differential treatment caused by the sensitive variable A.

The conditional independence conditions that Figure 4 expresses are the following:[4]

$$A \perp\!\!\!\perp X \tag{3}$$

$$A \perp\!\!\!\perp R \mid \mathbf{Y}' \tag{4}$$

$$Y \perp\!\!\!\perp \mathbf{Y}' \mid A, R. \tag{5}$$

We can view the additional requirement (3) as a non-proxy condition—A must be unconditionally independent of X.

4 PROXIES, OUTCOMES, AND INTERACTIONS

Several additional complications arise in realistic platforms. This section sharpens the discussion using the lens of algorithmic fairness and makes explicit what must be ruled out, what may be permitted, and how each issue surfaces in the test proposed in Section 5.

4.1 Features X and the sensitive attribute A are correlated

Treating A solely as a binary platform-identity flag is often insufficient because many observable covariates X partially reveal A. In the language of Kilbertus et al. [2018a], X may *contain proxies* for A. In marketplace settings, fulfillment mode, return policy, or seller ratings can correlate strongly with platform identity (e.g., *shipped by* the platform), as illustrated in Figure 1.

[4] That is, these are the d-sarated triples in the directed graph depicted in Figure 4.

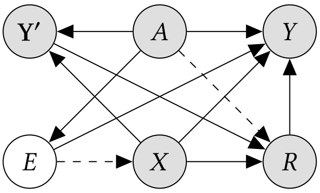


Fig. 5. Endogeneity and confounding. Latent decisions E (e.g., entry and pricing) affect X and Y and may be influenced by A, biasing naive estimates of $\mathbb{E}[Y \mid X, A, R]$. Valid construction of $\mathbf{Y}'$ requires randomized exposure or credible instruments.

We formalize proxy detection and usage as follows. For a candidate feature X_j, define a *proxy strength* functional

$$\text{ProxyStrength}(X_j; A) \in \{|\text{Corr}(X_j, A)|, I(X_j; A)\} \tag{6}$$

where $I(\bullet;\bullet)$ denotes mutual information. A feature is *proxy-admissible for direct use* in R if $\text{ProxyStrength}(X_j; A) \leq \tau$ for a pre-declared threshold τ. Let $W \subseteq X$ be the set of proxy-admissible features. All non-admissible features may still enter the prediction stage for $\mathbf{Y}'$ but *cannot* enter the ranking rule directly. This implements the resolving-variable principle of Kilbertus et al. [2018a]: any permissible dependence on A must operate via $\mathbf{Y}'$.

4.2 The value Y is not clearly defined

Our framework assumes that Y captures the ranking-relevant notion of "similarity". In practice, this similarity is given by anything that would lead to a different behavior of the consumer. This would be given, fo example, by:

$$\text{clicks: } Y \in \{0, 1\} \text{ for a click,}$$
$$\text{conversion: } Y \in \{0, 1\} \text{ for apurchase.}$$

In general, any metric that is a user action may be taken as an outcome measure Y.

4.3 The ranking R interacts with other ranking mechanisms

Large platforms deploy multiple steering mechanisms: homepage recommendations, search ordering, product-page placement (e.g., Buy Box), and conversational surfaces. Let $\mathcal{S}$ index surfaces and write $R = (R^{(s)})_{s \in \mathcal{S}}$. Then, it follows directly that if for some s we have that Definition 3.2 fails conditional on R_{-s}, then it also fails for $R = (R^{(s)})_{s \in \mathcal{S}}$. Thus, demonstrating preferencing on *any* surface suffices to demonstrate preferencing overall.

4.4 Features X and rankings R are endogenous

Estimating a calibrated $\mathbf{Y}'$ from observational logs is challenging because R is a function of (X, A) and because seller and platform decisions create latent confounding. In Figure 5, a latent variable E (e.g., entry and pricing decisions) affects both X and Y and may be influenced by A, and require careful causal inference.

4.5 Practical implications for enforcement.

In light of these complications, Section 5 requires (i) a proxy screen determining which features may enter R directly, (ii) an explicit construction of $\mathbf{Y}'$ with calibration diagnostics, and (iii) a test of the predicted outcomes under different sensitive attributes and the same ranking. Together these elements ensure that any residual dependence of R on A is either eliminated or channeled exclusively through a causally justified resolving variable.

5 DEMONSTRATING NON-SELF-PREFERENCING

This section expands the practical protocol that allows a hybrid platform to demonstrate that its ranking rule R satisfies our non-preferencing criterion from Definition 3.2. The regulator's null hypothesis is that the platform *does not* self-preference, i.e.,

$$H_0 : \quad A \perp\!\!\!\perp R \mid \mathbf{Y}' \quad \text{with calibrated } \mathbf{Y}' \text{ as in Definition 3.1.} \tag{7}$$

Rejection of (7) constitutes evidence of self-preferencing. The protocol below is designed to be auditable, modular, and feasible at scale.

5.1 A Regulatory Artifact

Auditing requires session-level logs that include, for each candidate i surfaced to a user request (e.g., a query or page view): (i) a snapshot of the ranking policy $\pi(x, \mathbf{y}')$ and the prediction algorithm producing the vectors $\mathbf{Y}'_i = (Y'_{ir})_{r \in \mathcal{R}}$, $i = 1, 2, \ldots, n$ at decision time, (ii) a minimal feature audit trace W_i (the subset of features that the platform claims enter R directly), (ii) the sensitive attribute A_i, and (iv) outcomes Y_i (click, add-to-cart, purchase, or whatever is designated as the desirable outcome).

Step 1: Proxy screening for features used directly in R. Let $X = (W, Z)$ split features into W (permitted to enter R directly) and Z (proxies that are *not* permitted to enter R directly). The platform must quantify the strength of association between each candidate feature and A ex ante, using correlation $(\mathrm{Corr}(X_j, A)|)$ or mutual information $(I(X_j; A))$. Denote ProxyStrength$(W_j; A)$ this measure of association the and publish the threshold τ such that all W satisfy ProxyStrength$(W_j; A) \leq \tau$.

Step 2: Constructing a calibrated *resolving variable* $\mathbf{Y}'$. Validity in Definition 3.1 is causal; mere supervised predictions are insufficient when R is endogenous. Two implementation routes are acceptable:

Randomized exposure (preferred) Reserve a small traffic slice (e.g., 0.01%) wherein ranking is randomized using a known stochastic policy $\pi_0(r \mid x, a)$. Learn $\mu_r(x, a) := \mathbb{E}[Y \mid X = x, A = a, R = r]$ on this slice. Set $Y'_r(x, a) := \mu_r(x, a)$ and report calibration diagnostics:

$$\mathrm{CalErr}_r := \mathbb{E}\left[(\mathbf{1}_{\{R=r\}} \cdot (Y - \mu_r(X, A)))^2\right], \qquad r \in \mathcal{R}. \tag{8}$$

Instrumental-variables or doubly-robust identification When randomization is infeasible, document instruments (policy shocks, eligibility thresholds) affecting R but not Y except through R, and estimate μ_r via a doubly robust (DR) estimator [Chernozhukov et al. 2018]. Let $p_r(x, a) := \mathbb{P}(R = r \mid X = x, A = a)$ be the propensity (known under holdout randomization or estimated otherwise). The DR functional

$$\psi_r(x, a) := \mu_r(x, a) + \frac{\mathbf{1}_{\{R=r\}}}{p_r(X, A)}\big(Y - \mu_r(X, A)\big) \tag{9}$$

is unbiased for $\mathbb{E}[Y \mid X = x, A = a, \mathrm{do}(R = r)]$ if either the propensity or outcome model is correct. Set $Y'_r(x, a) := \mu_r(x, a)$ and publish sensitivity of the next step to alternative specifications of μ_r and p_r.

Step 3: Structural fit test for $\pi\big(x, (\mu_r(x, a, c))_{r \in \mathcal{R}}\big)$. As a third estimate, provide evidence that the model fit on a smaller dataset also fits on all data on the platform. To devise this goodness-of-fit test, index requests by $s = 1, \ldots, m$. Request s surfaces a candidate set $\mathcal{I}_s$ with item features $\{(X_{si}, A_{si})\}_{i \in \mathcal{I}_s}$ and competitive context C_s (e.g., query, page, surface), yielding a ranking action r_s. The declared ranking policy specifies a conditional distribution over rankings

$$\pi(R_s \mid X_s, S_s, C_s), \quad \text{where}$$

Report inability to reject the conditional goodness-of-fit test

$$H_0^{\pi(\mu)}: \qquad \mathbb{P}(R_s \mid X_s, A_s, C_s) = \pi\big(R_s \mid X_s, S_s, C_s\big) \quad \text{for all } s = 1, 2, \ldots, m.$$

5.2 Relationship to outcome-based and conditioning-on-observables tests

The proposed SPS test operationalizes the outcome-based spirit in Aguiar et al. [2021], Jürgensmeier and Skiera [2023], and Reimers and Waldfogel [2023] while avoiding the pitfalls of naïvely conditioning on observables: any permissible dependence on A must flow through $\mathbf{Y}'$.

5.3 Putting it all together

As part of a *Non-Preferencing Card*, ship the following pieces for each relevant ranking product:

(1) *Proxy screen* ranking features to obtain W; publish thresholds for proxy association.
(2) *Obtain calibrated* $\mathbf{Y}'$ via randomized exposure or DR with instruments; publish calibration and sensitivity.
(3) *Test* out-of-sample goodness-of-fit of π and μ.

Passing all steps provides a concise, testable demonstration that observed dependence on platform identity, if any, operates exclusively through predicted match value, i.e., no self-preferencing.

6 CONCLUSION

This paper formalizes self-preferencing in algorithmic ranking as a causal fairness problem. We model platform identity A, features X, actions R, and outcomes Y in a directed graphical framework and define *non-preferencing* as the existence of a *resolving variable* $\mathbf{Y}'$ that is a calibrated predictor of a consumer action (in virtually all cases a click or purchase) under alternative platform ranking actions. This lens clarifies how classical objections to "unawareness" in Industrial Organization can be addressed without granting the platform carte blanche to use A: dependence on platform identity is permitted only to the extent that it changes predicted outcomes in a way that is causally justified.

We then translate the definition into an auditable protocol: proxy screening for direct features, construction and calibration of $\mathbf{Y}'$ using randomized or instrumental-variable-based identification tests of differential effects. Conceptually, our tests bridge outcome-based preferencing diagnostics in IO with fairness notions in the ranking literature, while making the causal pathways explicit.

We highlight that infrastructure to detect self-preferencing will benefit in particular the responsible development of consumer AI agents, which may exhibit much higher consumer steering than current platform technologies.

REFERENCES

Luis Aguiar, Joel Waldfogel, and Sarah Waldfogel. 2021. "Playlisting favorites: Measuring platform bias in the music industry." *International Journal of Industrial Organization*, 78, 102765. doi: https://doi.org/10.1016/j.ijindorg.2021.102765.

Solon Barocas and Andrew D Selbst. 2016. "Big data's disparate impact." *Calif. L. Rev.*, 104, 671.

Steven Berry, James Levinsohn, and Ariel Pakes. 1995. "Automobile prices in market equilibrium." *Econometrica : journal of the Econometric Society*, 63, 4, 841–890. Publisher: [Wiley, Econometric Society]. Retrieved Feb. 13, 2025 from http://www.jstor.org/stable/2171802.

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. "Putting fairness principles into practice: Challenges, metrics, and improvements." In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 453–459.

Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. "Equity of attention: Amortizing individual fairness in rankings." In: *The 41st international acm sigir conference on research & development in information retrieval*, 405–414.

L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2017. "Ranking with fairness constraints." *arXiv preprint arXiv:1704.06840.*

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Jan. 2018. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal*, 21, 1, (Jan. 2018), C1–C68. eprint: https://academic.oup.com/ectj/article-pdf/21/1/C1/27684918/ectj00c1.pdf. DOI: 10.1111/ectj.12097.

Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. "Evaluating stochastic rankings with expected exposure." In: *Proceedings of the 29th ACM international conference on information & knowledge management*, 275–284.

Chiara Farronato, Andrey Fradkin, and Alexander MacKay. May 2023. *Self-Preferencing at Amazon: Evidence from Search Rankings*. Tech. rep. (May 2023), 239–43. DOI: 10.1257/pandp.20231068.

Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. "Fairness-aware ranking in search & recommendation systems with application to linkedin talent search." In: *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, 2221–2231.

Olivia Hartzell and Andreas Haupt. 2025. *Platform preferencing and price competition I: Evidence from amazon*. (2025). https://hartzell.scholars.harvard.edu/sites/g/files/omnuum5151/files/2025-01/Platform_preferencing_i_statics.pdf.

Lukas Jürgensmeier and Bernd Skiera. 2023. "Measuring Fair Competition on Digital Platforms." *SSRN Electronic Journal*. DOI: 10.2139/ssrn.4393726.

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2018a. *Avoiding Discrimination through Causal Reasoning*. (2018). https://arxiv.org/abs/1706.02744 arXiv: 1706.02744 [stat.ML].

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Jan. 2018b. *Avoiding Discrimination through Causal Reasoning*. en. arXiv:1706.02744 [stat]. (Jan. 2018). DOI: 10.48550/arXiv.1706.02744.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2018. "Counterfactual Fairness." https://arxiv.org/abs/1703.06856 arXiv: 1703.06856 [stat.ML].

H Tai Lam. 2021. *Platform search design and market power*. (2021).

Kwok Hao Lee and Leon Musolff. 2021. "Entry Into Two-Sided Markets Shaped By Platform-Guided Search." *Working paper*, 1–64.

Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. 2018. "Causal Reasoning for Algorithmic Fairness." *arXiv preprint arXiv:1805.05859*. DOI: 10.48550/arXiv.1805.05859.

Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. "Controlling fairness and bias in dynamic learning-to-rank." In: *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, 429–438.

Razieh Nabi and Ilya Shpitser. 2018. "Fair Inference on Outcomes." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 1. Vol. 32. DOI: 10.1609/aaai.v32i1.11553.

OpenAI. 2025. *Agentic Commerce Protocol: Get Started*. https://developers.openai.com/commerce/guides/get-started/. Accessed: 2025-09-30. (2025).

Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services and amending Regulation (EU) No 1215/2012 (Platform-to-Business Regulation). Brussels, 11 July 2019. (June 20, 2019). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019R1150.

Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 October 2022 on contestable and fair markets in the digital sector (Digital Markets Act). Brussels, 12 October 2022. (Oct. 14, 2022). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%5C%3A32022R1925.

Imke Reimers and Joel Waldfogel. 2023. *A framework for detection, measurement, and welfare analysis of platform bias*. Tech. rep. National Bureau of Economic Research.

Kulvinder Singh and Thorsten Joachims. 2018. "Fairness of Exposure in Rankings." In: *Proceedings of the 2018 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2219–2228. DOI: 10.1145/3219819.3219952.

Ke Yang and Julia Stoyanovich. 2017. "Measuring fairness in ranked outputs." In: *Proceedings of the 29th international conference on scientific and statistical database management*, 1–6.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. "Fa* ir: A fair top-k ranking algorithm." In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1569–1578.