OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Designing Behaviourally Informed Market Mechanisms in the Streaming Industry: Efficiency, Engagement, and Fairness in Two-Sided Platforms

Akshat Garg
akshatpublish@gmail.com

ABSTRACT

Streaming platforms rely heavily on ad auctions and recommendation algorithms to allocate attention between users, advertisers, and creators. Traditional market design approaches assume rational, quasi-linear agents, yet user behaviour in digital environments systematically deviates from these assumptions due to cognitive limits, loss aversion, and engagement fatigue. This paper examines how integrating behavioural economics with market design can improve efficiency, engagement, and fairness in streaming ecosystems. Using examples from ad-supported platforms such as YouTube and subscription-driven services like Netflix and Spotify, the study highlights how Generalized Second-Price auctions and multi-layered recommendation systems may generate unintended distortions, including advertising weariness, algorithmic lock-in, and creator bias. It proposes constrained optimisation frameworks that incorporate user-specific tolerance thresholds, lifetime value considerations, and diversity-aware ranking rules. The findings suggest that allocation mechanisms grounded in both mathematical optimisation and behavioural realism are better positioned to sustain long-term platform welfare.

Keywords: Market design, behavioural economics, streaming platforms, ad auctions, algorithmic fairness

INTRODUCTION

Content consumption, over the past two decades, has undergone a paradigmatic shift from ownership-based legacy media to access-based personalized subscription models. This is evident from the emergence of Over-the-Top (OTT) platforms - such as Netflix, Youtube and Spotify - that have substantially eroded the market dominance of traditional television (Song, 2024), thus resulting in a global ‘cord-cutting’ phenomenon (Tefertiller, 2020). The economic scale of this phenomenon is profound. The global video streaming market was valued at USD 811 billion in 2025 and is projected to grow at a Compound Annual Growth Rate (CAGR) of 18.3% through 2032 (Sheynin & Brophy, 2025).

This disruption was catalysed by two key factors. Firstly, OTT platforms offered consumers unprecedented flexibility and accessibility in choosing content, in contrast to the rigid programming schedules and high advertising loads inherent to traditional television (Papathanasopoulos & Varoutas, 2024). Secondly, these services operate on data-driven personalization. OTT platforms analyze viewing habits using artificial intelligence and algorithms to curate individualized content and advertising recommendations (Gomez-Uribe & Hunt, 2015). This reduces search costs for the consumer and enhances user engagement. Thus, the core value proposition of streaming services is rooted in personalization.

OTT platforms operate like engineered markets, relying on recommendation systems as matching models and using dynamic auctions for ad placement. As per Bichler (2017), at the core of these systems are market design theory and mathematical optimization models - particularly linear programming - to ensure media content is efficiently allocated among platform users. Often, a key assumption in these systems is that platform users behave as rational agents. However, if personalization algorithms aim to perfect the user experience, why is the streaming industry grappling with high churn, subscription fatigue (Briel, 2025), and allegations that algorithms create "echo chambers" that limit user discovery (Fleder & Hosanagar, 2009)? The reason for this disconnect between algorithmic capability and user experience is that the optimization models and rational-choice assumptions underlying these platforms are designed for Homo Economicus ("Econs"), a concept first articulated by John Stuart Mill in the 1830s to describe a rational, self-serving model of human behaviour (Persky, 1995). Yet, the platforms' users are Homo Sapiens ("Humans"), who make predictable errors and are subject to limited attention, self-control problems, and cognitive biases. Hence, this paper aims to answer the following research question: **How can concepts from market design and behavioural economics be combined to improve efficiency, engagement, and fairness in the streaming industry?**

This paper aims to make two key contributions. Firstly, it synthesises the SAI (Smoothed Advertising Index) framework from Chae et al. (2019) with the lifetime value formulation from Ascarza et al. (2017) into a function that accounts for both immediate monetization and user retention, in order to optimise "behaviourally constrained" ad allocation. While prior work has incorporated user welfare into ad auctions through Markov Decision Process (MDP) formulations (Cai et al., 2024), this paper makes a distinct contribution by grounding the tolerance adjustment in Chae et al.'s (2019) latent-class weariness model. This allows the incorporated tolerance multiplier $T(u)$ to reflect heterogeneous behavioural types rather than a uniform user state, a distinction that matters crucially for personalisation and platform fairness. Secondly, it recommends a set of design principles - diversity-aware ranking, concave welfare functions, and governance tools for users to enhance user agency - to ensure fairness and efficiency in a two-sided streaming market.

BACKGROUND: MARKET DESIGN AND BEHAVIOURAL ECONOMICS

This paper's analysis is built upon the intersection of two distinct economic fields - market design and behavioural economics.

As Roth (2002) articulates in his seminal work, The Economist as Engineer, market design emerged as an

engineering-focused discipline concerned with the architecture and analysis of real-world allocation systems, particularly in complex environments where traditional supply-and-demand price mechanisms prove inadequate. For instance, two-sided matching mechanisms, such as the Gale-Shapley stable matching algorithm, were used for settings like kidney exchange (Roth et al., 2004) and school choice (Abdulkadiroglu and Sönmez, 2003), to efficiently determine a stable pairing between two sets of agents. Thus, mechanism design - often described as 'reverse game theory' - forms the theoretical core of market design. Rather than analyzing the outcomes of a predefined game, the designer specifies a desired social objective and then constructs the rules (or mechanisms) that incentivize rational, self-interested agents with private information to act in ways that achieve that precise goal (Bichler, 2017). There are two primary domains of such mechanisms - auctions & matching models.

Firstly, auctions employ sophisticated price-based protocols to allocate scarce resources, with foundational work establishing how to design auctions that maximize revenue or efficiency (Vickrey, 1961). Secondly, matching models address allocations where monetary transfers are infeasible, unethical, or undesirable. The objective here is to pair participants based on declared preferences to achieve not only efficient, but also stable outcomes.

By the 1990s, economists recognized the synergy between mathematical optimization and game-theoretic mechanism design (Roth, 2002), which provided a rigorous computational foundation for solving complex multi-object allocation problems, defined as the optimization of a linear objective function subject to a set of constraints. The National Resident Matching Program (NRMP), which uses the Roth-Peranson algorithm to place more than 40,000 medical trainees annually (Roth & Peranson, 1999), stands as a pivotal application of this perspective. This accentuates the economist's role as a practical market engineer to address existing market failures or construct entirely new market infrastructures (Roth, 2007).

That being said, as mentioned in the introduction, market design principles are engineered for *Homo economicus*, or "Econs" - agents who consistently maximize their quasilinear utility and react rationally to incentives. This assumption underpins strategy-proof mechanisms designed to achieve efficient resource allocation, such as the Vickrey-Clarke-Groves auctions and stable matching algorithms. However, empirical research in behavioural economics, a discipline that integrates insights from psychology with economic theory, reveals that real consumers are influenced by bounded rationality, cognitive biases and emotional heuristics (Thaler & Sunstein, 2008). The first significant departure from the idealized model of the "Econ" was articulated by Simon (1955), who introduced the concept of "bounded rationality." Simon argued that in a world of limited information, cognitive capacity, and time constraints, humans do not, and cannot, perform the god-like optimization assumed by classical models. Instead, they "satisfice", where they seek a solution that is "good enough," thereby replacing the principle of optimization with a heuristic. This insight established the initial problem: if agents are not fully rational, their psychological and cognitive factors become a first-order concern for economic models.

This foundation was significantly advanced by Kahneman and Tversky's (1979) Prospect Theory, which established that preferences are systematically influenced by reference points rather than absolute

outcomes. They also found that individuals typically experience losses about twice as strongly as comparable gains due to loss aversion.

Nevertheless, it was economist Richard Thaler who linked these psychological insights to economics, formalizing the field of behavioural economics. Thaler (1980) documented the endowment effect, which shows that individuals ask for much higher compensation to give up items they own than they would pay to acquire them. Furthermore, his work on mental accounting demonstrated that humans violate the core economic principle of fungibility. Thaler (1985) showed that individuals categorize and treat money non-fungibly based on its source or intended use, constraining spending within mental budget categories even when this produces suboptimal outcomes, a behaviour inconsistent with utility maximization.

Additional heuristics compound these effects. Tversky and Kahneman (1974) identified the availability heuristic, whereby individuals overweigh information that comes to mind easily when making judgments. Choice overload hinders optimal decision-making; Lyengar and Lepper (2000) found that consumers presented with extensive product assortments were 10% as likely to make purchases compared to those facing limited selections.

This body of research - from Simon's satisficing to Tversky's and Kahneman's heuristics and Thaler's formalizations - has now culminated in the applied field of choice architecture, or Nudge theory (Thaler & Sunstein, 2008). The core insight is that cognitive friction makes the design of the decision-making context a critical determinant of behavioural outcomes. For instance, Johnson and Goldstein (2003) found in their study of organ donation across Europe, countries with an opt-in system had consent rates around 12-25%, while opt-out countries achieved rates of 85-99.9%.

These cognitive biases are not marginal deviations but fundamental characteristics of human judgment and decision-making. Consequently, market mechanisms designed for idealized rational actors often produce outcomes that diverge from theoretical predictions when deployed in practice, underscoring the need to design mechanisms that accommodate rather than ignore behavioural realities.

This literature review thus establishes the central synergy and unresolved tension informing this paper. Market design provides the rigorous, mathematical, and institutional framework essential for building large-scale allocation systems. Its tools, derived from mechanism design and optimization, provide a blueprint for ad auctions and content-matching algorithms, designed to achieve efficient outcomes under rational assumptions. Conversely, behavioural economics provides a critical lens, exposing the psychological limitations and real-world complexities that govern how actual users interact with these systems. Understanding both perspectives - the logic of the system's architecture and the psychology of its users - is therefore essential. The intersection of these two disciplines is necessary to critically evaluate how efficiency, engagement, and fairness in the streaming industry could be improved.

DEFINITIONS

Before analysing streaming platforms through this dual lens, it is necessary to define the three evaluative objectives that structure this paper's analysis - efficiency, engagement and fairness - with adequate

precision to assess whether proposed mechanisms constitute genuine improvements.

Efficiency is defined here as retention-weighted allocative efficiency, which is the platform's capacity to maximise total expected surplus across users, advertisers and creators, discounted by the probability that each user remains on the platform. This definition does not frame efficiency as simply revenue-maximisation, because a platform that extracts maximum immediate advertising revenue by over-serving advertisements to weariness-prone users may appear efficient in the short-run, while destroying the retention base that sustains long-run value.

Engagement is defined as sustained engagement, which distinguishes between short-term and long-term engagement. Short-term engagement refers to immediate signals such as click-through rates, session length and content completion that platforms routinely optimize for because they are directly observable. Long-term engagement refers to continued platform use over time, captured by 30-day retention rates and subscription renewal behaviour. As the analysis of ad fatigue reveals, mechanisms that inflate short-term engagement metrics through repeated ad exposure or algorithmically addictive content recommendations can simultaneously erode long-term engagement by inducing weariness and psychological reactance. Throughout this paper, proposals are evaluated against long-term engagement as the primary criterion.

Fairness is defined across two dimensions, reflecting the two-sided structure of streaming markets. On the user side, fairness refers to equitable exposure to diverse content. A recommendation system that concentrates exposure on a narrow band of popular or platform-preferred content is unfair to users whose genuine preferences lie outside that band. On the creator side, fairness refers to equitable reach distribution across content producers, assessed by whether the distribution of algorithmic exposure across creators can be improved for worse-off creators without harming better-off ones, which is a criterion Do et al. (2021) formalise through Lorenz dominance.

These concepts of efficiency, engagement and fairness could oftentimes, be at odds with each other. For instance, redistributing recommendation exposure towards underserved creators may marginally reduce recommendation accuracy, creating a tension between fairness and efficiency. The reforms proposed in this paper do not resolve these tensions, but they make the tradeoffs explicit and propose design choices with outcomes that are more sustainable across all three dimensions than the status quo.

METHODOLOGY

This paper adopts a literature-based design approach. Rather than generating original empirical data, it synthesises existing research across market design, behavioural economics, and digital advertising to construct and evaluate mechanisms for streaming platform allocation. A mechanism is treated as an improvement if it satisfies two conditions. Firstly, it advances at least one of the three objectives of efficiency, engagement and fairness without worsening another, and secondly, it is supported by empirical evidence from at least one peer-reviewed study in the relevant domain.

Source selection prioritises peer-reviewed journal articles and primary technical reports, including published research from Netflix and Spotify's engineering teams. Industry publications and blog-format

explainers are used only for descriptive context and are not treated as evidence for causal or quantitative claims. Where multiple sources address the same phenomenon, priority is given to the most methodologically rigorous account.

The paper proceeds in two analytical stages. The first examines ad auctions on ad-supported platforms such as YouTube, drawing on market design principles and behavioural research to identify where the standard Generalised Second-Price mechanism diverges from real user behaviour, and deriving a behaviourally adjusted framework as a proposed correction. The second examines recommendation systems on subscription platforms such as Netflix and Spotify, using the same dual lens to identify market distortions and propose structural reforms.

It is important to note that the mechanisms proposed here are design arguments, not empirical findings. Claims are presented as implications of the framework rather than demonstrated outcomes, and the limitations of this approach are addressed in the conclusion.

THE MARKET DESIGN OF AD AUCTIONS IN THE STREAMING INDUSTRY

Ad-supported streaming platforms, such as YouTube, rely on advertising auctions as their economic backbone. As the world's largest ad-supported platform, YouTube generates most of its revenue through Generalized Second-Price (GSP) auctions (Agius, 2023), ranking advertisers by an Ad Rank that combines bid price with ad quality. In isolation, these auctions aim to maximize both allocative efficiency and platform revenue by awarding slots to the most relevant high bidders. Yet YouTube's mechanisms must account for constraints such as viewer tolerance, engagement, and retention. Systematic biases such as ad fatigue, psychological reactance, and loss aversion play a substantial role in shaping these outcomes and, consequently, the platform's long-term sustainability. Therefore, OTT platforms face a critical two-sided design challenge of maximizing advertiser revenue without triggering user churn.

The wearout pattern - where each additional ad impression creates diminishing but positive marginal returns - is generally well-known. This is also confirmed by Chae et al. (2018), who estimate a homogeneous response function.

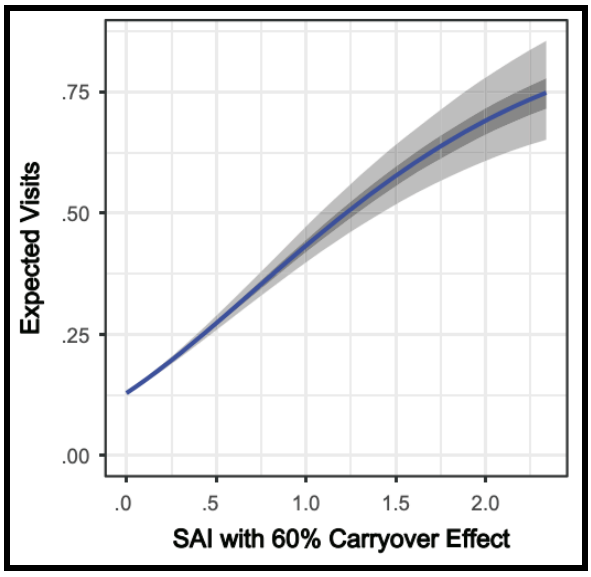

Fig 1. Response shape when all viewers are categorised into one class

The Smoothed Advertising Index (SAI), as seen in Figure 1, captures how recent ad exposures continue to influence current behaviour through a carryover effect. The carryover effect (set at 60% in their baseline specification) means that an exposure today still contributes 60% of its "attention weight" tomorrow, 36% the next day, and so forth. Formally, for individual *i* at time *t*, exposed to publisher *j*, the SAI is:

$$\mathrm{SAI}_{it} = \sum_{\tau<t} \lambda^{t-\tau} y_{i\tau},$$

where $0 < \lambda < 1$ is the carryover parameter, and y_{iT} is the number of impressions previously seen. This metric allows the model to detect both wearout (positive but diminishing marginal effects) and weariness (where the response function obtains a negative slope beyond a threshold level of exposure).

A central complication in applying market design principles to streaming platforms is that viewers do not respond uniformly to repeated advertising; instead, Chae et al. demonstrate, through empirical patterns, that assuming homogeneity eliminates evidence of weariness, in which the response function has a negative slope beyond a threshold level of exposure. Chae et al. estimate a fully Bayesian hierarchical mixture model, assigning each individual to one of five latent classes based on shared parameter values governing how SAI and timing affect the probability of visiting the advertiser's website. These classes reflect behavioural differences in (a) browsing frequency, (b) breadth of websites visited, and (c) topical alignment with the advertiser. For instance, as seen in Figure 2, Class 5 - the heaviest internet users - shows quasilinear, strongly positive responses, while Class 2 displays the clearest evidence of weariness, with response curves peaking at approximately three impressions within two days before turning negative. Class 1 consists of low-engagement users with negligible responsiveness; Classes 3 and 4 occupy

intermediate positions, with Class 3 exhibiting some weariness and Class 4 displaying mild wearout but no downturn.

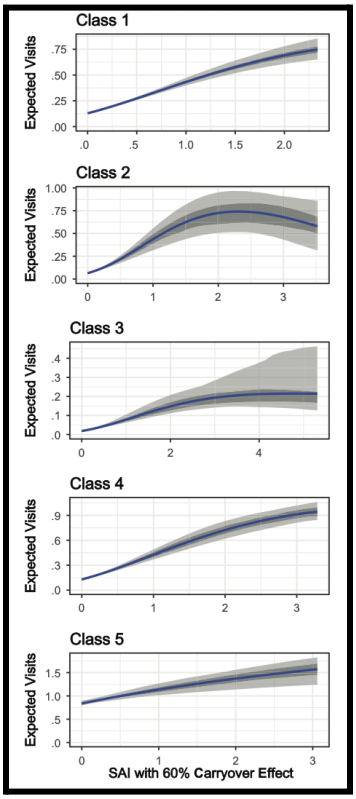

Fig 2. Ad Response Contours when viewers are categorised into different classes

The emergence of weariness in Class 2 - and, to a lesser extent, Classes 1 and 3 - reflects the fact that ad response is fundamentally shaped by behavioural biases that systematically violate the assumptions of rational choice embedded in classical auction theory.

Firstly, ad fatigue plays a central role. Cognitive Load Theory posits that humans have a limited capacity for attention. When advertisements interrupt primary tasks, such as watching content, they impose unnecessary cognitive load, which depletes working memory and reduces engagement with both the ads and the content (Paas et al., 2011). In their study involving 100 participants, Khatwani & P.R. (2025) demonstrated that a high-load condition, compared with a low-load condition, negatively affected all measured behavioural outcomes, including a substantial reduction in ad recall, less favourable viewer attitudes, and a pronounced drop in expressed purchase intentions. Additionally, empirical research in digital environments finds that algorithmically over-personalized or repetitive advertising leads to emotional exhaustion, increased irritation, and deliberate avoidance (e.g, skipping, muting, or ad-blocking), which are all hallmark indicators of ad fatigue (Stefanos Balaskas et al., 2025).

Another foundational behavioural mechanism is loss aversion, a core insight of prospect theory

(Kahneman & Tversky, 1979). From the viewer's perspective, each interruption by an unwanted ad represents a loss of uninterrupted viewing time. Since individuals are psychologically more sensitive to losses than to equivalent gains, the disutility of additional ads can increase more rapidly than the utility of attention or relevance, resulting in a negative slope in the response function at high exposure levels.

An additional mechanism accelerating weariness is psychological reactance - the negative effect that arises when individuals perceive their freedom to be restricted (Rosenberg & Siegel, 2025). In streaming, repetitive or unskippable ads are interpreted not simply as interruptions, but as constraints on user autonomy. Advertising studies show that ads perceived as forced or intrusive sharply increase irritation, reduce ad effectiveness, and lead to active avoidance behaviours such as skipping or abandoning the platform (Lin et al., 2021).

Moreover, demographic heterogeneity further amplifies these effects. One example is that younger audiences tend to be more tolerant of digital advertising than older generations (Walsh et al., 2024). According to Cashdollar et al. (2013), older viewers spend more time viewing non-central screen elements and take longer to disengage from unexpected visual interruptions. As a result, older adults are more likely to experience weariness because they have a lower ability to focus and incur greater attentional costs from surrounding stimuli.

The issue of weariness requires platforms to view ad allocation as a constrained optimisation problem, in which they maximise advertiser revenue while accounting for user-level psychological constraints. YouTube already uses several tools to manage user experience - GSP auctions that factor in ad quality and basic frequency caps. These features may alleviate irritation but rely on general rules and lack models of underlying behavioural types and predictive thresholds to determine when a viewer is about to experience weariness.

THE MARKET DESIGN OF MATCHING MECHANISMS IN THE STREAMING INDUSTRY

Another key role of streaming platforms is to connect millions of users with millions of content options in ways that satisfy both sides of the marketplace. Thus, recommendation systems function as matching mechanisms and shape welfare outcomes for users and creators. Unlike classical two-sided matching markets in the Gale-Shapley sense where both sides submit preferences and stability is the design objective, recommendation systems here are used as a metaphor for attention allocation. The platform allocates finite user attention across the content catalogue. Understanding recommendation systems requires recognizing that their primary function is to resolve the constraint of information abundance - a condition where the bottleneck is the user's finite cognitive capacity, not the content supply.

Behavioural constraints are crucial here. Research on the paradox of choice shows that an excessively large number of options reduces both the frequency of choice and overall satisfaction (Zhang et al., 2023). In this context, users avoid the cognitive cost of browsing huge catalogs. Data from Netflix indicate that only 20% of viewing starts with a search, whereas the remaining 80% is driven by algorithmic recommendations (Gomez-Uribe & Hunt, 2015). As a result, recommendation systems manage attention

by limiting the range of choices and lessening the burden of decision-making to improve efficiency. However, this situation offers the platform the power to decide what users can see, thus influencing the matches that occur.

Hence, the algorithm is actively shaping choice by structuring the environment, where the user's cognitive shortcuts are integral to the platform's allocative success as users attempt to 'satisfice' rather than engage in exhaustive search. It thus structures the environment through three layers.

Layer 1 - Candidate Generation Through Collaborative and Content-Based Filtering: The first layer of the matching pipeline involves reducing the number of items from millions to a manageable set of plausible candidates. Two foundational techniques, collaborative filtering (CF) and content-based filtering (CBF), are used in this process (GeeksforGeeks, 2025).

Collaborative filtering, such as Spotify's 'Fans Also Like', leverages the logic of social proof. It constructs a large user-item interaction matrix and uses mathematical techniques such as matrix factorization to infer latent preferences and recommend items consumed by users with similar taste profiles (Murrell, 2025). Its strength lies in discovering unexpected connections, but it is inherently limited by the "cold-start problem" - it cannot effectively recommend content or creators lacking historical engagement data (James, 2025).

Content-based filtering, conversely, relies on item attributes. Platforms such as Netflix employ advanced methods, including deep learning for visual aesthetics, to generate thousands of micro-genres (Steck et al., 2021), while Spotify analyzes raw audio waveforms using convolutional neural networks to create "sonic profiles" (Falah et al., 2025). The value of CBF lies in its ability to serve new items and users immediately. However, its limitation is the risk of creating a filter bubble by endlessly reinforcing a single, narrow taste, thus limiting user discovery.

Real systems utilize hybrid approaches to maximize both accuracy and diversity. Spotify's flagship feature, 'Discover Weekly', exemplifies this by blending CF, natural language processing (for analyzing web text about music), and audio models to create three independent recommendation sources (Maheshwari, 2023), which can then be combined to address the cold-start problem and introduce novelty.

Layer 2 - Ranking & Optimization: The second layer, Ranking and Optimization, is where the platform's objectives are realized as the matching mechanism's reward function. Once candidate items are generated, they are assigned a score and ranked by predicted value to a user.

This process is fundamentally an optimization problem. For each user, the algorithm must assign a score to each item based on multiple, sometimes conflicting, objectives such as retention, usage duration on the platform, user satisfaction, and generated revenue (Pastukhov, 2025). Platforms manage the inherent exploration-exploitation trade-off using contextual bandit algorithms that systematically decide when to recommend a "sure bet" (exploitation) and when to surface a less

"sure" or less familiar item to gather new data and test content (exploration) (McInerney et al., 2018). The platform, in setting the parameters of this bandit, effectively decides how much exploration to permit.

The ranking layer defines the reward function of the matching system; thus, the fundamental quality and nature of the realized content matches depend on it. For instance, if the platform optimizes primarily for short-term metrics like immediate click-through rate, it will favour instantly "addictive" or easily consumed content, potentially disfavouring slow-burning, high-quality, or niche content that contributes to long-term user satisfaction.

Layer 3 - Presentation**:** The final layer of the matching pipeline is presentation, an aspect of market-design that fundamentally shapes which matches are realized. This is achieved through several key mechanisms.

Firstly, position bias refers to the prominent placement of items in top rows, like Netflix's Top Picks. These items receive disproportionately higher engagement. By controlling these positions, the platform directly controls high-value match opportunities.

Secondly, hierarchical matching refers to organizing content into genre rows or categories. This addresses the cognitive overload problem by narrowing the user's immediate choice set while simultaneously pre-filtering user exposure.

Lastly, defaults are features such as Netflix's autoplay or Spotify's playlist continuation, leveraging status quo bias to set the path of least resistance and steer sequential consumption.

While recommendation systems enhance efficiency by mitigating cognitive load, they simultaneously introduce market distortions and welfare concerns regarding fairness and discovery. Firstly, the heavy reliance on past behaviour creates algorithmic lock-in, often resulting in filter bubbles or recommendation chambers. When the ranking algorithm optimizes solely for predictive accuracy, it disproportionately recommends content similar to what the user has already consumed, leading to a homogeneous experience. This tendency has been shown to reduce sales diversity, as market structure favours already popular items, thereby diminishing users' exposure to niche or independent content (Fleder & Hosanagar, 2009). This loss of serendipity undermines long-term user welfare by limiting cultural and informational discovery.

Secondly, a major conflict of interest arises from the platform's dual role as both matchmaker (recommender) and market participant (content producer) when markets are vertically integrated. Evidence suggests that platforms can engage in strategic recommendation bias, steering users toward content that maximizes platform profit rather than pure user utility. For instance, a platform may intentionally bias rankings toward its own originals or content with cheaper royalty rates to reduce costs, a practice that can erode the market power of external content providers (Bourreau & Gaudin, 2018). When the system pushes platform-preferred matches over user-optimal matches, the incentives for independent creators are weakened, and the supply of external content is jeopardised.

Hence, to operationalise weariness in ad auctions, platforms can extend the standard ad-ranking rule by incorporating a behavioural adjustment grounded in advertising and behavioural economics research. Ads are ranked by an effective score

$$S_{i,u} = b_i \cdot q_i \cdot T(u),$$

where b_i is advertiser i's bid and q_i is an ad quality term, consistent with GSP mechanisms used by platforms such as YouTube. The key addition is $T(u)$, a user-specific tolerance multiplier that down-weights ads for users likely to experience weariness. Drawing from the Smoothed Advertising Intensity framework in Chae, Bruno, and Feinberg (2019), cumulative exposure is defined as

$$\mathrm{SAI}_u = \sum_{\tau<t} \lambda^{t-\tau} y_{u\tau},$$

where $0 < \lambda < 1$ captures memory carryover from past impressions. Behavioural evidence on ad fatigue and psychological reactance (Brehm, 1966; Li, Edwards & Lee, 2002) motivates a declining tolerance function

$$T(u) = \exp(-\alpha_k \cdot \mathrm{SAI}_u),$$

where α_k is class-specific, reflecting heterogeneous sensitivity to repeated ads as identified by the latent-class model of Chae et al. (2019).

Beyond ranking, platforms must decide whether to show an ad at all, which requires reconciling immediate auction revenue with long-run user value. The platform's objective is to maximize a composite payoff function that accounts for both immediate monetization and user retention. Following (Ascarza et al., 2017), who frame customer-centric decision-making around expected lifetime value, the platform chooses ad delivery to maximize:

$$V_{i,u} = R_{i,u} + \gamma \cdot E[V_u(t)] \cdot S(t|\mathrm{SAI}_u) \cdot T(u),$$

Here, i indexes the advertiser whose ad is under consideration, and u indexes the user being served. $R_{i,u}$ refers to the immediate auction revenue. The expected lifetime value term $E[v_u(u)]$ depends only on u, as it represents the stochastic present value of future transactions with user u, where the expectation is taken over u's uncertain future retention path. The retention probability $S(t \mid SAI_u)$ conditions this expectation on the user's current smoothed advertising intensity, capturing the empirical relationship between cumulative

ad exposure and churn risk. *T(u)* is a user-specific tolerance multiplier that decays with cumulative ad exposure. The discount factor $0 < \gamma < 1$ converts future value to present terms.

Churn management is central to this formulation, whereby retention campaigns must balance targeted interventions against contact fatigue. Their framework illustrates how proactive ad delivery decisions involve fundamental tradeoffs: showing an ad yields immediate revenue $R_{i,u}$ but carries a latent cost if exposure increases churn risk through cumulative weariness, manifested in both the declining tolerance multiplier *T(u)* and reduced retention probability $S(t \mid SAI_u)$.

Thus, ads are withheld once the expected marginal benefit of immediate revenue is outweighed by the retention cost imposed by excess cumulative exposure. The decision threshold is:

$$\text{Show ad if } R_{i,u} > -\gamma \cdot E[V_u(t)] \cdot \frac{\partial[S(t|\text{SAI}_u) \cdot T(u)]}{\partial \text{SAI}_u},$$

The partial derivative is taken with respect to SAI_u, holding γ, $E[v_u(u)]$, and advertiser-specific terms constant. It is negative by construction since both $S(t \mid SAI_u)$ and *T(u)* are decreasing in SAI_u, making the right-hand side positive. Therefore, the decision threshold is when the present value of future transactions exceeds the lifetime value loss from incremental weariness. This ensures that allocative efficiency aligns with user retention, psychological credibility, and platform long-run profitability, formalising how classical auction mechanisms can internalise behavioural constraints that standard revenue-maximising designs ignore.

However, this shift toward behavioural alignment must extend beyond the commercial layer of advertising to the algorithmic matching of content itself. The ultimate goal for streaming platforms is to move beyond myopic engagement maximization toward a system that enhances efficiency, engagement, and fairness across the two-sided market. Below are three evidence-based structural reforms.

A. Diversity-Aware Ranking

It is essential to find a balance between personalized and diversified recommendations. By incorporating explicit diversity constraints into the ranking function, platforms can ensure a minimum threshold of variety in aspects such as genre and content age among the top recommendations. These constraints can be represented as a bi-objective optimization problem, in which the solution space consists of a Pareto-optimal front of non-dominated solutions. In this setup, increasing personalization tends to reduce diversity and vice versa (Areeb et al., 2023). Diversity constraints could significantly improve the quality of choice and broaden exposure, thereby favouring long-term user welfare by mitigating filter bubbles and addressing the cold-start problem.

B. Concave Welfare Functions

To foster a healthier content supply, creator welfare needs to be prioritised alongside user utility. The well-being of creators depends on the reach of their content and the economic incentives they derive from

the platform, both of which are at risk due to recommendation bias. Thus, platforms could adopt algorithmic fairness frameworks to ensure a fair distribution of exposure. Do et al. (2021) propose that concave welfare functions, by assigning diminishing marginal weight to the utility gains of already well-exposed creators, increase the ranking priority of underexposed creators while maintaining recommendation quality at lower computational cost than constraint-based methods. By targeting the Lorenz dominance criterion, this approach ensures that the distribution of creator exposure in the revised system unambiguously dominates that of the status quo across all quantiles. By targeting not only user utility but also creator welfare, fair ranking algorithms ensure long-term ecosystem sustainability while being economically viable.

C. User Controls to Break Algorithmic Lock-In

Users must be empowered by platforms with explicit governance tools that enhance their agency, thereby counteracting algorithm lock-in. Examples of such tools include a "Diversity Slider" to control the balance between predictive accuracy and content novelty, and a "Preference Reset Mode" that allows users to temporarily de-weight past consumption history. These mechanisms could mitigate the formation of filter bubbles and foster more authentic preferences.

By adopting this behaviourally adjusted matching framework, platforms transform into a true, welfare-enhancing marketplace rather than limiting user discovery. This thus ensures a richer, more diverse, and more sustainable content ecosystem.

CONCLUSION

As matching models and advertising auctions play a key role in how streaming platforms operate, examining their dependence on rational-choice assumptions and the impact of behavioural deviations has become essential. This paper explores how market design and behavioural economics could work together to increase efficiency, engagement, and fairness in the streaming industry, and argues that integrating mathematical optimization with behavioural insights produces mechanisms that are both economically feasible and psychologically credible.

This paper highlighted that, in the context of ad-supported platforms such as YouTube, current Generalized Second-Price (GSP) auctions largely overlook the carryover effects of advertising and how these effects vary across users. While some users may display quasi-linear engagement, others reach a tipping point of weariness at certain thresholds, which can be attributed to high cognitive load, loss aversion, and psychological reactance. Furthermore, these thresholds are not universal; they vary with latent engagement classes and age-related attentional costs. Thus, streaming platforms that do not account for these effects may experience diminished engagement and potential platform churn. Likewise, the study of recommendation systems beyond ad auctions on platforms such as Netflix and Spotify revealed a key issue. The three-layered processes for generating, ranking, and presenting recommendations help users manage the overwhelming number of options available to them, but they can also introduce market distortions such as algorithmic lock-in and recommendation bias. Streaming platforms risk stifling genuine user discovery and prioritising their own profit margins and vertical integration over fairness for

creators. Therefore, these findings suggest that, to ensure long-term ecosystem sustainability, streaming platforms must transition from simple optimisation to a constrained optimisation model that accounts for users' finite cognitive capacity and emotional heuristics.

The proposed behavioural adjustments to the ad-ranking rule and payoff function - incorporating user-specific tolerance multipliers and lifetime value optimization - enable ad auctions to internalize psychological realities like weariness. Structural reforms such as diversity-aware ranking, concave welfare functions, and user controls to break algorithmic lock-in further help counteract filter bubbles and restore creator fairness in the face of recommendation bias by leveraging the synergy between behavioural economics and market design principles.

Several limitations, however, bound the scope of these contributions. The class-specific decay parameter α_k is drawn from Chae et al.'s (2019) estimates on display advertising data; whether these behavioural classes and sensitivity values transfer to a streaming context, where content consumption patterns and ad interruption dynamics differ substantially from banner advertising, remains an open empirical question requiring re-estimation on platform-specific data. Similarly, the retention probability function $S(t \mid SAI_u)$ linking cumulative ad exposure to churn risk requires proprietary user-level data unavailable to this study. A follow-up would involve an A/B test varying frequency caps across user behavioural classes on an ad-supported streaming platform, enabling estimation of the parameter α_k and validation of the decision threshold against observed 30-day retention outcomes. Moreover, while Areeb et al. (2023) and Do et al. (2021) provide quantitative support for diversity-aware ranking and concave welfare functions respectively, the proposed user-facing controls lack direct empirical validation for their quantitative effect on consumption patterns or long-term user welfare in a real streaming environment. A natural next step could be a field experiment or simulation study measuring and quantifying the welfare effects of user-facing controls on filter bubble formation and creator exposure distribution.

Ultimately, the synthesis of fields of behavioural economics and market design principles indicates that for an allocation system to be truly efficient, fair, and engaging for both users and creators, it must move beyond the idealized model of the rational 'Econ' to accommodate the psychological heuristics and cognitive constraints that govern real user interaction.

REFERENCES

Abdulkadiroglu, A., & Sönmez, T. (2003). *School Choice: A Mechanism Design Approach*. https://www.tayfunsonmez.net/wp-content/uploads/2013/10/AbdulkadirogluSonmez-AER2003.pdf

Agius, N. (2023, October 13). *What is RGSP? Google's Randomized Generalized Second-Price ad auctions explained*. Search Engine Land. https://searchengineland.com/google-rgsp-randomized-generalized-second-price-ad-auctions-explained-433053

Areeb, Q. M., Nadeem, M., Sohail, S. S., Imam, R., Doctor, F., Himeur, Y., Hussain, A., & Amira, A. (2023). Filter bubbles in recommender systems: Fact or fallacy—A systematic review. *WIREs Data Mining and Knowledge Discovery*, *13*(6). https://doi.org/10.1002/widm.1512

Ascarza, E., Fader, P. S., & Hardie, B. G. S. (2017). Marketing Models for the Customer-Centric Firm. *International Series in Operations Research & Management Science*, *254*, 297–329. https://doi.org/10.1007/978-3-319-56941-3_10

Bichler, M. (2017). *Market Design - A Linear Programming Approach to Auctions and Matching*. Cambridge University Press. https://doi.org/10.1017/9781316779873

Bourreau, M., & Gaudin, G. (2018). Streaming Platform and Strategic Recommendation Bias. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3290617

Briel, R. (2025, August 20). *US Streaming platforms shift focus to retention as churn rates surge*. Broadband TV News. https://www.broadbandtvnews.com/2025/08/20/us-streaming-platforms-shift-focus-to-retention-as-churn-rates-surge/

Cashdollar, N., Fukuda, K., Bocklage, A., Aurtenetxe, S., Vogel, E. K., & Gazzaley, A. (2013). Prolonged disengagement from attentional capture in normal aging. *Psychology and Aging*, *28*(1), 77–86. https://doi.org/10.1037/a0029899

Cai, Y., Feng, Z., Liaw, C., Mehta, A., & Velegkas, G. (2024). User Response in Ad Auctions: An MDP Formulation of Long-term Revenue Optimization. *Proceedings of the ACM Web Conference 2024*, 111–122. https://doi.org/10.1145/3589334.3645495

Chae, I., Bruno, H. A., & Feinberg, F. M. (2018). Wearout or Weariness? Measuring Potential Negative Consequences of Online Ad Volume and Placement on Website Visits. *Journal of Marketing Research*, *56*(1), 57–75. https://doi.org/10.1177/0022243718820587

Do, V., Corbett-Davies, S., Atif, J., & Usunier, N. (2021). Two-sided fairness in rankings via Lorenz dominance. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2110.15781

Falah, N., Yousefimehr, B., & Ghatee, M. (2025, June 11). *Predicting Music Track Popularity by Convolutional Neural Networks on Spotify Features and Spectrogram of Audio Waveform*. Arxiv.org. https://arxiv.org/html/2505.07280v1

Fleder, D., & Hosanagar, K. (2009). Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Management Science*, *55*(5), 697–712. https://doi.org/10.1287/mnsc.1080.0974

GeeksforGeeks. (2025, July 12). *Collaborative Filtering in Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/collaborative-filtering-ml/

Gomez-Uribe, C. A., & Hunt, N. (2015). The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems*, *6*(4), 1–19. https://doi.org/10.1145/2843948

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, *79*(6), 995–1006. https://doi.org/10.1037//0022-3514.79.6.995

James, P. (2025, February 25). *What is the Cold Start Problem in Recommender Systems?* FreeCodeCamp.org. https://www.freecodecamp.org/news/cold-start-problem-in-recommender-systems/

Johnson, E. J., & Goldstein, D. G. (2003, November 21). *Do Defaults Save Lives?* Papers.ssrn.com. https://ssrn.com/abstract=1324774

Kahneman, D., & Tversky, A. (1979). Prospect theory: an Analysis of Decision Under Risk. *Econometrica*, *47*(2), 263–292. https://doi.org/10.2307/1914185

Khatwani, M., & P.R, K. (2025). Cognitive Load and Advertisement Effectiveness: Neurocognitive Perspective. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5443775

Lin, H. C.-S., Lee, N. C.-A., & Lu, Y.-C. (2021). The Mitigators of Ad Irritation and Avoidance of YouTube Skippable In-Stream Ads: An Empirical Study in Taiwan. *Information*, *12*(9), 373. https://doi.org/10.3390/info12090373

Maheshwari, C. (2023). *Music Recommendation on Spotify using Deep Learning* (p. arXiv). https://arxiv.org/pdf/2312.10079

McInerney, J., Lacker, B., Hansen, S., Higley, K., Bouchard, H., Gruson, A., & Mehrotra, R. (2018, October 2). *Explore, Exploit, Explain: Personalizing Explainable Recommendations with Bandits | Spotify Research*. Spotify Research. https://research.atspotify.com/publications/explore-exploit-explain-personalizing-explainable-recommendations-with-bandits

Murrell, T. (2025, June 17). *Matrix Factorization: The Bedrock of Collaborative Filtering Recommendations | Shaped Blog*. Shaped.ai. https://www.shaped.ai/blog/matrix-factorization-the-bedrock-of-collaborative-filtering-recommendations

Paas, F., Renkl, A., & Sweller, J. (2011). *Cognitive Load Theory : a Special Issue of educational Psychologist*. Springer Science+Business Media, LLC.

Papathanasopoulos, A., & Varoutas, D. (2024). On the competition between Video OTT platforms vs Traditional TV: A niche case study in Greece. *Telematics and Informatics Reports*, *16*, 100166. https://doi.org/10.1016/j.teler.2024.100166

Pastukhov, D. (2025, September 1). *Inside Spotify's Recommendation System: A Complete Guide (2025 Update)*. Music-Tomorrow.com. https://www.music-tomorrow.com/blog/how-spotify-recommendation-system-works-complete-guide

Persky, J. (1995). Retrospectives: The Ethology of Homo Economicus. *Journal of Economic Perspectives*, *9*(2), 221–231. https://doi.org/10.1257/jep.9.2.221

Rosenberg, B. D., & Siegel, J. T. (2025). Psychological reactance theory: An introduction and overview. *Motivation Science*, *11*(2), 133–138. https://doi.org/10.1037/mot0000376

Roth, A. E. (2002). The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics. *Econometrica*, *70*(4), 1341–1378. https://doi.org/10.1111/1468-0262.00335

Roth, A. E. (2007, October 1). *The Art of Designing Markets*. Harvard Business Review. https://hbr.org/2007/10/the-art-of-designing-markets

Roth, A. E., & Peranson, E. (1999). The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design. *American Economic Review*, *89*(4), 748–780. https://doi.org/10.1257/aer.89.4.748

Roth, A. E., Sönmez, T., & Ünver, M. U. (2004). Kidney Exchange. *The Quarterly Journal of Economics*, *119*(2), 457–488. https://doi.org/10.1162/0033553041382157

Sheynin, N., & Brophy, M. (2025). *The Future of Streaming Platforms: Key Trends and Outlook*. Alpha-Sense.com. https://www.alpha-sense.com/blog/trends/streaming-platforms-key-trends-and-outlook/#key-trends-in-streaming

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. https://doi.org/10.2307/1884852

Singh, A. (2024, October). *Gale-Shapley Algorithm Explained | Built In*. Built In. https://builtin.com/articles/gale-shapley-algorithm

Song, J. (2024). The Evolution and Impact of Streaming Services: Changing the Media Landscape. *Global Media Journal*, *22*(72). https://doi.org/10.36648/1550-7521.22.72.470

Steck, H., Baltrunas, L., Elahi, E., Liang, D., Raimond, Y., & Basilico, J. (2021). Deep Learning for Recommender Systems: A Netflix Case Study. *AI Magazine*, *42*(3), 7–18. https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/18140

Stefanos Balaskas, Konstantakopoulou, M., Ioanna Yfantidou, & Komis, K. (2025). Algorithmic Burnout and Digital Well-Being: Modelling Young Adults' Resistance to Personalized Digital Persuasion. *Societies*, *15*(8), 232–232. https://doi.org/10.3390/soc15080232

Tefertiller, A. (2020). Cable cord-cutting and streaming adoption: Advertising avoidance and technology acceptance in television innovation. *Telematics and Informatics*, *51*, 101416. https://doi.org/10.1016/j.tele.2020.101416

Thaler, R. (1980). Toward a Positive Theory of Consumer Choice. *Journal of Economic Behavior & Organization*, *1*(1), 39–60. https://doi.org/10.1016/0167-2681(80)90051-7

Thaler, R. (1985). Mental Accounting and Consumer Choice. *Marketing Science*, *4*(3), 199–214. http://www.jstor.org/stable/183904

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: improving decisions using the architecture of choice*. Yale University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. http://www.jstor.org/stable/1738360

Vickrey, W. (1961). COUNTERSPECULATION, AUCTIONS, AND COMPETITIVE SEALED TENDERS. *The Journal of Finance*, *16*(1), 8–37. https://doi.org/10.1111/j.1540-6261.1961.tb02789.x

Walsh, C., Chee-Read, A., Prouix, M., Linnehan, A., Venicio, B., & Harrison, P. (2024, September 13). *Consumer Insights: Advertising, US 2024 | Forrester*. Forrester.com.

https://www.forrester.com/report/consumer-insights-advertising-us-2024/RES181477

Zhang, G., Cao, J., & Dong, L. (2023). Examining the influence of information overload on consumers' purchase in live streaming: A heuristic-systematic model perspective. *PLOS ONE*, *18*(8), e0284466–e0284466. https://doi.org/10.1371/journal.pone.0284466