Designing Detection Algorithms for AI-Generated Content: Consumer Inference, Creator Incentives, and Platform Strategy

Jieteng Chen* T. Tony Ke† Jiwoong Shin‡

September 24, 2025

Abstract

Generative AI has transformed content creation, enhancing efficiency and scalability across media platforms. However, it also introduces substantial risks, particularly the spread of misinformation that can undermine consumer trust and platform credibility. In response, platforms deploy detection algorithms to distinguish AI-generated from human-created content, but these systems face inherent trade-offs: aggressive detection lowers false negatives (failing to detect AI-generated content) but raises false positives (misclassifying human-created content), discouraging truthful creators. Conversely, conservative detection protects creators but weakens the informational value of labels, eroding consumer trust. We develop a model in which a platform sets the detection threshold, consumers infer credibility from labels when deciding whether to engage, and creators choose whether to adopt AI and how much effort to exert to create content. A central insight is that equilibrium structure shifts across regimes as the threshold changes. At low thresholds, consumers trust human labels and partially engage with AI-labeled content, disciplining AI misuse and boosting engagement. At high thresholds, this inference breaks down, AI adoption rises, and both trust and engagement collapse. Thus, the platform's optimal detection strategy balances these forces, choosing a threshold that preserves label credibility while aligning creator incentives with consumer trust. Our analysis shows how detection policy shapes content creation, consumer inference, and overall welfare in two-sided content markets.

Keywords: Algorithmic detection, AI-generated content, Misinformation, Consumer inference, Platform design, Content moderation, Two-sided markets

*PhD Candidate in Marketing, Business School, Chinese University of Hong Kong, jieteng.chen@link.cuhk.edu.hk
†Professor of Marketing, Business School, Chinese University of Hong Kong, tonyke@cuhk.edu.hk.
‡Professor of Marketing, School of Management, Yale University, jiwoong.shin@yale.edu.

1 Introduction

Generative artificial intelligence (AI) is transforming content creation at an unprecedented scale, producing text, images, audio, and video with remarkable efficiency. From crafting marketing copy and product descriptions to social media posts and even contributing to news articles, generative AI greatly enhances efficiency and creativity. BuzzFeed, for example, has integrated AI-driven tools to generate personalized articles at scale, significantly boosting user engagement. Similarly, marketing firms, e-commerce platforms, and news agencies increasingly rely on AI to draft advertising copy, product descriptions, and news summaries.

However, the ease with which AI can generate realistic and persuasive content also introduces significant risks, especially the spread of misinformation. In practice, AI-generated misinformation has already demonstrated its potential to undermine public trust and distort online discourse. For instance, during the recent controversy surrounding Princess Kate Middleton, several AI-generated photographs circulated widely on social media, fueling public confusion and speculation before official clarifications emerged.[1] Such incidents highlight how AI-generated content poses a direct threat to consumer welfare by blurring the boundary between fact and fabrication, influencing consumer beliefs, and ultimately undermining consumer trust in online content and platform credibility.

In response to these concerns, platforms such as Facebook, Instagram, TikTok, and LinkedIn have begun deploying detection algorithms to distinguish human-created content from AI-generated content.[2] It is important to note that platforms typically cannot easily determine whether a piece of information is factually accurate or misinformation. Even when fact-checking is possible, it requires substantial human resources and time to verify, often allowing misinformation to spread before corrective action can be taken. Instead, these detection algorithms operate indirectly: they focus on identifying statistical patterns, linguistic features, or metadata that are characteristics of AI-generated output. These features may include stylistic inconsistencies, atypical word choice, or the absence of typical human writing nuances in text, as well as unnatural texture patterns, irregularities in lighting and shading, distorted facial features, or inconsistencies in background elements in images, artifacts commonly associated with the generative AI process. Therefore, rather than adjudicating the truthfulness of content, platforms act as gatekeepers, classifying and labeling content based on its likely origin. By labeling AI-generated content, platforms aim to

[1] *BBC News*, March 11, 2024 "Kate Photo withdrawn by five news agencies amid 'manipulation' concerns." https://www.bbc.com/news/uk-68526972

[2] See (1) https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/, (2) https://newsroom.tiktok.com/en-us/partnering-with-our-industry-to-advance-ai-transparency-and-literacy, (3) https://www.linkedin.com/help/linkedin/answer/a6282984.

inform consumers about the potential risks associated with AI-generated content.

A central challenge in deploying detection algorithms, essentially a classification problem, is balancing two types of errors: *false positives* and *false negatives.* A more aggressive detection policy reduces exposure to AI-generated misinformation but increases *false positives*, mistakenly flagging human-created content as AI-generated and potentially discouraging legitimate creators. A more conservative policy mitigates false positives but raises the risk of *false negatives*, failing to identify AI-generated content and potentially allowing misinformation to spread undetected. These trade-offs have important implications for platform strategy, content quality, and overall market outcomes. Moreover, detection policies influence not only direct classification outcomes but also consumer inferences about content authenticity. A conservative policy may foster widespread skepticism, even toward content labeled as human-created. Conversely, a more aggressive policy may preserve engagement with flagged posts if consumers view misclassifications as plausible. Through these mechanisms, detection policies shape not only the mechanical outcomes of classification, but also consumers' beliefs and strategic behavior across different detection regimes, ultimately affecting consumer engagement, content creation incentives, and overall welfare.

To analyze these issues, we develop a theoretical model of strategic interaction among three key players: consumers, content creators, and the platform. On the demand side, consumers derive utility only from high-quality truthful content. Content quality is directly observable when consumers view the content, but truthfulness is never verifiable. As a result, consumers rely on platform-generated labels as signals of credibility and decide whether to engage based on the observed quality and the label. In our model, engagement denotes the consumer's decision to allocate attention to view and interact with content. Once she engages, she consumes the content and realizes utility. Thus, we use the terms engagement and consumption interchangeably. This demand-side inference is central to shaping equilibrium outcomes on the platform. On the supply side, there are two types of content creators: "truthful" creators, who produce only truthful and authentic content, and "deceptive" creators, who generate fake content and misinformation. These types are fixed and unobservable. Creators make two key decisions: how much effort to exert in content creation, and whether to adopt AI tools. Exerting effort increases the likelihood of producing high-quality content that appeals to consumers but incurs a cost. AI tools reduce the cost of content creation by automating tasks such as drafting text, generating images, or synthesizing ideas. This benefit applies to both types of creators, but it also increases the likelihood of being flagged as AI-generated content by the platform's detection algorithm. Because misinformation can be made to appear high quality more easily since persuasive effects can be generated through emotional or sensational framing without the effort of ensuring factual accuracy, deceptive creators have an inherent advantage. We capture this asymmetry by assuming that, for any given effort level, deceptive creators are more likely to produce high-quality content. The platform, operating in a two-sided market, intermediates between creators and consumers using a probabilistic detection algorithm. Each piece of content is assigned a score reflecting the likelihood of being AI-generated based on observable features, with AI-generated content typically receiving higher scores. The platform then applies a threshold: content with the score above the threshold is labeled as AI-generated, while content below is labeled as human-created. This threshold governs the trade-off between false positives (mislabeling human content) and false negatives (failing to detect AI content). The platform sets the threshold to maximize a weighted summation of consumer surplus and creator profits. This framework allows us to characterize the platform's optimal detection policy while accounting for strategic behaviors on both sides of the market. We examine whether and when detection can mitigate the spread of misinformation, how AI/human labels shape consumer inference and engagement, and how platform detection strategy affects creator production effort and AI adoption decision. Finally, we explore how the optimal detection threshold changes as generative AI and detection technologies evolve.

We begin with a benchmark model without platform detection to isolate the strategic interaction between creators and consumers. Equilibrium behavior varies with the cost of AI adoption. When AI is either very cheap or very expensive, all creators choose ßidentical AI adoption strategies, leading to pooling equilibria where consumers engage with high-quality content only if the fraction of truthful creators is sufficiently high. However, when AI costs are moderate, strategic divergence emerges: truthful creators avoid AI and exert low effort, while deceptive creators adopt AI and exert high effort at reduced costs to exploit the ease of producing compelling misinformation. This separation results in consumer skepticism and a semi-separating equilibrium in which consumers engage with high-quality content only probabilistically.

We then extend the model to incorporate platform detection, where content is probabilistically labeled as AI- or human-generated. These labels influence consumer inference only when AI adoption differs across creator types, which arises under moderate AI costs, as in the benchmark case. In this regime, truthful creators do not adopt AI, but deceptive creators use AI probabilistically, making the AI label a noisy signal of misinformation. More importantly, detection supports a richer set of semi-separating equilibria, where the nature of the equilibrium shifts as the detection threshold changes. Under a low detection threshold (*aggressive*), AI-generated misinformation is less likely to be mislabeled as human-created (i.e., the false negative rate is low), making the human label a credible signal of authenticity. Consumers always engage with human-labeled content, but engage with AI-labeled content probabilistically—they sometimes engage and sometimes do not. We refer to this as the *semi-A* equilibrium, where the AI-labeled content is partially trusted. As the detection threshold rises (a more *conservative* policy), the equilibrium structure undergoes a regime shift: human-created content is now rarely mislabeled (i.e., the false positive rate is low), so the AI label becomes a stronger signal of misinformation. We refer to this regime as the *semi-H* equilibrium. In this *semi-H* regime, consumers avoid AI-labeled content entirely and engage with human-labeled content probabilistically. This endogenous shift in equilibrium structure is central to understanding how detection policy shapes outcomes.

Building on these equilibrium patterns, our analysis yields several key insights into how platform detection policy shapes market outcomes. First, the relationship between the detection threshold and consumer engagement is governed not by the threshold level itself, but by which equilibrium regime the platform induces. When the threshold lies in the *semi-A* region, the aggressive detection policy lowers false negatives, strengthens the informativeness of the human label, deters AI adoption by deceptive creators, and encourages effort from truthful creators, leading to higher consumer engagement. However, once the threshold crosses into the *semi-H* region (i.e., very high threshold), the conservative detection policy erodes label informativeness, raises AI adoption, and diminishes engagement. This highlights the importance of anticipating regime shifts in designing detection policies. Second, the mechanism driving these patterns is consumer inference. Detection labels influence how consumers assess content credibility, which in turn shapes their engagement decisions. Creators respond to these demand-side signals: when AI usage can be inferred from labels, deceptive creators face lower returns to AI adoption, reducing misinformation. But when detection becomes too aggressive or too conservative, inference breaks down and strategic misuse of AI intensifies. Third, these strategic feedback effects imply that detection conservativeness does not always benefit creators. We show that while a high threshold reduces misclassification risk of human-created content, it can also depress engagement and profits by weakening label credibility. Finally, the platform's optimal threshold balances these forces: it neither minimizes false positives nor false negatives alone, but simultaneously maximizes the informational value of labels to align creator incentives with consumer trust. As the detection accuracy improves, the platform can afford to relax its detection policy without sacrificing credibility; conversely, when AI becomes cheaper to use, more aggressive oversight is required to preserve engagement and welfare.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 presents the model. Section 4 analyzes the benchmark case without platform detection.

Section 5 characterizes the equilibrium of the main model and examines the platform's optimal detection strategy. Section 6 provides extensions, and Section 7 concludes.

2 Literature Review

Our research connects to multiple strands of literature at the intersection of algorithmic design, platform strategy, content moderation, and AI-generated content. First, our work relates to the literature on algorithm design and the economic implications of algorithms. Some work considers a principal's algorithm design problem in the presence of strategic agents who can manipulate the information provided to algorithm (Eliaz and Spiegler, 2019; Björkegren et al., 2020), while other research investigates the strategic interaction between multiple algorithms (Liang, 2019; Salant and Cherry, 2020; Montiel Olea et al., 2022) or between firms deploying algorithms, such as demand prediction algorithms (Miklós-Thal and Tucker, 2019; O'Connor and Wilson, 2021), and pricing algorithms (Calvano et al., 2020; Hansen et al., 2021; Klein, 2021; Brown and MacKay, 2023). In addition, current literature also explores the impact of predictive AI and recommender systems on user beliefs and behavior (Che and Hörner, 2018; Zhong, 2023; Zhou and Zou, 2023; Choi et al., 2024; Wang, 2025; Ning et al., 2025). In contrast to the recommendation setting, our detection algorithm does not steer consumer choice directly but instead alters beliefs about content authenticity, thereby shaping equilibrium inference, AI adoption, and effort decisions.

Our study also contributes to the literature on the firms' strategic design of algorithms and inherent tradeoffs. Cao et al. (2024) investigate the exploration and exploitation tradeoff in recommendation algorithms and show that a monopolistic firm has a stronger incentive to explore consumers' interests than competing firms. Iyer and Ke (2024) study the bias-variance tradeoff in model selection of algorithmic target advertising and find that competing firms adopt biased algorithms to soften competition. Closely related to our work, Iyer et al. (2024) examine the precision and recall tradeoff in targeting, conceptually analogous to the false positives-false negatives tradeoff: higher precision reduces false positives at the cost of lower recall. They show that firms favor high-precision but low-recall algorithms to reduce competition. We extend this line by modeling detection as a platform-level intervention: a probabilistic classifier that shapes equilibrium outcomes on two-sided markets through strategic consumer inference and creator behavior.

A growing literature examines content platforms hosting user-generated or decentralized content, often through the lens of two-sided markets. These studies explore how to influence content creation incentives by revenue sharing (Jain and Qian, 2021), recommendation (Qian and Jain, 2024; Zou et al., 2024), or promotion (Ren, 2024). Some research has shown that platforms can provide extra information and influence market outcomes through certification or endorsement (Hui et al., 2023; Bairathi et al., 2025; Chen et al., 2025). Our model contributes to this literature by showing how detection systems, through platform certificates such as labeling, act as powerful levers of platform control. Rather than removing content or banning users, the platform classifies content with varying error probabilities and influences both sides of the market: it affects creators' incentives to adopt AI and exert effort, and alters consumer beliefs about content credibility.

Our paper also enriches the broader literature on content moderation and misinformation detection. Prior work examines centralized versus decentralized enforcement (Wu, 2024; Chang et al., 2024), trade-offs between moderation and user growth (Liu et al., 2022), and balancing consumer engagement and ad revenue (Madio and Quinn, 2024). While much of this literature focuses on hard interventions like censorship or content removal, we highlight indirect governance through labeling. Rather than judging truth directly, platforms use imperfect classifiers to signal content origin, allowing consumers to interpret these signals strategically. Shin and Yu (2021) also study how imperfect binary signals influence belief updating and consumer demand. Relatedly, Yang et al. (2024) examine false positives and false negatives in a communication game, where a disinformation detector affects a sender's incentive to lie. This links to recent work on diagnostic testing by Dai and Singh (2025), who study how labs set diagnosis thresholds by weighing false positives against false negatives. In parallel, we show that platforms optimize detection threshold not for accuracy alone, but to manage the economic consequences of consumer inference and creator behavior in a two-sided market.

Finally, we contribute to the emerging literature on AI-generated content and its detection. Some works explore the capabilities of AI in areas such as content creation and personalized advertising (Zou et al., 2025; Kapoor and Kumar, 2025), while other work investigates the detectability of machine-generated reviews and their impact on trust and persuasion in digital platforms (Crothers et al., 2023; Ma and Luo, 2024; Shin et al., 2025). In contrast, we consider the potential negative consequences of AI usage on misinformation creation and offer insights into the optimal design of AI-generated content detection algorithms for content platforms.

3 Model Setup

We study a content platform that intermediates interactions between unit masses of creators and consumers. Creators produce content for user consumption, but not all content is equally valuable:

consumers care about both quality (e.g., narrative coherence, emotional resonance) and authenticity. While quality is revealed upon consumption, authenticity, specifically, whether content contains misinformation, is not directly observable. This asymmetry creates uncertainty for consumers and poses a design problem for the platform: how to supply information to consumers through detection algorithms that label content as AI- or human-generated.

Creator Types and Content Production

In our model, content is characterized along two dimensions: quality and authenticity. The quality dimension captures attributes such as newsworthiness as well as clarity, coherence, and polish, which directly influence consumer enjoyment. The authenticity dimension, orthogonal to quality, reflects whether the content is truthful or misinformation. This two-dimensional structure underscores that content varies both in its quality and truthfulness, and both dimensions jointly determine its value to consumers.

Building on this foundation, we model creator heterogeneity along the authenticity dimension. Each creator is of type $j \in \{T, D\}$: a fraction $\lambda \in (0, 1)$ are "truthful" creators ($j = T$) who exclusively produce truthful and authentic content, while the remaining $1 - \lambda$ are "deceptive" creators ($j = D$) who generate only fake content and misinformation. This dichotomy captures a structural distinction in the creator population: some consistently aim to inform and engage audiences with credible material, while others rely on sensational or misleading content to attract attention or extract value. A creator's type is private information, but the overall distribution, $\lambda = \Pr(j = T)$, is common knowledge among all agents, including consumers and the platform.

Each creator chooses an effort level $e_j \in [0, 1]$ and decides whether to adopt AI tools $a_j \in \{0, 1\}$ that reduce the cost of content generation.[3] Higher effort increases the probability of producing high-quality content, defined as content that appeals to consumers along observable dimensions such as clarity, coherence, or emotional resonance. A type-j creator who exerts effort e_j produces quality q with the following probabilities:

[3]In practice, AI adoption takes two main forms: assistive use, where creators leverage tools for supportive tasks such as editing drafts and enhancing clarity; and generative use, where AI tools produce substantial content such as drafting full articles or synthesizing images. Platforms, such as Meta, typically prioritize labeling generative outputs due to their higher detectability and greater misinformation risks, while often allowing assistive usage without labeling. For analytical tractability, our model focuses on generative AI adoption and represents it as a binary choice. Distinguishing between assistive and generative modes of adoption, and their implications for creator incentives and label informativeness, is an important direction for future work.

$$q = \begin{cases} \overline{q} & \text{with probability } \ \eta_j \cdot e_j, \\ \underline{q} = 0 & \text{otherwise,} \end{cases}$$

where $\eta_T = 1$ for truthful creators and $\eta_D = r > 1$ for deceptive creators.[4] The parameter $r > 1$ reflects the fundamental asymmetry that deceptive creators do not face the constraint of factual accuracy. Because they are free to invent narratives, they can make content appear engaging or "high quality" through tactics such as sensationalized framing and emotional triggers, without incurring the effort required to ensure accuracy. By contrast, truthful creators must invest effort not only in presentation but also in maintaining fidelity to facts. This gives deceptive creators an inherent advantage: for the same effort level, they are more likely to generate content that appears high quality, even if it is false. This asymmetry is consistent with empirical evidence showing that misinformation spreads faster and more broadly than truthful content due to novelty and high-arousal emotions such as fear, disgust, and surprise (Vosoughi et al., 2018).

The cost of effort is quadratic and depends on whether the creator adopts AI:

$$C(e; a) = c(a) \cdot \frac{e^2}{2}, \quad \text{where } \ c(a) = \begin{cases} c & \text{if } \ a = 0 \\ c/\theta & \text{if } \ a = 1 \end{cases},$$

where c denotes the baseline marginal cost of effort, and $\theta > 1$ captures the relative efficiency gain from AI, where $c > r^2 \cdot \theta$.[5] For example, $\theta = 2$ implies that AI reduces the marginal cost of effort by half. This reflects the assumption that AI tools enhance efficiency by lowering the marginal cost of generating high-quality content. In addition, creators who adopt AI incur a fixed cost $K > 0$. This structure captures the productivity benefit of AI: while using AI requires upfront investment, it enables more efficient content production by reducing the cost of exerting effort.

Creators benefit from attracting consumer engagement, which translates into both economic incentives (e.g., ad revenues, subscription fees) and non-economic motivations such as social influence or recognition. We normalize the revenue per unit of consumer engagement to one, so a type-j

[4]If $\eta_T > \eta_D$, the equilibrium behaviors of creators and consumers are reversed: truthful creators become at least as likely as deceptive creators to adopt AI—in other words, AI is more closely associated with truthful content rather than misinformation. Consequently, consumers are more willing to engage with AI-labeled content than human-labeled content. The platform can still influence consumers and creators' behavior through the informativeness of labels by adjusting the threshold. We note this case for completeness, although it does not realistically describe online content markets, where misinformation has consistently been shown to be easier to present as engaging than factual content.

[5]Throughout the analysis, we impose this minor technical assumption that $c > r^2 \cdot \theta$ to ensure that all equilibrium effort choices remain interior.

creator's total revenue is equal to the total consumer demand for their content, denoted by D_j. Therefore, the profit for a type-j creator is given by,

$$\pi_j(e_j, a_j) = D_j - C(e_j; a_j) - a_j \cdot K.$$

Consumer Utility

Consumers randomly pick one piece of content and derive utility only if the content is both high-quality and truthful. They enjoy visually polished, well-crafted, and informative content, but also care deeply about its authenticity. Accordingly, we assume that consumers receive utility $u = \overline{q}$ from high-quality content ($\overline{q}$) only if it is produced by a truthful creator, and zero otherwise. This reflects the idea that while superficial appeal may attract initial attention, credibility ultimately matters. Misinformation, even when engaging in form, can cause tangible harm.[6] We therefore model consumers as forward-looking agents who value not just content quality, but also informational integrity.

While content quality q is revealed upon observing the content, the creator type (and thus the content authenticity) is not. Consumers infer authenticity from observable signals, specifically the platform-generated label indicating whether content is likely AI- or human-created, and update their belief accordingly. Consumers have an outside option with $v_o > 0$. They choose to engage with content only when their expected utility based on posterior belief exceeds v_o:

$$\mathrm{E}[u|q, L] = q \cdot \Pr(j = T|q, L) \geq v_o,$$

where L is the label assigned by the platform, which affects consumers' beliefs. This structure introduces an inference problem: consumption decisions depend on how credible each label is as a signal of truthfulness, which in turn depends on creators' equilibrium strategies and the platform's detection rule.

Algorithmic Detection

The platform uses a probabilistic detection algorithm that assigns each content a score $s \in [0, 1]$, indicating the predicted probability that it was AI-generated. The score distribution depends on the true source of content: AI-generated content yields scores drawn from a cumulative distribution $F_A(s)$ with density function $f_A(s)$, while human-created content yields scores drawn from $F_H(s)$

[6] For example, misinformation about COVID-19 treatments led to widespread confusion and adverse health outcomes (Bridgman et al., 2020). See also Allcott and Gentzkow (2017) on the influence of fake news during elections.

with density $f_H(s)$. We assume the monotone likelihood ratio property (MLRP): higher scores are more indicative of AI generation. Formally, for $s_2 > s_1$,

$$\frac{f_A(s_2)}{f_H(s_2)} > \frac{f_A(s_1)}{f_H(s_1)}.$$

This property ensures that higher values of s are more likely to originate from AI-generated content, capturing the statistical relationship between AI usage and the algorithm's signal.

Based on this score, the platform sets a threshold $x \in (0, 1)$ to classify content and assign a binary label ($L \in \{L_A, L_H\}$): content with $s > x$ is labeled as AI-generated ($L = L_A$), while content with $s \leq x$ is labeled as human-created ($L = L_H$).[7] Because the classification is imperfect, the platform faces a fundamental trade-off between false positives (mislabeling human content as AI-generated) and false negatives (failing to detect AI-generated content). Specifically, the probability of a false positive is $1 - F_H(x)$, and the probability of a false negative is $F_A(x)$. Raising the threshold x reduces the false positives but boosts the false negatives, and vice versa. Table 1 summarizes the detection outcomes by content origin and label assignment.

		Label	
		Human ($L = L_H$)	AI ($L = L_A$)
AI Usage	Human-generated ($a = 0$): F_H	True Negative: $F_H(x)$	**False Positive**: $1 - F_H(x)$
	AI-generated ($a = 1$): F_A	**False Negative**: $F_A(x)$	True Positive: $1 - F_A(x)$

Table 1: Detection Outcomes Based on Threshold x

Platform's Objective

As platforms often rely on creators to produce content and monetize consumer engagement, we consider the platform chooses the detection threshold x to maximize a weighted summation of consumer surplus and type-T creator profits,

$$\Pi(x) = w \cdot CS + (1 - w) \cdot \pi_T,$$

where $w \in [0, 1]$ captures the platform's relative emphasis on consumer surplus versus creator-side incentives.[8] Consumer surplus CS is defined as the expected utility from consumed content, net

[7] The main model focuses on the detection threshold x as the platform's key decision variable. In extension 6.1, we extend the model to examine how changes in the F_A and F_H affect platform's decision.

[8] We will later show that the platform's optimal detection strategy is invariant to the choice of w, and that including type-D creators' profits in the platform's payoff does not affect the result.

of the outside option v_0: $CS \equiv E[u|q, L] - v_0$. A consumer engages with content if their expected utility exceeds an outside option v_0. Thus, surplus reflects the value derived from consumption, accounting for both content quality and credibility as inferred through platform labels.

Strategies, Timing, and Equilibrium Concept

We analyze a symmetric Perfect Bayesian Equilibrium in which all agents of the same type adopt identical strategies. A type-$j \in \{T, D\}$ creator's strategy is defined by a (possibly mixed) choice over AI adoption and effort level, represented by a probability distribution $\sigma_j(a, e) \in \Delta(\{0, 1\} \times [0, 1])$. On the demand side, consumers choose a strategy $\delta_c = (\delta_H, \delta_A) \in [0, 1]^2$, where δ_H and δ_A denote the probability of consuming high-quality content labeled as human-created (L_H) and as AI-generated (L_A), respectively. Consumers form beliefs about the content's type based on labels and maximize expected utility conditional on those beliefs.

The timing of the game unfolds as follows. In the first stage, the platform sets its detection algorithm by choosing a classification threshold $x \in (0, 1)$. Given this threshold, creators then decide whether to adopt AI tools and how much effort to exert in content creation in the second stage. Once content is produced, the platform applies its detection algorithm to label content as either AI-generated or human-created in the third stage. Finally, consumers randomly pick one piece of content from the platform, observe its realized quality and label, form beliefs about the content's source, and decide whether to engage. Payoffs are realized at this final stage. Figure 1 summarizes the game sequence.

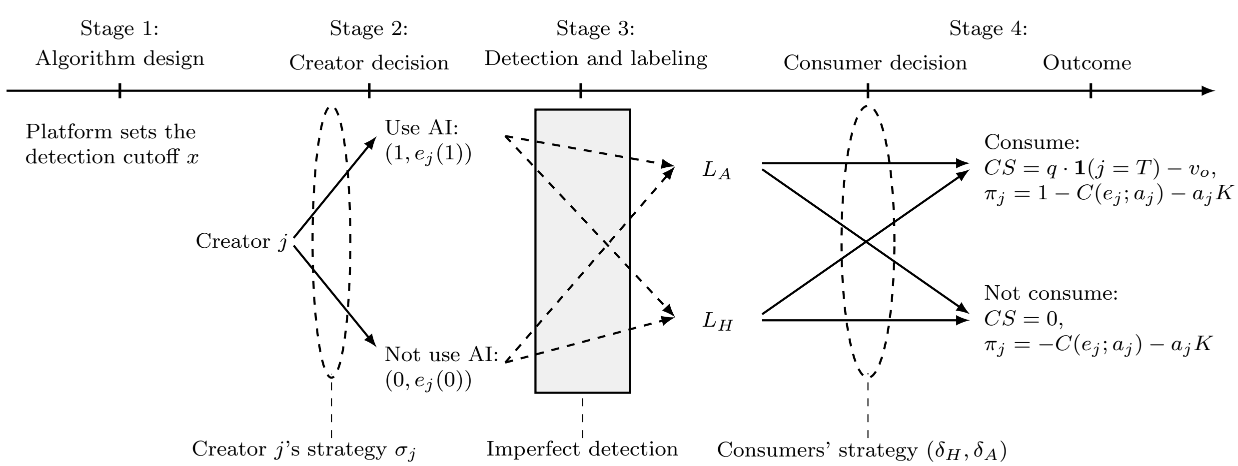


Figure 1: Timing of the Game

In the creators-consumers subgame (given any detection strategy x set by the platform), we focus on Perfect Bayesian Equilibrium, in which the platform's labels may serve as a noisy signal for the type of creator. Given the fraction of truthful creators λ, the creators' anticipated strategies $\widetilde{\sigma}_j$ for $j \in \{T, D\}$, and the observed label $L \in \{L_A, L_H\}$, consumers update their posterior beliefs using Bayes' rule whenever plausible. Formally, the resulting equilibrium must satisfy three conditions: (i) each individual type-j creator's strategy σ_j maximizes her profit given the consumer' strategy $\delta_c = (\delta_H, \delta_A)$ and the strategy of the other creators $\sigma_{j'}$ of both types $j \in \{T, D\}$; (ii) the consumer maximizes expected utility given beliefs and the equilibrium strategies of both creator types σ_j^* for $j \in \{T, D\}$; and (iii) beliefs are updated consistently with Bayes' rule wherever possible.

Multiple equilibria may exist in this subgame. To refine the subgame equilibria, we apply a Pareto-dominance criterion: an equilibrium is eliminated if there exists another subgame equilibrium that all three parties (the truthful creator, the deceptive creator, and the consumer) weakly prefer, with at least one party strictly better off.

4 Benchmark: The Case without Platform Detection

We begin our analysis by examining a benchmark where the platform does not employ any detection algorithm. This allows us to isolate the strategic interaction between creators and consumers, abstracting away the influence of platform-generated labels. In this environment, consumers observe only the realized quality of the content upon engagement, without knowing whether it was AI-generated or human-created. Consequently, upon encountering high-quality content $\overline{q}$, consumers must update their beliefs about its authenticity solely based on the anticipated equilibrium behavior of creators. Specifically, let $\widetilde{\sigma}_j(a, e)$ denote the anticipated strategy of a type-$j \in \{T, D\}$ creator, defined over AI adoption decision $a \in \{0, 1\}$ and effort level $e \in [0, 1]$. We introduce the following notations:

$$\bar{a}_j = \Pr(a_j = 1) = \int_0^1 \widetilde{\sigma}_j(1, e)\, de,$$
$$\bar{e}_j(a) = \mathbb{E}_{\widetilde{\sigma}_j}[e \mid a] = \frac{\int_0^1 e \cdot \widetilde{\sigma}_j(a, e)\, de}{\int_0^1 \widetilde{\sigma}_j(a, e)\, de},$$

where $\bar{a}_j$ is the probability that a type-j creator adopts AI (i.e., $a_j = 1$), and $\bar{e}_j(a)$ is the expected effort level conditional on AI adoption $a \in \{0, 1\}$. Given this, the probability that a type-j creator produces high-quality content is:

$$\Pr(q = \bar{q} \mid j) = \eta_j \cdot [\bar{a}_j \cdot \bar{e}_j(1) + (1 - \bar{a}_j) \cdot \bar{e}_j(0)],$$

where $\eta_T = 1$ and $\eta_D = r > 1$, capturing the idea that deceptive creators more easily produce superficially high-quality (but fake) content. Given this structure, a consumer's belief that a high-quality piece of content is from a truthful creator is:

$$\mu_0(\widetilde{\sigma}) = \frac{\lambda \cdot [\bar{a}_T \cdot \bar{e}_T(1) + (1 - \bar{a}_T) \cdot \bar{e}_T(0)]}{\lambda \cdot [\bar{a}_T \cdot \bar{e}_T(1) + (1 - \bar{a}_T) \cdot \bar{e}_T(0)] + (1 - \lambda) \cdot r \cdot [\bar{a}_D \cdot \bar{e}_D(1) + (1 - \bar{a}_D) \cdot \bar{e}_D(0)]}.$$

In equilibrium, $\widetilde{\sigma}$ will coincide with creators' equilibrium strategy σ, and thus a consumer will consume high-quality content if and only if $\mu_0(\widetilde{\sigma})\bar{q} - v_o \geq 0$. Given this decision rule (or consumer's strategy δ_c), creators choose effort to maximize their profit. The optimal effort levels under each AI adoption decision are:

$$e_T(0; \delta_c) = \arg\max_{e_T} \left\{ e_T \delta_c - \frac{c}{2} e_T^2 \right\} = \delta_c / c, \quad e_T(1; \delta_c) = \arg\max_{e_T} \left\{ e_T \delta_c - \frac{c}{2\theta} e_T^2 \right\} = \theta \delta_c / c.$$
$$e_D(0; \delta_c) = \arg\max_{e_D} \left\{ r e_D \delta_c - \frac{c}{2} e_D^2 \right\} = r \delta_c / c, \quad e_D(1; \delta_c) = \arg\max_{e_D} \left\{ r e_D \delta_c - \frac{c}{2\theta} e_D^2 \right\} = r \theta \delta_c / c.$$

Because creators choose from a discrete set of actions, we slightly abuse notation by letting $\sigma_j = (\bar{a}_j, e_j(1), e_j(0))$ denote a type-j creator's strategy, where $\bar{a}_j$ is the probability of adopting AI, and $e_j(1)$ and $e_j(0)$ denote effort levels with and without AI, respectively.

We next characterize different types of equilibria that may arise in this benchmark environment.

Pooling Equilibrium

When the share of truthful creators λ is sufficiently high, consumers are optimistic and always consume high-quality content ($\delta_c = 1$). Under these conditions, both types of creators may find it optimal to adopt the same AI adoption strategy, resulting in pooling equilibria.

Two such equilibria arise depending on the cost of adopting AI (K). If K is relatively small, there exists a pooling equilibrium where all creators adopt AI tools, $\bar{a}_T = \bar{a}_D = 1$. We refer to this as the "$Pool - 1$" equilibrium. Conversely, if K is high, a pooling equilibrium arises in which neither type adopts AI, i.e., $\bar{a}_T = \bar{a}_D = 0$, which we denote "$Pool - 0$".

A third pooling equilibrium, "$Pool - \emptyset''$, also exists in which creators exert no effort and consumers choose not to consume $\delta_c = 0$, leading to market collapses. However, this equilibrium is strictly Pareto-dominated whenever other equilibria exist.

Lemma 1. *There are three different types of pooling equilibria. Let* $\underline{K} \equiv \frac{\theta-1}{2c}$ *and* $\overline{K} \equiv \frac{r^2(\theta-1)}{2c}$.

(1) (Pool-1) If $\lambda \geq \underline{\lambda} \equiv \frac{r^2 v_o}{r^2 v_o + (\bar{q} - v_o)}$ *and* $0 < K \leq \underline{K}$*, a pooling equilibrium exists with* $\delta_c = 1$*,* $\bar{a}_T = \bar{a}_D = 1$*,* $e_T(1) = \theta/c$*, and* $e_D(1) = r\theta/c$*.*

(2) (Pool-0) If $\lambda \geq \underline{\lambda}$ *and* $K \geq \overline{K}$*, a pooling equilibrium exists with* $\delta_c = 1$*,* $\bar{a}_T = \bar{a}_D = 0$*,* $e_T(0) = 1/c$*, and* $e_D(0) = r/c$*.*

(3) (Pool-$\emptyset$*) For any* λ*, there always exists a pooling equilibrium where* $\delta_c = 0$ *and no effort or AI adoption.*[9] *This outcome is Pareto-dominated when other equilibria are feasible.*

Separating Equilibrium

When the cost of adopting AI is intermediate $\underline{K} \leq K \leq \overline{K}$ and $\lambda \geq \overline{\lambda} \equiv \frac{r^2 v_o \theta}{r^2 v_o \theta + (\bar{q} - v_o)}$, there can exist a pure-strategy separating equilibrium in which the two types of creators adopt different strategies. In the resulting separating equilibrium, truthful creators avoid AI ($\bar{a}_T = 0$), while deceptive creators adopt AI to exploit its cost advantages ($\bar{a}_D = 1$). Consumers, observing high-quality content, infer that it is likely to come from truthful types and therefore choose to consume ($\delta_c = 1$). The following lemma formalizes the existence condition of such a separating equilibrium. It highlights that the two types of creators' AI adoption decisions differ only when the cost of AI is neither too low nor too high.

Lemma 2. *If* $\underline{K} \leq K \leq \overline{K}$ *and* $\lambda \geq \overline{\lambda} \equiv \frac{r^2 v_o \theta}{r^2 v_o \theta + (\bar{q} - v_o)}$*, a separating equilibrium exists with* $\delta_c = 1$*,* $\bar{a}_T = 0$*,* $e_T(0) = 1/c$*, and* $\bar{a}_D = 1$*,* $e_D(1) = r\theta/c$*.*

Semi-separating Equilibrium

There also exists a semi-separating equilibrium where the type-T creators choose not to adopt AI ($\bar{a}_T = 0$), the type-D creators mix between using AI and not using AI ($0 < \bar{a}_D < 1$). Consumers, upon observing high-quality content, are indifferent between consuming and not consuming it.[10] This semi-separating equilibrium exists under intermediate conditions: the fraction of truthful creators is in the intermediate range $\underline{\lambda} < \lambda < \overline{\lambda}$ and the cost of AI adoption must also be moderate $0 < K < \overline{K}$. When λ is too high, consumers are overly optimistic and fully consume, leading to pooling or separating outcomes. When λ is too low, they become overly skeptical and disengage

[9] One can assume that upon observing a high-quality piece of content, $\widetilde{\mu} = \Pr(j = T \mid \bar{q}) = 0$ for supporting this Pool-$\emptyset$ equilibrium.

[10] There exists another possible semi-separating equilibrium in which the type-T creators are indifferent between using AI $\bar{a}_T \in (0, 1)$, and the type-D creators always use AI $\bar{a}_D = 1$. This equilibrium can only exist when $0 < K < \underline{K}$, and is Pareto-dominated by the Pool-1 equilibrium.

entirely. Similarly, if K is too low, all creators adopt AI, and if K is too high, neither creators have an incentive to adopt AI. When K is moderate $0 < K < \overline{K}$ and λ is intermediate $\underline{\lambda} < \lambda < \overline{\lambda}$, if consumers fully consume, then the type-D creators would adopt AI to exploit its cost advantage and thus generate more fake content, which deters consumers from fully consuming. Consequently, there does not exist a pure-strategy separating equilibrium in this parameter range. Moreover, this semi-separating equilibrium is Pareto-dominated by the Pool-1 equilibrium if $0 < K \leq \underline{K}$ (where $\underline{K} < \overline{K}$), as both types of creators adopt AI and consumers always consume high-quality content, leading to strictly higher surplus for all parties.

Lemma 3. *If $\underline{\lambda} < \lambda < \overline{\lambda}$ and $0 < K < \overline{K}$, there exists a semi-separating equilibrium where consumers are indifferent, and their strategy is given by $\delta_c = \frac{1}{r} \cdot \sqrt{\frac{2cK}{\theta - 1}}$. Type-$T$ creators choose $(0, e_T(0))$, and type-D creators mix between $(1, e_D(1))$ and $(0, e_D(0))$ with probability $\bar{a}_D \in (0,1)$. Moreover, if K is small such that $0 < K \leq \underline{K}$, this semi-separating equilibrium is Pareto-dominated by the Pool-1 equilibrium.*

This semi-separating equilibrium should be viewed as describing the aggregate mixing rate in equilibrium. Mixed-strategy equilibrium in game theory admits two standard interpretations (Fudenberg and Tirole, 1991). The first is individual-level randomization: each consumer engages with high-quality content with probability δ_c, and Type-D creators use AI with probability $\overline{a}_D$. The second is a population interpretation: a fraction δ_c of consumers always engage with high-quality content while the rest never do; a fraction $\overline{a}_D$ of Type-D creators choose $(1, e_D(1))$ and the rest follow $(0, e_D(0))$. Both interpretations yield the same aggregate outcome, and the latter is more natural in the current model of a large market. The semi-separating equilibria in the main model can follow the same interpretation.

Equilibrium Characterization without Platform Detection

Lemmas 1, 2, and 3 jointly characterize, in the full range of parameters, all possible equilibrium outcomes in the benchmark setting without platform detection. These equilibrium outcomes vary systematically with two key primitives: the share of truthful creators λ and the cost of AI adoption K. Figure 2 summarizes which equilibrium arises for each parameter region, focusing on those that are Pareto-optimal whenever multiple equilibria coexist. When the fraction of truthful creators is low ($\lambda < \underline{\lambda}$), consumers are sufficiently pessimistic that they refuse to engage with any content, leading to a collapse in demand. We therefore focus on the more relevant region $\lambda \geq \underline{\lambda}$, where at least some consumers are willing to engage with content.

For low AI cost ($0 < K \leq \underline{K}$), both types of creators adopt AI to reduce effort costs, and consumers are optimistic. This leads to a pooling equilibrium with full consumption and is indicated as Pool-1 in Figure 2. As the AI cost increases to an intermediate range ($\underline{K} < K < \overline{K}$), type-$T$ creators avoid using AI, while type-D creators still have an incentive to use it for its cost advantages. This divergence in their behaviors makes high-quality content a noisy signal of source type, introducing uncertainty into consumer inference, since high-quality content can originate from either type. The resulting equilibrium depends on how confident consumers are in the proportion of truthful creators. (1) If λ falls in the intermediate range ($\underline{\lambda} < \lambda < \overline{\lambda}$), consumers are uncertain about the truthfulness of high-quality content and thus only consume it with a probability $0 < \delta_c < 1$, leading to a semi-separating equilibrium. (2) If $\overline{\lambda} \leq \lambda < 1$, the prevalence of truthful creators makes consumers still willing to consume high-quality content, leading to a pure strategy separating equilibrium, where different types of creators follow distinct AI adoption decisions. Finally, when the AI cost is high ($K \geq \overline{K}$), neither type of creator adopts AI, and their behavior converges again. Consumers return to fully consuming content, resulting in a second pooling equilibrium: Pool-0 in Figure 2.

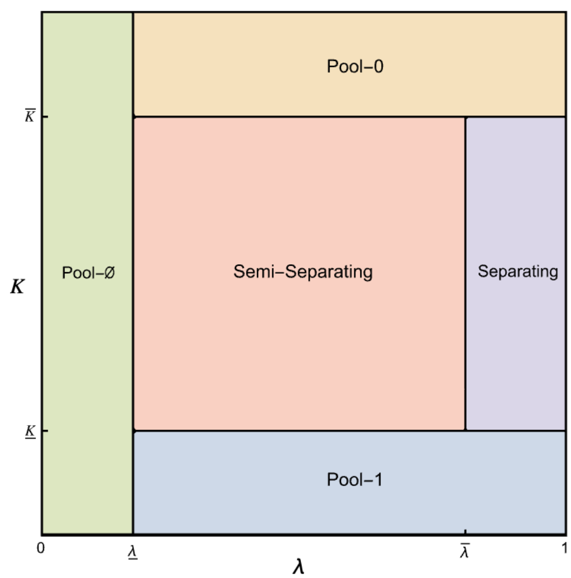


Figure 2: Equilibrium of Creators-Consumers Subgame without Platform Detection

Throughout the following main analysis, we focus on the intermediate range of λ by making the following assumption, where consumers remain skeptical about the truthfulness of high-quality content and the benchmark outcome is a semi-separating equilibrium in the absence of detection. This modeling choice allows us to concentrate on the region where algorithmic labeling is most consequential for shifting market equilibrium behavior.[11]

Assumption 1. $\underline{\lambda} < \lambda < \widetilde{\lambda}$.

5 Main Model: Platform Detection

We now extend the benchmark analysis to incorporate the platform's detection algorithm, which classifies content as either human-created (L_H) or AI-generated (L_A). The platform chooses a detection threshold $x \in (0, 1)$, which determines how aggressively the algorithm flags content as AI-generated. This threshold affects not only the distribution of labels observed by consumers but also the inference they draw from those labels, thereby affecting the creators' incentives.

The platform's detection algorithm assigns each piece of content a score indicating the likelihood that it was AI-generated. These scores follow different distributions depending on the content's true origin: $F_A(s)$ for AI-generated content and $F_H(s)$ for human-created content. The platform classifies content as AI-generated (L_A) if its score exceeds the threshold x, and as human-created (L_H) otherwise. Two key functions characterize the algorithm's classification accuracy. The cumulative distribution function $F_H(x)$ denotes the probability that human-created content receives a score below the threshold x, and is therefore (correctly) labeled as L_H. Conversely, $F_A(x)$ is the probability that AI-generated content also falls below x and is thus (incorrectly) labeled as L_H.

These functions jointly determine the accuracy and credibility of content labeling. A lower threshold reduces false negatives by decreasing $F_A(x)$, meaning fewer AI-generated contents are mislabeled as human. However, it simultaneously increases false positives by decreasing $F_H(x)$, meaning more human-generated contents are mislabeled as AI. Conversely, a higher threshold raises false negatives as $F_A(x)$ increases, and decreases false positives $(1 - F_H(x))$ as $F_H(x)$ increases. The platform thus faces a fundamental trade-off between minimizing false positives and minimizing false negatives. These relationships are summarized in Table 2.

This classification trade-off has direct implications for consumer beliefs and the informativeness of content labels. A lower threshold enhances detection aggressiveness but increases false positives, potentially eroding informativeness in the AI label (L_A) as more human-created content is misclassified. A higher threshold reduces false positives but allows more AI-generated content to evade detection, weakening the reliability of the human label (L_H). In either extreme, label credibility

[11] In the online Appendix, we analyze cases where λ is either very low ($\lambda < \underline{\lambda}$) or very high ($\lambda > \widetilde{\lambda}$). In either case, the platform's labeling algorithm does not affect consumers' consumption decisions (except the case when λ is not extremely high, where a separating equilibrium can arise. Even in this separating equilibrium, creators' AI adoption strategies and consumers' consumption strategy are fixed regardless of the platform's detection threshold).

Detection Threshold	$F_H(x)$	$F_A(x)$	False Positives $(1 - F_H(x))$	False Negatives $(F_A(x))$
Lower x (More Aggressive)	Decreases (↓)	Decreases (↓)	Increases (↑)	Decreases (↓)
Higher x (More Conservative)	Increases (↑)	Increases (↑)	Decreases (↓)	Increases (↑)

Table 2: Effect of Detection Threshold on Classification Outcomes

deteriorates, diminishing the informativeness of labels and impairing consumers' ability to assess content authenticity. These shifts in belief, in turn, influence consumers' engagement decisions and alter creators' strategic choices about AI adoption and effort. In the remainder of this section, we analyze how the platform's choice of detection threshold shapes equilibrium outcomes by interacting with consumer inference and creator incentives.

5.1 Detection Threshold, Belief Updating, and Incentives

We now examine how the platform's detection threshold x shapes the informativeness of content labels and how that, in turn, affects consumer inference and strategic creator behavior. Although creators move first in the game, their decisions are shaped by expectations of how consumers will interpret content labels. These labels, in turn, are probabilistic signals generated by the detection algorithm based on the platform's chosen threshold x.

The labeling mechanism assigns either L_H (human-created) or L_A (AI-generated) to any high-quality content based on the content's detection score. Recall that $F_H(x)$ is the probability that human-generated content is labeled as L_H, while $F_A(x)$ is the probability that AI-generated content is mistakenly labeled L_H. Upon seeing a label $L \in \{L_H, L_A\}$ attached to high-quality content, the consumer updates her posterior beliefs about whether it was created by a truthful creator based on these classification probabilities and anticipated creator behavior $\widetilde{\sigma}$. Let $m_j(H)$ and $m_j(A)$ denote the probability that a type-$j \in \{T, D\}$ creator generates high-quality content labeled as L_H and L_A, respectively. These probabilities depend on the creator's AI adoption probability $\overline{a}_j$, expected effort level $\bar{e}_j(a)$, and the accuracy under the detection algorithm, $F_H(x)$ and $F_A(x)$.

For example, $m_T(H)$ accounts for two possibilities: a type-T creator (i) uses AI, generates high-content content with probability $\bar{a}_T\bar{e}_T(1)$, and the algorithm misclassifies it as human-created with probability $F_A(x)$; or (ii) the creator does not use AI, generates high-content content with probability $(1 - \bar{a}_T)\bar{e}_T(0)$, and the algorithm correctly labels it as human-created with probability $F_H(x)$. The sum of these two cases gives $m_T(H)$. The other entries follow analogously. Table 3 summarizes all four labeling probabilities.

	$L = L_H$	$L = L_A$
$j = T$	$\bar{a}_T \bar{e}_T(1) F_A(x) + (1 - \bar{a}_T)\bar{e}_T(0) F_H(x)$	$\bar{a}_T \bar{e}_T(1)(1 - F_A(x)) + (1 - \bar{a}_T)\bar{e}_T(0)(1 - F_H(x))$
$j = D$	$r\left[\bar{a}_D \bar{e}_D(1) F_A(x) + (1 - \bar{a}_D)\bar{e}_D(0) F_H(x)\right]$	$r\left[\bar{a}_D \bar{e}_D(1)(1 - F_A(x)) + (1 - \bar{a}_D)\bar{e}_D(0)(1 - F_H(x))\right]$

Table 3: Probabilities of Content Labeling $m_j(L)$

Using these expressions, we can calculate consumers' posterior beliefs: $\mu_H(x)$ as the probability that human-labeled high-quality content is truthful, and $\mu_A(x)$ as the probability that AI-labeled high-quality content is truthful.[12]

$$\mu_H(x) = \frac{\lambda m_T(H)}{\lambda m_T(H) + (1-\lambda) m_D(H)}, \quad \mu_A(x) = \frac{\lambda m_T(A)}{\lambda m_T(A) + (1-\lambda) m_D(A)}. \tag{1}$$

Proposition 1 (Consumer Inference and Label Informativeness)**.** *Labels are strictly informative; that is, $\mu_H(x) > \mu_A(x)$ if and only if*

$$\frac{(1-\bar{a}_T)\bar{e}_T(0)}{\bar{a}_T \bar{e}_T(1) + (1-\bar{a}_T)\bar{e}_T(0)} > \frac{(1-\bar{a}_D)\bar{e}_D(0)}{\bar{a}_D \bar{e}_D(1) + (1-\bar{a}_D)\bar{e}_D(0)}. \tag{2}$$

Under this condition, both $\mu_H(x)$ and $\mu_A(x)$ strictly decrease in x: $\frac{\partial \mu_H(x)}{\partial x} < 0, \ \frac{\partial \mu_A(x)}{\partial x} < 0.$

This proposition characterizes when labels are informative and how informativeness responds to the detection threshold. If truthful creators are relatively more likely to avoid AI than deceptive creators, then human-labeled content (L_H) becomes informative, indicating a truthful origin.[13] Moreover, as the detection threshold x increases, the informativeness of the label L_H erodes but that of L_A enhances. More AI-generated content slips through as L_H (higher $F_A(x)$), and less human content is mislabeled as L_A (lower $1 - F_H(x)$). As a result, both beliefs $\mu_H(x)$ and $\mu_A(x)$ decline.

We now show that the condition for informativeness in Equation (2) always holds in equilibrium. Specifically, deceptive creators are more inclined to adopt AI than truthful creators, creating a systematic asymmetry in AI usage incentives.

Lemma 4 (AI Adoption Incentives)**.** *In any equilibrium, the type-D creators are at least as likely as the type-T creators to adopt AI: $\bar{a}_D \geq \bar{a}_T$. If $0 < \bar{a}_T < 1$, then $\bar{a}_D = 1$; if $0 < \bar{a}_D < 1$, then $\bar{a}_T = 0$.*

[12]For notational simplicity, we write $\mu_H(x)$ and $\mu_A(x)$ instead of the more precise $\mu_H(x \mid \widetilde{\sigma})$ and $\mu_A(x \mid \widetilde{\sigma})$, suppressing dependence on anticipated creator strategies when context permits.

[13]If both types either always adopt AI ($\bar{a}_T = \bar{a}_D = 1$) or never adopt ($\bar{a}_T = \bar{a}_D = 0$), labels provide no information about type and posterior beliefs revert to the prior ($\mu_H(x) = \mu_A(x) = \mu_0$), regardless of the detection threshold.

This asymmetry arises because deceptive creators benefit more from AI's efficiency gains and are less concerned with reputational loss of being labeled as AI. While both types face the same consumer inference and share the same payoff structure (expected credibility μ times content quality q), deceptive creators have a mechanical advantage: they are more likely to produce high-quality content (i.e., $r > 1$). Even if skeptical consumers assign lower credibility to AI-labeled content, deceptive creators can still expect a higher payoff through their greater ability to produce high-quality content. Truthful creators, by contrast, must exert more effort to generate high-quality content and thus rely more on maintaining credibility to attract consumption. This divergence leads to systematically stronger AI adoption incentives for deceptive creators, ensuring that Proposition 1 holds in equilibrium: content labels remain informative whenever detection is used.

5.2 Equilibrium Characterization

We now characterize the set of Perfect Bayesian Equilibria in the main model with platform detection. When the platform detection provides informative signals about content authenticity, creators' incentives to adopt AI diverge, generating the possibility of semi-separating equilibria. These outcomes are of primary interest because they capture meaningful interaction between platform policy and strategic behavior. In contrast, pooling equilibria, where both types of creators adopt the same AI strategy, arise only if the cost of AI is either very high $K \geq \overline{K}$ or very low $K \leq \underline{K}$, making content labels uninformative (as shown in Proposition 1).

Our focus is therefore on the interior case where AI costs are moderate ($\underline{K} < K < \overline{K}$) and detection meaningfully affects both consumer belief and creator strategy. In these semi-separating equilibria, type-T creators do not adopt AI to preserve credibility, while type-D creators mix between adopting and not adopting AI, balancing efficiency gains against reputational loss. Consumers, in turn, update their beliefs based on the observed label and choose whether to consume or not.[14]

Creator Payoffs

In any candidate semi-separating equilibrium, the type-T creators avoid AI entirely ($\overline{a}_T = 0$), and exert optimal effort $e_T(0)$ to maximize expected payoff. Type-D creators mix between adopting and not adopting AI with probability $\overline{a}_D \in (0,\ 1)$. Let (δ_H, δ_A) denote the consumer's mixed content consumption strategy upon seeing L_H and L_A, respectively. The type-j creators' expected

[14]There exist other possible semi-separating equilibria where the type-T creators mix between using AI $\bar{a}_T \in (0, 1)$, the type-D creators use AI $\bar{a}_D = 1$, and consumers mix on either high-quality content with L_A or that with L_H. The same as the benchmark, these equilibria exist for $0 < K < \underline{K}$, and are Pareto-dominated by the Pool-1 equilibrium.

payoffs:

$$\pi_j(a=1) = \max_{e_j} \left\{ \eta_j \cdot e_j \cdot [F_A(x)\delta_H + (1-F_A(x))\delta_A] - \frac{c \cdot e_j^2}{2\theta} \right\} - K,$$
$$\pi_j(a=0) = \max_{e_j} \left\{ \eta_j \cdot e_j \cdot [F_H(x)\delta_H + (1-F_H(x))\delta_A] - \frac{c \cdot e_j^2}{2} \right\},$$

where $\eta_T = 1$ and $\eta_D = r > 1$ as before. In equilibrium, type-D creators must be indifferent between these two actions: $\pi_D(a=1) = \pi_D(a=0)$.

Consumer Mixing Behavior

In any semi-separating equilibrium, consumers mix over one label while either always consuming or never consuming the other. The next result characterizes this structure.

Lemma 5 (Consumer Mixing)**.** *In any semi-separating equilibrium, if consumers mix on L_A ($0 < \delta_A < 1$), then they always consume L_H content ($\delta_H = 1$). Conversely, if $0 < \delta_H < 1$, then $\delta_A = 0$.*

This lemma implies that semi-separating equilibria fall into two distinct regimes: (i) In a *semi-A* equilibrium, consumers are indifferent only over AI-labeled L_A content: $\delta_A \in (0,1)$ and $\delta_H = 1$. (ii) In a *semi-H* equilibrium, they are indifferent only over human-labeled L_H content: $\delta_H \in (0,1)$ and $\delta_A = 0$. The intuition follows directly from the nature of binary state (truthful or fake) and belief monotonicity, which constrains consumer strategies to two indifference-based outcomes: if consumers mix upon observing L_A, they must be indifferent between consuming AI-labeled content and not, which implies that $\mu_A \cdot \overline{q} = v_o$ holds with equality. Since $\mu_H > \mu_A$ by Proposition 1, content labeled L_H must be strictly more attractive, so $\delta_H = 1$. Similarly, if consumers mix on L_H, they must always reject L_A content.

Figure 3 illustrates the underlying semi-separating equilibria structure. Consumers do not observe the creator's type or AI adoption directly. Instead, the platform's detection algorithm probabilistically labels content based on the chosen threshold x, potentially generating misclassification. Consumers then observe the label and decide to consume with probability δ_H or δ_A, depending on the label observed. The figure captures the multi-stage strategic interaction among creators, the platform detection algorithm, and consumer inference that supports semi-separating equilibria.

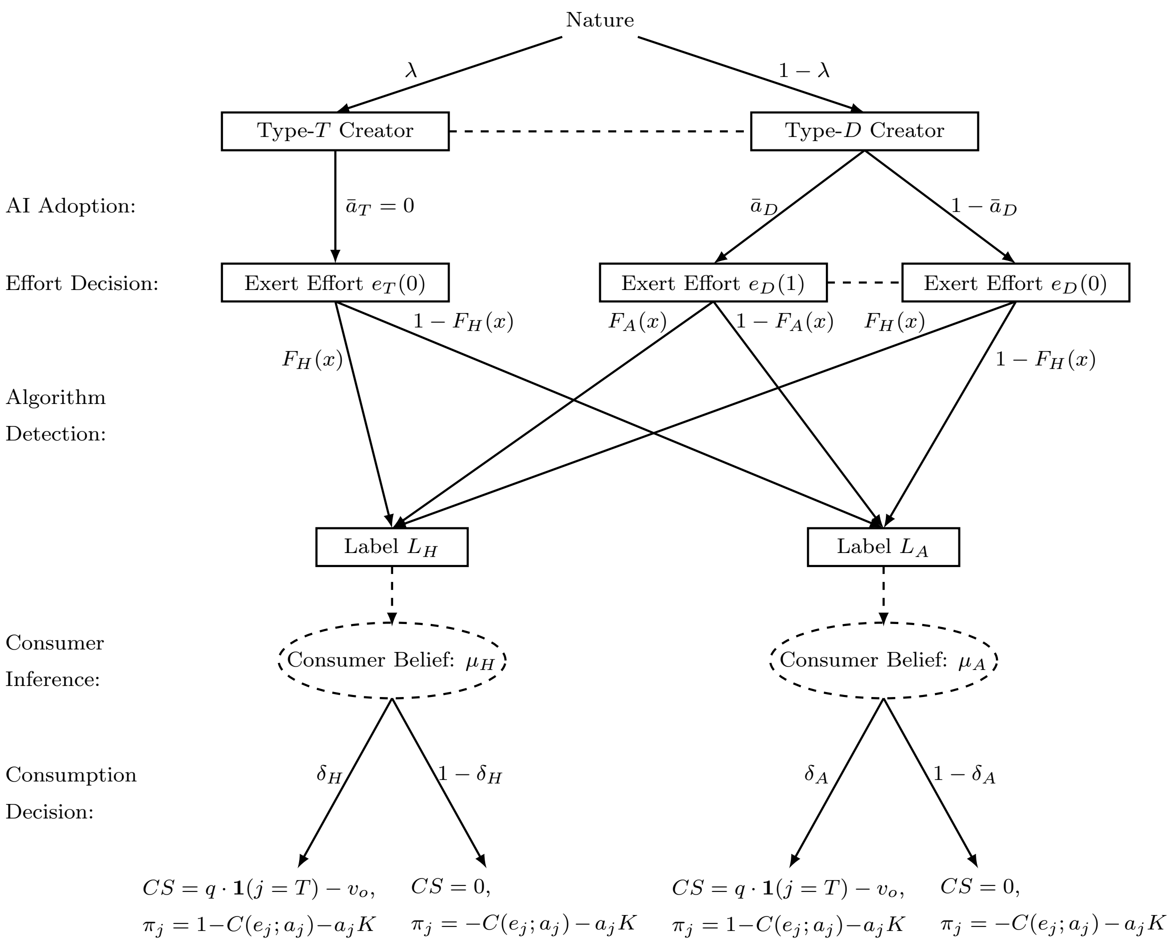


Figure 3: Game Tree of Semi-separating Equilibria

Effort Choices and Creator Strategies

Given (δ_H, δ_A), the type-T creator's optimal effort is:

$$e_T(a = 0) = \frac{F_H(x)\delta_H + (1 - F_H(x))\delta_A}{c}.$$

Type-D creators' optimal effort levels, conditional on their AI adoption decision, must be:

$$e_D(1) = \frac{r\theta\left[F_A(x)\delta_H + (1 - F_A(x))\delta_A\right]}{c} \quad \text{and} \quad e_D(0) = \frac{r\left[F_H(x)\delta_H + (1 - F_H(x))\delta_A\right]}{c}.$$

In equilibrium, the type-D creator's AI adoption probability $\bar{a}_D$ must make consumers indifferent upon seeing AI-labeled content or human-labeled content. Then, consumers' indifference condition pins down the type-D creator's equilibrium mixing probability $\bar{a}_D$. We characterize the existence conditions for both *semi-A* and *semi-H* equilibria in Proposition 2.

Proposition 2 (Equilibrium Characterization)**.** *Under Assumption 1, the equilibrium depends on the cost of AI adoption K and the platform's detection threshold x as follows:*

(1) If $0 < K \leq \underline{K}$, the unique equilibrium is the Pool-1 equilibrium: both types adopt AI.

(2) If $\underline{K} < K < \overline{K}$, two different types of semi-separating equilibria exist:

- *(i)* ***(Semi-A):*** *If $x \in (0, x^*]$, consumers mix over L_A content ($\delta_A \in (0,1)$) and fully consume L_H ($\delta_H = 1$). Type-T creators avoid AI, and type-D creators mix AI adoption with probability $\bar{a}_D^{semi_A} \in (0,1)$.*
- *(ii)* ***(Semi-H):*** *If $x \in (x^*, 1)$, consumers never consume L_A ($\delta_A = 0$) and mix over L_H ($\delta_H \in (0,1)$). Type-T creators avoid AI, and type-D creators mix AI adoption with probability $\bar{a}_D^{semi_H} \in (0,1)$.*

(3) If $K \geq \overline{K}$, the unique equilibrium is the Pool-0 equilibrium: both types avoid AI.

Proposition 2 and Figure 4 highlight how equilibrium outcomes depend on both AI adoption costs and the platform's detection threshold. When K is either very low ($K \leq \underline{K}$) or prohibitively high ($K \geq \overline{K}$), both types of creators make the same AI adoption decision (either both adopt or both abstain), making the label uninformative and resulting in a pooling equilibrium. However, in the intermediate regime, detection plays a strategic role, and the equilibrium depends critically on the detection threshold x.

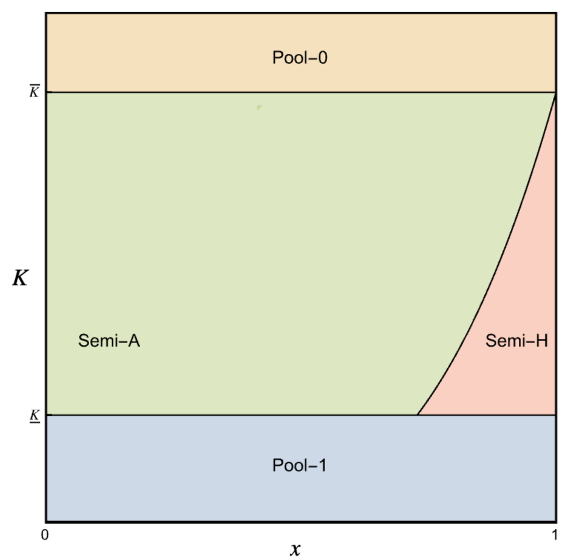


Figure 4: Equilibrium Characterization with Detection

Under an *aggressive* detection $0 < x \leq x^*$, AI-generated content is less likely to be mislabeled as human-created (i.e., false negative rate is low), making the human label highly informative. As a result, consumers trust human-labeled content but remain skeptical of content with an AI label, leading to a *semi-A* equilibrium. In contrast, under a *conservative* detection policy $x^* < x < 1$, human-created content is rarely misclassified (i.e., false positive rate is low), and the AI label becomes a strong indicator of AI-generated content. Therefore, consumers never consume AI-labeled content and mix over human-labeled content, resulting in a *semi-H* equilibrium.

5.3 Impact of Detection Threshold on Equilibrium Outcomes and Welfare

We now analyze how the platform's detection threshold x affects equilibrium outcomes. As shown in Proposition 2, semi-separating equilibria arise when AI adoption is asymmetric across creator types and labels remain informative. In this section, we examine how changes in x shape equilibrium behavior for both consumers and creators. These effects, taken together, determine the platform's optimal detection policy and total welfare.

Effect of Detection on Consumers' Equilibrium Strategy

Detection affects how consumers interpret content labels. As established in Proposition 1, both posterior beliefs $\mu_H(x)$ and $\mu_A(x)$ decline as x increases: the human label becomes less informative, while the AI label becomes a more definitive indication of misinformation. This shift does not imply uniform erosion of informativeness, but rather a redistribution: L_H loses credibility as AI-generated content slips through, while L_A becomes a stronger signal of misinformation.

Lemma 6 (Detection and Consumer Equilibrium Strategy)**.** *Consumer equilibrium strategy* $(\delta_H(x), \delta_A(x))$ *varies with the detection threshold* x *as follows:*

(1) In semi-A region ($x \leq x^*$)*,* $\delta_H(x) = 1$*, and* $\delta_A(x)$ *can be non-monotonic: when* x *is small, it increases with* x *if and only if* $f_A(0)/f_H(0)$ *is small; when* x *is large, it declines with* x*.*

(2) In semi-H region ($x > x^*$)*,* $\delta_A(x) = 0$*, and* $\delta_H(x)$ *strictly decreases in* x*.*

Lemma 6 captures how detection policy reshapes consumer behaviors by altering posterior beliefs and the informativeness of labels. In *semi-A* equilibria (aggressive detection), consumers fully trust human-labeled content ($\delta_H = 1$) and consume the AI-labeled content with probability $\delta_A(x) \in (0, 1)$. Figure 5 shows the non-monotonicity of δ_A in the *semi-A* regime. As x initially increases from zero, the algorithm becomes relatively accurate and the AI label becomes more informative, which can increase the truthful creator's effort and thus, consumer engagement (raising δ_A); as x further increases, the AI label becomes increasingly associated with deception, diluting consumer trust and decreasing consumer engagement (lowering δ_A). Once detection becomes sufficiently conservative ($x > x^*$), the equilibrium shifts to *semi-H* regime, where the AI label is fully avoided ($\delta_A = 0$), and consumer demand becomes concentrated on L_H. Even then, trust in L_H falls as the detection threshold rises, since a higher x increases the likelihood that AI content is misclassified as human. Thus, $\delta_H(x)$ declines in this regime as well.

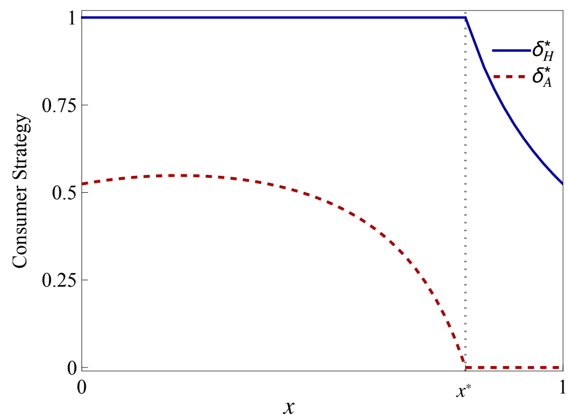


Figure 5: Effect of Detection Threshold x on Consumer Strategy

Overall, the transition from *semi-A* to *semi-H* reflects a structural shift in consumer inference: from partial trust in both labels to selective engagement with human-labeled content only. The regime shift at x^* marks a collapse in AI-labeled content demand and a reduction in overall engagement. As detection becomes overly conservative, both $\mu_H(x)$ and $\mu_A(x)$ fall, not because both labels lose value equally, but because the growing prevalence of misclassification undermines confidence in the labeling system altogether.

Effect of Detection on Creators' AI Adoption and Effort

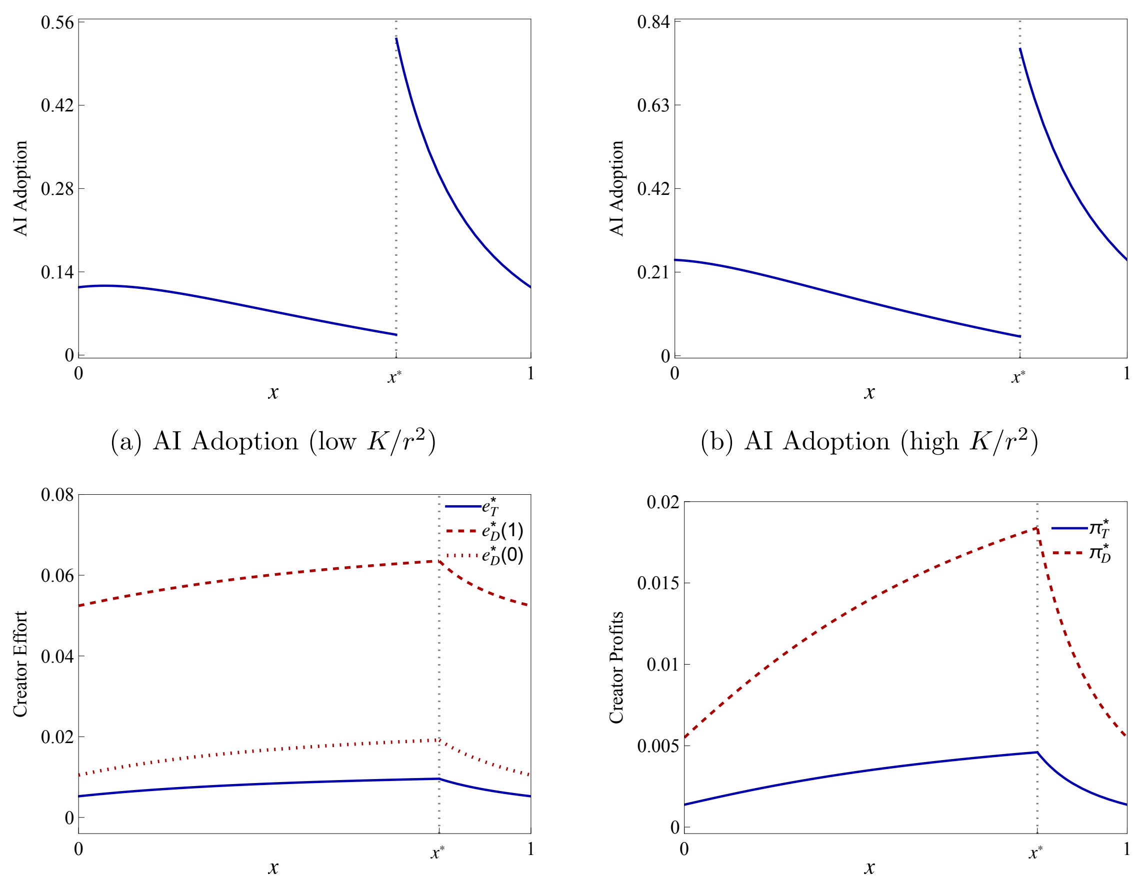
The detection threshold also shapes creators' AI adoption decisions and their effort choice. Under an aggressive detection (i.e., low x), the probability that AI-generated content is labeled as such is high, reducing the relative appeal of AI adoption due to lower demand for L_A content. Note that $\bar{a}_D(x)$ denotes the equilibrium probability that type-D creators adopt AI.

Lemma 7 (Detection and AI Adoption)**.** *In equilibrium,*

(1) In semi-A region ($x \leq x^*$), $\bar{a}_D(x)$ may be non-monotonic if K/r^2 is small; otherwise (i.e., K/r^2 is high), it strictly decreases in x.

(2) In semi-H region ($x > x^*$), $\bar{a}_D(x)$ strictly decreases in x, with a discrete jump at x^*.

These dynamics are depicted in the upper panels (a) and (b) of Figure 6. In the *semi-A* regime, when AI is relatively cheap (low K/r^2), initial increases in x reduce the penalty of AI usage and permit opportunistic AI adoption due to the low cost. However, further increases in x erode AI label credibility, reduce consumer engagement with AI-labeled content, and thus may cause $\bar{a}_D$ to decline. On the other hand, when AI is more costly (high K/r^2), this opportunistic AI adoption does not arise due to the cost, and as x increases, it consistently discourages AI usage, and $\bar{a}_D$ decreases monotonically. Once detection crosses the boundary into the *semi-H* regime ($x > x^*$), AI adoption jumps sharply in both cost cases, which can be explained by the equilibrium regime shift. At $x = x^*$, consumers cease engaging AI-labeled content and start to consume human-labeled content only probabilistically. This shift erodes the reward for avoiding AI (as δ_H starts to decline), prompting deceptive creators to increase their AI adoption rate. In *semi-H* regime, as credibility deteriorates and consumers avoid AI-labeled content, further raising x reduces creator incentives to adopt AI and results in a continued decline in $\bar{a}_D(x)$.



(a) AI Adoption (low K/r^2)

(b) AI Adoption (high K/r^2)

(c) Creator Effort

(d) Creator Profit

Figure 6: Effect of Platform Detection on Creator Behavior

Moreover, creators adjust their effort in response to equilibrium demand, which depends on both label distributions and consumer beliefs.

Lemma 8 (Creators' Effort and Profits)**.** *In equilibrium:*

(1) In semi-A region ($x \leq x^*$), both types increase their effort levels with x.

(2) In semi-H region ($x > x^*$), all effort levels decline as x increases.

Moreover, creator profits move in the same direction as effort levels.

Panels (c) and (d) of Figure 6 illustrate these relationships. In the *semi-A* regime, higher x allocates more content into the human-labeled category, increasing expected demand. This "label reallocation" effect dominates any loss in label credibility, incentivizing both types to exert greater effort. In contrast, in the *semi-H* regime, a more conservative detection erodes the credibility of human labels. As consumers grow more skeptical, expected demand falls, prompting creators to scale back efforts. Panel (d) confirms that this decline in effort leads directly to lower profits.

Put all these lemmas 7, 8 together, the detection threshold has a fundamental strategic impact on creator behavior: an intermediately conservative policy can moderate type-D creators' AI usage (by threatening them with detection while still rewarding truthful creators' content quality via an informative labeling), but an overly conservative policy backfires by encouraging rampant AI-generated fake content by shifting to *semi-H* equilibrium. This behavioral pattern directly affects the welfare outcomes on the platform.

5.4 Welfare Implications and Optimal Detection Threshold

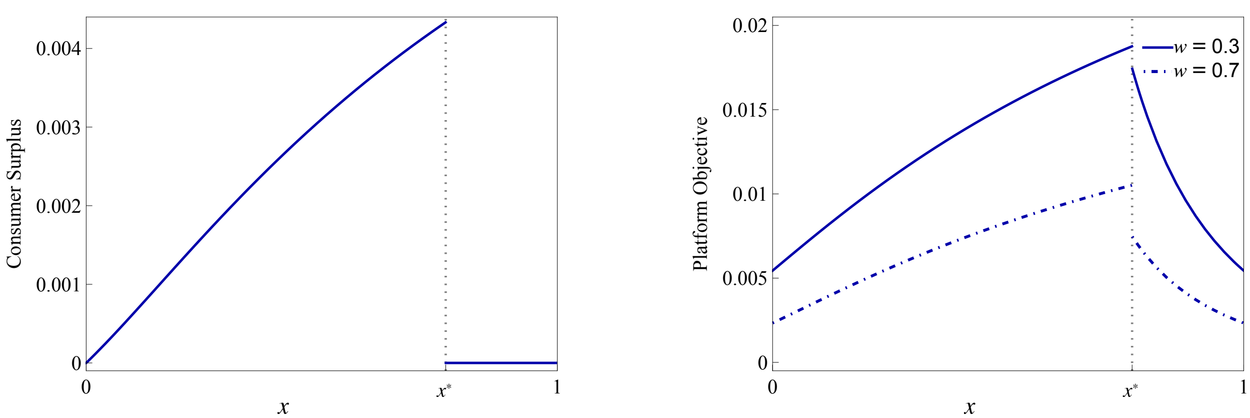
Following Lemma 8 and Proposition 3, it is clear that consumer welfare and creator profit (particularly the profits of truthful creators) are both non-monotonic in the detection threshold x. Let $CS(x)$ denote consumer surplus. When x is very low (under overly aggressive detection), fewer AI-generated posts are mislabeled as human-created (reducing false negatives). However, excessively aggressive detection also increases false positives, leading to the misclassification of truthful human content, causing unwarranted skepticism and lower consumption. At the other extreme, an overly conservative detection (a very high x) deteriorates content credibility, reducing the overall engagement and lowering creators' profit. These behavioral adjustments feed directly into market welfare outcomes. The next result summarizes our main findings on how welfare outcomes vary with x, and investigates the optimal detection policy.

Proposition 3 (Welfare and Optimal Detection Threshold)**.** *Consumer surplus CS increases in x under semi-A region ($x \leq x^*$), discretely drops at $x = x^*$, and remains flat under semi-H region ($x > x^*$). Moreover, the platform's objective, maximizing a weighted summation of consumer surplus and type-T creators' profit, is uniquely maximized at x^*. At this threshold:*

(1) The resulting equilibrium is the semi-A equilibrium, in which consumers fully consume L_H labeled content but do not consume L_A labeled content: $\delta_c^{semi_A} = (1, 0)$.

(2) Both consumer surplus and the profit of truthful creators attain their maximum values.

Figure 7 illustrates these results. Panel (a) plots the pattern of consumer surplus. In the *semi-A* regime ($x \leq x^*$), consumers fully engage with L_H content and selectively consume L_A content. As x increases, fewer human-generated posts are misclassified as AI (i.e., false positives decline), increasing the share of trustworthy L_H content and thereby improving consumption. Thus, consumers at first benefit from more posts under the reliable L_H label, and truthful creators benefit from greater engagement. Consumer surplus rises with x in this region, and so do truthful creators' profits. However, as x approaches x^*, false negatives become more prominent, causing more AI-generated content to be mislabeled as L_H. This undermines the credibility of the human label. Once x crosses the threshold x^* (shift to *semi-H* regime), the informativeness of AI labels improves, and consumers cease engaging with L_A content and become wary of even L_H content. This results in a sharp drop in consumer surplus at $x = x^*$, with no further gain beyond this point ($x > x^*$), since AI-labeled content is already ignored entirely, additional conservativeness only weakens engagement with L_H.



(a) Consumer Surplus (b) Platform Objective

Figure 7: Effect of Detection Threshold on Welfare

Panel (b) shows the platform's optimal detection threshold, which maximizes the weighted summation of consumer surplus and type-T creators' profits for different weights w. Across all weights, the objective is single-peaked and uniquely maximized at x^*. This threshold balances label credibility and informativeness: it deters AI misuse while preserving the credibility of L_H content.

On the creator side, profits follow a similar pattern. In the *semi-A* regime ($x \leq x^*$), type-T creators benefit from credible human labels, and type-D creators face disincentives to adopt AI, reducing misinformation. When $x > x^*$, the equilibrium shifts to *semi-H* regime and the credibility of human label becomes diluted; even human-labeled posts face skepticism. This harms both consumer engagement and type-T creators' payoffs. Thus, total surplus is maximized at an interior threshold x^*, the unique point where engagement, credibility, and welfare are jointly optimized. A more aggressive detection wastes legitimate content; a more conservative detection invites misuse and distrust.

6 Extensions

In this extension, we explore how the platform's optimal detection strategy responds to changes in the underlying environment. Specifically, we examine two key drivers: (i) improvements in detection technology, and (ii) changes in the cost of AI adoption.

6.1 Detection Technology Development

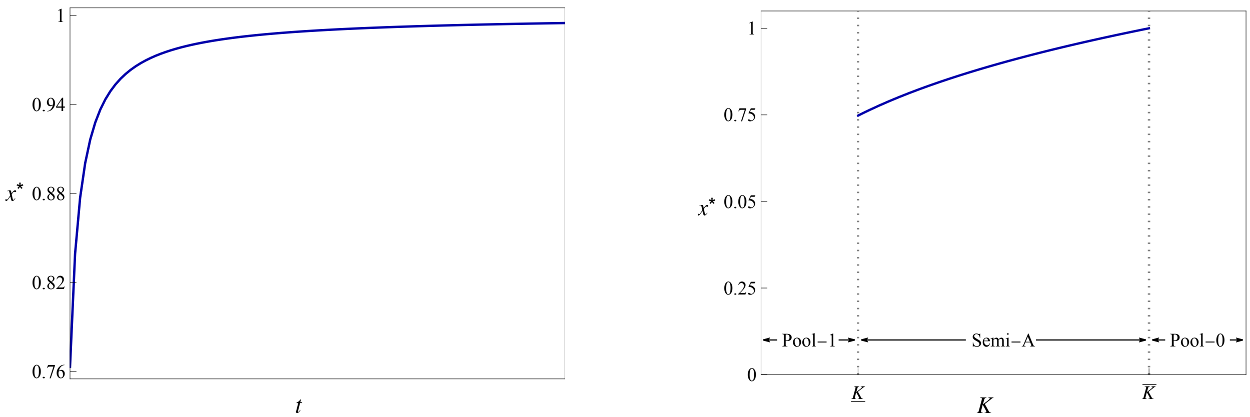
We first investigate how improvements in detection technology affect the platform's optimal detection strategy. We introduce a factor t to parameterize the performance of the detection algorithm, with higher t indicating more accurate classification. Formally, we assume:

$$\frac{\partial F_A(x;t)}{\partial t} < 0 \quad \text{and} \quad \frac{\partial F_H(x;t)}{\partial t} > 0, \tag{3}$$

so that as t increases, the algorithm becomes more accurate: both false negatives (F_A) and false positives ($1 - F_H$) decrease.

Proposition 4 (Effect of Algorithm Technology on Optimal Detection Strategy)**.** *Given $\underline{K} < K < \overline{K}$, the platform's optimal detection strategy x^* increases with t.*

Proposition 4 and panel (a) of Figure 8 show that as the detection algorithm becomes more accurate, the platform can afford to relax its detection policy. This is because a sophisticated algorithm reduces the risk of false positives and false negatives. By increasing x^*, the platform can still maintain consumer trust and deter type-D creators from adopting AI without incurring a large increase in misinformation. Technological improvement thus enables greater flexibility to be more conservative without triggering excessive AI adoption and misinformation.



(a) Detection Technology

(b) AI Adoption Cost

Figure 8: Comparative Statics of Optimal Detection Threshold

6.2 Effect of AI Adoption Cost

We next examine how the platform's optimal detection strategy responds to changes in the cost of AI adoption in Proposition 5 and panel (b) of Figure 8. We have the following proposition.

Proposition 5 (Effect of AI Adoption Cost on Optimal Detection Strategy)**.**

(i) If $K \leq \underline{K}$ or $K \geq \overline{K}$, the platform does not employ detection.

(ii) If $\underline{K} < K < \overline{K}$, the platform's optimal detection threshold x^* increases with K.

When adoption costs are either very low ($K \leq \underline{K}$) or very high ($K \geq \overline{K}$), both types of creators make the same AI adoption decision, so detection does not alter the equilibrium and is therefore unnecessary. By contrast, when K falls within an intermediate range, the platform actively employs detection. In this region, as AI tools become cheaper, creators, particularly of type-D, face stronger incentives to use them. The platform must then reoptimize its detection policy in response to these shifts in incentives from declining AI costs. A more aggressive policy becomes necessary to preserve content credibility and deter widespread AI usage that could undermine engagement and trust.

7 Conclusion

Generative AI has revolutionized content creation, but it also raises serious concerns about the spread of misinformation and its potential harm to consumer welfare. This paper analyzes how content platforms design detection algorithms to manage the risk of AI-generated misinformation. Our model demonstrates how detection policy shapes consumer inference, creator incentives, and overall welfare by incorporating strategic interactions among creators, consumers, and the platform.

We first identify a core trade-off: aggressive detection increases the informativeness of the human-label, making it a more credible signal of truthful content, but also increases the risk of misclassifying legitimate human content, thereby discouraging truthful creators. In contrast, conservative detection reduces false positives but weakens consumer inference, allowing deceptive creators to adopt AI more intensively. The platform's optimal detection strategy must balance these forces, not by simply minimizing classification errors, but by choosing the threshold that best aligns incentives, discouraging excessive AI adoption, and sustaining consumer engagement. We then show that improvements in detection accuracy allow platforms to adopt more conservative policies without sacrificing welfare, while cheaper AI technologies require a more aggressive detection policy to maintain credibility.

These results highlight a broader insight that the platform's detection policy is not merely a classification tool. It fundamentally alters the strategic behavior of both sides of the market. By influencing the informativeness of content labels and consumer inference, detection rules endogenously affect AI adoption, effort, and consumption decisions. Designing optimal detection strategies, therefore, requires accounting for these equilibrium outcomes and recognizing that algorithmic accuracy alone is not sufficient. What matters is how detection reshapes incentives in two-sided content markets, where platform policies influence not only the informativeness of content labels but also the strategic responses of creators and consumers alike.

As generative AI continues to evolve, understanding the economic foundations of detection strategies will become increasingly critical for platform design and policy. Our analysis offers guidance for practitioners designing content moderation systems: detection should be viewed not in isolation, but as part of an incentive architecture that governs user behavior and trust. By explicitly modeling how detection shapes market-level outcomes, our work contributes to the broader literature on platform governance and algorithmic interventions. We hope these insights will encourage future research at the intersection of economics, technology, and platforms, and inform the development of responsible detection strategies in an increasingly AI-driven environment.

Appendix

Proof of Lemma 1: Pooling Equilibrium in Benchmark

Proof. We show the existence condition of each pooling equilibrium.

(i) Pool-1 Equilibrium: Given $\delta_c = 1$, both types prefer adopting AI if the gain from lower effort cost exceeds the AI cost:

$$\begin{cases} \max\limits_{e_T} \left\{ e_T \cdot \delta_c - \frac{c}{2\theta} \cdot e_T^2 \right\} - K \geq \max\limits_{e_T} \left\{ e_T \cdot \delta_c - \frac{c}{2} \cdot e_T^2 \right\} \\ \max\limits_{e_D} \left\{ r \cdot e_D \cdot \delta_c - \frac{c}{2\theta} \cdot e_D^2 \right\} - K \geq \max\limits_{e_D} \left\{ r \cdot e_D \cdot \delta_c - \frac{c}{2} \cdot e_D^2 \right\} \end{cases} \Leftrightarrow 0 < K \leq \frac{\theta - 1}{2c} \equiv \underline{K}.$$

Creators' optimal efforts under AI are $e_T(1) = \theta/c$ and $e_D(1) = r\theta/c$. Consumers consume if their posterior belief exceeds the threshold: $\mu_0(\widetilde{\sigma}) = \frac{\lambda e_T(1)}{\lambda e_T(1) + (1-\lambda) r e_D(1)} \geq \frac{v_o}{\overline{q}} \Leftrightarrow \lambda \geq \underline{\lambda} \equiv \frac{r^2 v_o}{r^2 v_o + (\overline{q} - v_o)}$.

(ii) Pool-0 Equilibrium: Both types prefer not adopting AI if $K \geq \overline{K} \equiv \frac{r^2(\theta-1)}{2c}$ by similar logic in case (i) above. Their optimal efforts without AI are $e_T(0) = 1/c$ and $e_D(0) = r/c$. Consumers will again consume if $\lambda \geq \underline{\lambda}$.

(iii) Pool-$\emptyset$ Equilibrium: If no effort is exerted, the content is uniformly low-quality. Therefore, consumers' posterior belief upon seeing high-quality content is off-equilibrium-path and thus can be specified as 0 and choose not to consume. Given this, creators have no incentive to exert effort, confirming the equilibrium. □

Proof of Lemma 2: Separating Equilibrium in Benchmark

Proof. Given $\delta_c = 1$, a separating equilibrium where type-T creators avoid AI and type-D creators adopt AI exists if and only if

$$\begin{cases} \max\limits_{e_T} \left\{ e_T \cdot \delta_c - \frac{c}{2\theta} \cdot e_T^2 \right\} - K \leq \max\limits_{e_T} \left\{ e_T \cdot \delta_c - \frac{c}{2} \cdot e_T^2 \right\} \\ \max\limits_{e_D} \left\{ r \cdot e_D \cdot \delta_c - \frac{c}{2\theta} \cdot e_D^2 \right\} - K \geq \max\limits_{e_D} \left\{ r \cdot e_D \cdot \delta_c - \frac{c}{2} \cdot e_D^2 \right\} \end{cases} \Leftrightarrow \underline{K} \leq K \leq \overline{K}.$$

Then, the optimal effort levels are $e_T(0) = \frac{1}{c}$ and $e_D(1) = \frac{r\theta}{c}$. Given this, consumers will consume high-quality content if and only if the λ is sufficiently high, $\mu(\widetilde{\sigma}) = \frac{\lambda e_T(0)}{\lambda e_T(0) + (1-\lambda) r e_D(1)} \geq \frac{v_o}{\overline{q}} \Leftrightarrow \lambda \geq \overline{\lambda} \equiv \frac{r^2 v_o \theta}{r^2 v_o \theta + (\overline{q} - v_o)}$. □

Proof of Lemma 3: Semi-separating Equilibrium in Benchmark

Proof. In equilibrium, type-D creators are indifferent between adopting AI or not, which requires

$$\max_{e_D}\left\{r\cdot e_D\cdot\delta_c-\frac{c}{2\theta}\cdot e_D^2\right\}-K=\max_{e_D}\left\{r\cdot e_D\cdot\delta_c-\frac{c}{2}\cdot e_D^2\right\}\ \Rightarrow\ \delta_c=\frac{1}{r}\sqrt{\frac{2cK}{\theta-1}}.$$

This implies $0<\delta_c<1$ if and only if $0<K<\overline{K}$.

Given δ_c, the optimal efforts are $e_T(0)=\frac{\delta_c}{c}$, $e_D(0)=\frac{r\delta_c}{c}$, and $e_D(1)=\frac{r\theta\delta_c}{c}$. In equilibrium, the type-D creator's strategy $\bar{a}_D$ makes consumers indifferent between consuming high-quality content and taking the outside option, such that $\frac{\lambda e_T(0)}{\lambda e_T(0)+(1-\lambda)r[\bar{a}_D e_D(1)+(1-\bar{a}_D)e_D(0)]}=\frac{v_o}{\bar{q}}$, which implies that $\bar{a}_D=\frac{(\bar{q}-v_o)\lambda-(1-\lambda)r^2v_o}{(1-\lambda)r^2v_o(\theta-1)}$. To ensure $\bar{a}_D\in(0,1)$, we have $\underline{\lambda}<\lambda<\overline{\lambda}$.

Finally, note that the semi-separating equilibrium is Pareto-dominated by the Pool-1 equilibrium for $0<K\leq\underline{K}$. Under semi-separating equilibrium: $\pi_T=\frac{\delta_c^2}{2c}$, $\pi_D=\frac{(r\delta_c)^2\theta}{2c}-K$, $CS=0$. Under the Pool-1: $\pi_T=\frac{\theta}{2c}-K$, $\pi_D=\frac{r^2\theta}{2c}-K$, $CS=\frac{\theta}{c}\left[\lambda(\bar{q}-v_0)-(1-\lambda)r^2v_o\right]>0$, which are higher than their counterparts in semi-separating equilibrium. □

Proof of Proposition 1: Consumer Inference and Label Informativeness

Proof. We compare the posterior beliefs $\mu_H(\widetilde{\sigma})$ and $\mu_A(\widetilde{\sigma})$. The difference is

$$\mu_H(\widetilde{\sigma})-\mu_A(\widetilde{\sigma})=\frac{\lambda(1-\lambda)r\left(F_H(x)-F_A(x)\right)\left[\bar{a}_D\bar{e}_D(1)\cdot(1-\bar{a}_T)\bar{e}_T(0)-(1-\bar{a}_D)\bar{e}_D(0)\cdot\bar{a}_T\bar{e}_T(1)\right]}{\left[\lambda m_T(H)+(1-\lambda)m_D(H)\right]\cdot\left[\lambda m_T(A)+(1-\lambda)m_D(A)\right]},$$

which implies that $\mu_H(\widetilde{\sigma})-\mu_A(\widetilde{\sigma})\geq 0\Leftrightarrow\bar{a}_D\bar{e}_D(1)\cdot(1-\bar{a}_T)\bar{e}_T(0)-(1-\bar{a}_D)\bar{e}_D(0)\cdot\bar{a}_T\bar{e}_T(1)\geq 0$. The expression is 0 when $\bar{a}_T=\bar{a}_D=1$, or $\bar{a}_T=\bar{a}_D=0$. Otherwise, $\mu_H(\widetilde{\sigma})>\mu_A(\widetilde{\sigma})$ holds if

$$\frac{(1-\bar{a}_T)\bar{e}_T(0)}{\bar{a}_T\bar{e}_T(1)+(1-\bar{a}_T)\bar{e}_T(0)}>\frac{(1-\bar{a}_D)\bar{e}_D(0)}{\bar{a}_D\bar{e}_D(1)+(1-\bar{a}_D)\bar{e}_D(0)}.$$

This condition is satisfied under the following: (i) $0<\bar{a}_T<1$ and $\bar{a}_D=1$; (ii) $0<\bar{a}_D<1$ and $\bar{a}_T=0$, and (iii) $\bar{a}_T=0$ and $\bar{a}_D=1$. To analyze how these beliefs change with x, we differentiate:

$$\frac{\partial\mu_H(\widetilde{\sigma})}{\partial x}\propto-\frac{d}{dx}\left(\frac{F_A(x)}{F_H(x)}\right),\quad\frac{\partial\mu_A(\widetilde{\sigma})}{\partial x}\propto-\frac{d}{dx}\left(\frac{1-F_A(x)}{1-F_H(x)}\right).$$

Using MLRP ($\frac{f_A(t)}{f_H(t)}<\frac{f_A(x)}{f_H(x)}$ for $t<x$), we have:

$$\frac{d}{dx}\left(\frac{F_A(x)}{F_H(x)}\right)=\frac{f_A(x)F_H(x)-F_A(x)f_H(x)}{F_H^2(x)}=\frac{1}{F_H^2(x)}\left[f_A(x)\int_0^x f_H(t)dt-f_H(x)\int_0^x f_A(t)dt\right]>0,$$

$$\frac{d}{dx}\left(\frac{1-F_A(x)}{1-F_H(x)}\right) = \frac{f_H(x)(1-F_A(x)) - f_A(x)(1-F_H(x))}{(1-F_H(x))^2}$$
$$= \frac{1}{(1-F_H(x))^2}\left[f_H(x)\int_x^1 f_A(t)dt - f_A(x)\int_x^1 f_H(t)dt\right] > 0, \tag{4}$$

Therefore, both μ_H and μ_A strictly decrease in x. □

Proof of Lemma 4: Creators' AI Adoption Incentive

Proof. Given consumers' strategy (δ_H, δ_A), the type-T creators adopt AI if

$$\max_{e_T}\left\{e_T\left[F_A(x)\delta_H + (1-F_A(x))\delta_A\right] - \frac{c}{2\theta}e_T^2\right\} - K \geq \max_{e_T}\left\{e_T\left[F_H(x)\delta_H + (1-F_H(x))\delta_A\right] - \frac{c}{2}e_T^2\right\}$$
$$\Leftrightarrow K \leq \frac{\theta}{2c}\left[F_A(x)\delta_H + (1-F_A(x))\delta_A\right]^2 - \frac{1}{2c}\left[F_H(x)\delta_H + (1-F_H(x))\delta_A\right]^2.$$

Similarly, type-D creators adopt AI if

$$K \leq r^2\Big\{\frac{\theta}{2c}\left[F_A(x)\delta_H + (1-F_A(x))\delta_A\right]^2 - \frac{1}{2c}\left[F_H(x)\delta_H + (1-F_H(x))\delta_A\right]^2\Big\}.$$

Since $r > 1$, the condition for type-T creators to adopt AI always implies the condition for type-D creators. Moreover, if type-T creators are indifferent (i.e., $0 < \overline{a}_T < 1$), then type-D creators strictly prefer to adopt AI, i.e., $\overline{a}_D = 1$. If type-D creators are indifferent ($0 < \overline{a}_D < 1$), then inequality for type-D binds, and inequality for type-T does not hold, implying $\overline{a}_T = 0$. □

Proof of Proposition 2: Equilibrium Characterization

Proof. In this proof, we summarize the main logic and relegate the full derivations to the Online Appendix for brevity. We first define $\phi(x) = \frac{r^2}{2c}\left(\theta F_A^2(x) - F_H^2(x)\right)$, which measures the net profit gain from using AI for type-D creators given $\delta_H = 1$ and $\delta_A = 0$.

(i) Semi-A region: Given consumers' strategy (δ_H, δ_A) with $\delta_H = 1$, the type-D creator is indifferent between adopting AI and not, and this indifference condition pins down $\delta_A^{semi_A} = \frac{F_H(1-F_H)r - F_A(1-F_A)\theta r + \sqrt{\left[(1-F_A)^2\theta - (1-F_H)^2\right]2cK - (F_H-F_A)^2 r^2\theta}}{(1-F_A)^2\theta - (1-F_H)^2}$, which requires $0 < K < \overline{K}$ and $0 < x \leq x^* \equiv \phi^{-1}(K)$, where x^* is the detection threshold at which the type-D creator is indifferent between using AI and not at the fixed cost K given $\delta_c = (1,0)$, thereby delimiting the semi-A and semi-H regimes. Given this, optimal efforts are $e_T^{semi_A} = \frac{F_H + (1-F_H)\delta_A^{semi_A}}{c}$, $e_D^{semi_A}(0) = \frac{r[F_H + (1-F_H)\delta_A^{semi_A}]}{c}$, and $e_D^{semi_A}(1) = \frac{r\theta[F_A + (1-F_A)\delta_A^{semi_A}]}{c}$. Then, type-$D$ creators' mixing probability $\overline{a}_D^{semi_A}$ ensures $\mu_A = \frac{v_o}{\overline{q}}$. Lastly, we show that semi-A equilibrium is Pareto-dominated by Pool-1 when $0 < K \leq \underline{K}$, as it yields lower creators' profits and consumer surplus due to $\delta_A^{semi_A} < 1$.

(ii) Semi-H region: Type-D creators' indifference implies that $\delta_H^{semi_H} = \frac{1}{r}\sqrt{\frac{2cK}{F_A^2(x)\theta - F_H^2(x)}}$, which requires $0 < K < \overline{K}$ and $x^* < x < 1$. Creators' optimal effort follows analogously. Type-D creators' mixing probability $\bar{a}_D^{semi_H}$ ensures $\mu_H = \frac{v_o}{\bar{q}}$. Last, we show that the semi-H equilibrium is Pareto-dominated by Pool-1 for $0 < K \leq \underline{K}$, as creators' and consumers' payoffs are strictly lower. □

Proof of Lemma 6: Detection and Consumer Equilibrium Strategy

Proof. ***(i) Semi-A region:*** In this case, $\delta_A^{semi_A}$ solves the indifference condition $\pi_D(1) = \pi_D(0)$. By the implicit function theorem, $\frac{d\delta_A^{semi_A}}{dx} = -\frac{(\partial\pi_D(1)/\partial x) - (\partial\pi_D(0)/\partial x)}{(\partial\pi_D(1)/\partial\delta_A) - (\partial\pi_D(0)/\partial\delta_A)} \propto \left(\frac{\partial\pi_D(0)}{\partial x} - \frac{\partial\pi_D(1)}{\partial x}\right) \propto \psi(x)$, where $\psi(x) \equiv \left[F_H(x) + (1 - F_H(x))\delta_A^{semi_A}\right] f_H(x) - \theta\left[F_A(x) + (1 - F_A(x))\delta_A^{semi_A}\right] f_A(x)$. We evaluate $\psi(x)$ at the boundaries. As $x \to 0$, note that $\delta_A^{semi_A} > 0$ and $f_H(x), f_A(x) > 0$, so $\psi(x) \to f_H(x)\delta_A^{semi_A}\left(1 - \theta\frac{f_A(x)}{f_H(x)}\right)$. Under MLRP, this limit is positive if $\frac{f_A(0)}{f_H(0)}$ is sufficiently small. As $x \to x^*$ (with $\phi(x^*) = K > 0$), we have $\delta_A^{semi_A} \to 0$ and $\frac{f_A(x^*)}{f_H(x^*)} > \frac{F_A(x^*)}{F_H(x^*)} > \frac{1}{\sqrt{\theta}}$, implying $\psi(x^*) = F_H(x^*)f_H(x^*) - \theta \cdot F_A(x^*)f_A(x^*) < 0$. Thus, $\delta_A^{semi_A}(x)$ can be non-monotonic in x.

(ii) Semi-H region: Differentiating $\delta_H^{semi_H}$ yields $\frac{d\delta_H^{semi_H}}{dx} \propto -\frac{d}{dx}\left(F_A^2(x)\theta - F_H^2(x)\right)$. It suffices to show $F_H^2(x)\left[\left(\frac{F_A(x)}{F_H(x)}\right)^2\theta - 1\right]$ increases with x. Since $F_H(x)$ and $\frac{F_A(x)}{F_H(x)}$ both increases in x from Equation (4), the expression is strictly increasing, implying $\delta_H^{semi_H}$ decreases in x. □

Proof of Lemma 7: Detection and AI Adoption

Proof. ***(i) Semi-A region:*** Let $z(x) = \frac{1-F_A(x)}{1-F_H(x)}$ denote the survival function ratio, and define $y(z)$ as a transformed effort ratio capturing equilibrium behavior: $y(z) = \frac{e_D(1)}{e_D(0)} \cdot \frac{z(x)}{\theta}$, where $y < z$ due to $e_D(1)/\theta < e_D(0)$ in semi-A equilibrium. The variable $y(z)$ combines the ratio of effort across AI and non-AI strategies, the label survival rate, and the cost advantage of AI. This allows us to express the type-D indifference condition $\pi_D(1) = \pi_D(0)$ as:

$$(z-1)^2\left(\frac{y^2}{z^2} - \frac{1}{\theta}\right) = \zeta \cdot \left(z - \frac{y}{z}\right)^2, \text{ where } \zeta \equiv \frac{2Kc}{\theta r^2}. \tag{5}$$

This equation defines $y(z)$ implicitly as a function of z, and hence of x. The AI adoption rate is:

$$\overline{a}_D^{semi_A}(x) = \frac{\lambda(\overline{q} - v_o) - (1-\lambda)rv_o^2}{(1-\lambda)rv_o^2(\theta y - 1)}.$$

Here, $\bar{a}_D^{semi_A}(x)$ is strictly decreasing in $y(z)$ ($\frac{d\,\bar{a}_D^{semi_A}(x)}{dy} < 0$), and $z(x) = \frac{1-F_A(x)}{1-F_H(x)}$ increases in x ($\frac{dz}{dx} > 0$). Thus, the sign of $\frac{d\bar{a}_D^{semi_A}(x)}{dx}$ is the opposite of $\frac{dy}{dz}$:

$$\frac{d\bar{a}_D^{semi_A}}{dx} = \frac{d\bar{a}_D^{semi_A}}{dy} \cdot \frac{dy}{dz} \cdot \frac{dz}{dx} \quad \propto -\frac{dy}{dz}.$$

Therefore, the sign of $\left(\frac{dy}{dz}\right)$ determines whether AI adoption increases or decreases in x. The following two Claims together establish the comparative statics result in the semi-A regime.

Claim 1 (Non-monotonic case)**.** *If $\frac{K}{r^2} < \frac{\theta-1}{8c}$, then $y(z)$ can be non-monotonic in z: it can either first decrease and then increase, or always decrease. Therefore, $\bar{a}_D^{semi_A}(x)$ can be non-monotonic.*

Proof. See the Online Appendix. □

Claim 2 (Monotonic case)**.** *If $\frac{K}{r^2} \geq \frac{\theta-1}{8c}$, then $y(z)$ is strictly increasing in z, implying that $\bar{a}_D^{semi_A}(x)$ decreases in x.*

Proof. See the Online Appendix. □

(ii) Semi-H region: For $x > x^*$, consumers ignore AI-label content entirely and $\bar{a}_D^{semi_H} = \frac{\lambda(\bar{q}-v_o)-(1-\lambda)r^2v_o}{r^2v_o(1-\lambda)}\left[\theta\left(\frac{F_A(x)}{F_H(x)}\right)^2 - 1\right]^{-1}$, which strictly decreases in x because $\frac{F_A(x)}{F_H(x)}$ increases with x. ***(iii) Discontinuity at x^*:*** Finally, as $x \to x^*$, the AI adoption rate jumps up: $\lim_{x\to x^{*-}} \bar{a}_D^{semi_A} < \lim_{x\to x^{*+}} \bar{a}_D^{semi_H}$ due to $\frac{F_A(x)}{F_H(x)} < \frac{1-F_A(x)}{1-F_H(x)}$. □

Proof of Lemma 8: Creators' Effort and Profits

Proof. ***(i) Semi-A region:*** In this region, we have $\pi_D(1) = \pi_D(0)$. We prove the monotonicity by contradiction. Suppose instead $\pi_D(1)$ or $\pi_D(0)$ has a non-monotonic relationship with x and admits a critical point $x' \in (0, x^*]$. Then, both derivatives must vanish at x', which means $\frac{d\pi_D(1)}{dx} = \frac{d\pi_D(0)}{dx} = 0$. This leads to $\frac{d\delta_A^{semi_A}}{dx} = -\frac{f_A(x)(1-\delta_A^{semi_A})}{1-F_A(x)} = -\frac{f_H(x)(1-\delta_A^{semi_A})}{1-F_H(x)}$, which implies $\frac{f_A(x)}{1-F_A(x)} = \frac{f_H(x)}{1-F_H(x)}$. However, it contradicts MLRP, which ensures that $\frac{f_A(x)}{f_H(x)}$ increases in x, implying $\frac{f_H(x)}{1-F_H(x)} > \frac{f_A(x)}{1-F_A(x)}$.

To prove monotone increasing behavior in x, consider the behavior near $x = x^*$. We show $\lim_{x\to x^{*-}} \frac{d\pi_D(1)}{dx} = \lim_{x\to x^{*-}} \frac{d\pi_D(0)}{dx} > 0$. By the chain rule, we have $\frac{d\pi_D(0)}{dx} \propto f_H(x)(1-\delta_A^{semi_A}) + (1 - F_H(x))\frac{d\delta_A^{semi_A}}{dx}$. As $x \to x^{*-}$, we have $\delta_A^{semi_A} \to 0$ and thus $f_H(x^*) + (1 - F_H(x^*))\frac{d\delta_A^{semi_A}}{dx}\Big|_{x=x^*} = \frac{\theta F_A(x^*)[f_H(x^*)(1-F_A(x^*)-f_A(x^*)(1-F_H(x^*))]}{\theta F_A(x^*)(1-F_A(x^*))-F_H(x^*)(1-F_H(x^*))} > 0$. Hence, π_D (and thus $\pi_T = \frac{1}{r^2}\pi_D$) strictly increases in x. Since equilibrium profits are quadratic in effort, the associated effort levels must also rise in x.

(i) Semi-H region: In this regime, creators' efforts are $e_D(1) = \frac{\theta}{rc}\sqrt{\frac{2cK(F_A(x)/F_H(x))^2}{(F_A(x)/F_H(x))^2\theta-1}}$, and $e_D(0) = \frac{1}{rc}\sqrt{\frac{2cK}{(F_A(x)/F_H(x))^2\theta-1}}$. which decrease with x because $F_A(x)/F_H(x)$ increases in x (from Equation (4)). Then, the type-T creators' effort and the two types of creators' profits also decrease with x by the same logic in case (i) above. □

Proof of Proposition 3: Welfare and Optimal Detection Threshold

Proof. For $0 < x \leq x^*$, a semi-A equilibrium exists. Consumer surplus is:

$$\begin{aligned} CS &= \lambda e_T^{semi_A}(q - v_o) - (1-\lambda)r\left[\bar{a}_D^{semi_A} e_D^{semi_A}(1) + (1-\bar{a}_D^{semi_A})e_D^{semi_A}(0)\right] v_o \\ &= \left\{\lambda(q-v_o) - (1-\lambda)r^2\left[\bar{a}_D^{semi_A}(\Gamma(x)-1)+1\right] v_o\right\} e_T^{semi_A}, \end{aligned}$$

where $\Gamma(x) \equiv \frac{e_D^{semi_A}(1)}{e_D^{semi_A}(0)} > 1$ decreases in x, and $e_T^{semi_A}$ increases in x (from Lemma 8). To show $CS(x)$ increases in x, it suffices to prove $\bar{a}_D^{semi_A}(\Gamma(x)-1)$ decreases in x. Then, we have $\bar{a}_D^{semi_A} = \frac{\lambda(q-v_o)-(1-\lambda)r^2 v_o}{(1-\lambda)r^2 v_o\left[\Gamma(x)\frac{1-F_A(x)}{1-F_H(x)}-1\right]}$, which implies that

$$\frac{d}{dx}\left(\bar{a}_D^{semi_A}\left(\Gamma(x)-1\right)\right) \propto \left[\frac{F_H(x)-F_A(x)}{1-F_H(x)}\frac{d\Gamma(x)}{dx} - \Gamma(x)\left(\Gamma(x)-1\right)\frac{d}{dx}\left(\frac{1-F_A(x)}{1-F_H(x)}\right)\right] < 0.$$

For $x^* < x < 1$, the semi-H equilibrium exists, and consumer surplus is constant at 0. Since type-T profit is also maximized at x^*, the platform's objective is maximized at x^*. □

Proof of Proposition 4: Effect of Algorithm Technology on Optimal Detection Strategy

Proof. By introducing t, the optimal x^* solves $\phi(x;t) - K = \frac{r^2}{2c}\left(\theta F_A^2(x;t) - F_H^2(x;t)\right) - K = 0$. By the implicit function theorem, $dx^*/dt = -(\partial\phi(x;t)/\partial t)/(\partial\phi(x;t)/\partial x)|_{x=x^*}$. Since $\frac{\partial\phi(x;t)}{\partial x} > 0$ whenever $\phi(x;t) > 0$, it suffices to show that

$$\frac{\partial\phi(x,t)}{\partial t} = \frac{r^2}{c}\left(\theta F_A(x;t)\frac{\partial F_A(x;t)}{\partial t} - F_H(x;t)\frac{\partial F_H(x;t)}{\partial t}\right) < 0.$$ □

Proof of Proposition 5: Effect of AI Adoption Cost on Optimal Detection Strategy

Proof. When $\underline{K} < K < \overline{K}$, the optimal detection strategy x^* is determined by $\phi(x) = K$, where $\phi(x) = \frac{r^2}{2c}\left(\theta F_A^2(x) - F_H^2(x)\right)$. By the implicit function theorem, we have $\frac{dx^*}{dK} = \frac{1}{\partial\phi(x)/\partial x}\bigg|_{x=x^*} > 0$, since $\partial\phi(x)/\partial x > 0$ for all x such that $\phi(x) > 0$. □

References

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.

Bairathi, M., Zhang, X., and Lambrecht, A. (2025). The value of platform endorsement. *Marketing Science*, 44(1):84–101.

Björkegren, D., Blumenstock, J. E., and Knight, S. (2020). Manipulation-proof machine learning. *arXiv preprint arXiv:2004.03865*.

Bridgman, A., Merkley, E., Loewen, P. J., Owen, T., Ruths, D., Teichmann, L., and Zhilin, O. (2020). The causes and consequences of covid-19 misperceptions: Understanding the role of news and social media. *Harvard Kennedy School Misinformation Review*, 1(3).

Brown, Z. Y. and MacKay, A. (2023). Competition in pricing algorithms. *American Economic Journal: Microeconomics*, 15(2):109–156.

Calvano, E., Calzolari, G., Denicolo, V., and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297.

Cao, H. H., Ma, L., Ning, Z. E., and Sun, B. (2024). How does competition affect exploration vs. exploitation? a tale of two recommendation algorithms. *Management Science*, 70(2):1029–1051.

Chang, D., Segura, A., and Zhang, P. (2024). Decentralizing content moderation. *Available at SSRN 4709599*.

Che, Y.-K. and Hörner, J. (2018). Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics*, 133(2):871–925.

Chen, B., Li, K., and Guan, X. (2025). Platform certification and consumer verification. *Working Paper*.

Choi, W. J., Liu, Q., and Shin, J. (2024). Predictive analytics and ship-then-shop subscription. *Management Science*, 70(2):1012–1028.

Crothers, E. N., Japkowicz, N., and Viktor, H. L. (2023). Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002.

Dai, T. and Singh, S. (2025). Overdiagnosis and undertesting for infectious diseases. *Marketing Science*, 44(2):353–373.

Eliaz, K. and Spiegler, R. (2019). The model selection curse. *American Economic Review: Insights*, 1(2):127–140.

Fudenberg, D. and Tirole, J. (1991). *Game theory*. MIT press.

Hansen, K. T., Misra, K., and Pai, M. M. (2021). Frontiers: Algorithmic collusion: Supra-competitive prices via independent algorithms. *Marketing Science*, 40(1):1–12.

Hui, X., Saeedi, M., Spagnolo, G., and Tadelis, S. (2023). Raising the bar: Certification thresholds and market outcomes. *American Economic Journal: Microeconomics*, 15(2):599–626.

Iyer, G. and Ke, T. T. (2024). Competitive model selection in algorithmic targeting. *Marketing Science*, 43(6):1226–1241.

Iyer, G., Yao, Y. J., and Zhong, Z. Z. (2024). Precision-recall tradeoff in competitive targeting. *Working Paper*.

Jain, S. and Qian, K. (2021). Compensating online content producers: A theoretical analysis. *Management Science*, 67(11):7075–7090.

Kapoor, A. and Kumar, M. (2025). Frontiers: Generative ai and personalized video advertisements. *Marketing Science*.

Klein, T. (2021). Autonomous algorithmic collusion: Q-learning under sequential pricing. *The RAND Journal of Economics*, 52(3):538–558.

Liang, A. (2019). Games of incomplete information played by statisticians. *arXiv preprint arXiv:1910.07018*.

Liu, Y., Yildirim, P., and Zhang, Z. J. (2022). Implications of revenue models and technology for content moderation strategies. *Marketing Science*, 41(4):831–847.

Ma, L. and Luo, L. (2024). Wisdom of the ai crowd? can we detect ai-generated product reviews? *Working paper*.

Madio, L. and Quinn, M. (2024). Content moderation and advertising in social media platforms. *Journal of Economics & Management Strategy*.

Miklós-Thal, J. and Tucker, C. (2019). Collusion by algorithm: Does better demand prediction facilitate coordination between sellers? *Management Science*, 65(4):1552–1561.

Montiel Olea, J. L., Ortoleva, P., Pai, M. M., and Prat, A. (2022). Competing models. *The Quarterly Journal of Economics*, 137(4):2419–2457.

Ning, Z. E., Shin, J., and Yu, J. (2025). Targeted advertising as implicit recommendation: strategic mistargeting and personal data opt-out. *Marketing Science*, 44(2):390–410.

O'Connor, J. and Wilson, N. E. (2021). Reduced demand uncertainty and the sustainability of collusion: How ai could affect competition. *Information Economics and Policy*, 54:100882.

Qian, K. and Jain, S. (2024). Digital content creation: An analysis of the impact of recommendation systems. *Management Science*, 70(12):8668–8684.

Ren, Q. (2024). Navigating the creator economy: An analysis of content promotion and view allocation policies on digital content platforms. *Marketing Science*.

Salant, Y. and Cherry, J. (2020). Statistical inference in games. *Econometrica*, 88(4):1725–1752.

Shin, J. and Yu, J. (2021). Targeted advertising and consumer inference. *Marketing Science*, 40(5):900–922.

Shin, M., Kim, J., and Shin, J. (2025). The adoption and efficacy of large language models: Evidence from consumer complaints in the financial industry. *Available at SSRN 5004194*.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *science*, 359(6380):1146–1151.

Wang, C.-Y. (2025). Balancing breadth and depth in optimal design of recommended choice sets. *Working Paper*.

Wu, Y. (2024). Creation, consumption, and control of sensitive content. *Marketing Science*, 43(4):885–902.

Yang, W., Yao, Y. J., and Zhou, P. (2024). Strategic misinformation generation and detection. *Working Paper*.

Zhong, Z. (2023). Platform search design: The roles of precision and price. *Marketing Science*, 42(2):293–313.

Zhou, B. and Zou, T. (2023). Competing for recommendations: The strategic impact of personalized product recommendations in online marketplaces. *Marketing Science*, 42(2):360–376.

Zou, T., Shi, Z. J., and Wu, Y. (2025). Welfare implications of democratization in content creation. *Working Paper*.

Zou, T., Wu, Y., and Sarvary, M. (2024). Designing recommendation systems on content platforms: Trading off quality and variety. *Available at SSRN*.

Online Appendix for "Designing Detection Algorithms for AI-Generated Content: Consumer Inference, Creator Incentives, and Platform Strategy"

Jieteng Chen, T. Tony Ke, and Jiwoong Shin

September 24, 2025

A.1 Detailed proof for Proposition 1

We provide an omitted portion of the proof for Proposition 1. In the main proof, we have that $\frac{\partial \mu_H(\widetilde{\sigma})}{\partial x} \propto -\frac{d}{dx}\left(\frac{F_A(x)}{F_H(x)}\right)$ and $\frac{\partial \mu_A(\widetilde{\sigma})}{\partial x} \propto -\frac{d}{dx}\left(\frac{1-F_A(x)}{1-F_H(x)}\right)$ without detailed steps. We provide detailed steps.

$$\begin{aligned}\frac{\partial \mu_H(\widetilde{\sigma})}{\partial x} &= -\frac{\lambda(1-\lambda)rF_H^2(x)\left[\overline{a}_D\overline{e}_D(1)\cdot(1-\overline{a}_T)\overline{e}_T(0)-(1-\overline{a}_D)\overline{e}_D(0)\cdot\overline{a}_T\overline{e}_T(1)\right]}{\left[\lambda m_T(H)+(1-\lambda)m_D(H)\right]^2}\frac{d}{dx}\left(\frac{F_A(x)}{F_H(x)}\right)\\ &\propto -\frac{d}{dx}\left(\frac{F_A(x)}{F_H(x)}\right).\end{aligned}$$

$$\begin{aligned}\frac{\partial \mu_A(\widetilde{\sigma})}{\partial x} &= -\frac{\lambda(1-\lambda)r(1-F_H(x))^2\left[\overline{a}_D\overline{e}_D(1)\cdot(1-\overline{a}_T)\overline{e}_T(0)-(1-\overline{a}_D)\overline{e}_D(0)\cdot\overline{a}_T\overline{e}_T(1)\right]}{\left[\lambda m_T(A)+(1-\lambda)m_D(A)\right]^2}\frac{d}{dx}\left(\frac{1-F_A(x)}{1-F_H(x)}\right)\\ &\propto -\frac{d}{dx}\left(\frac{1-F_A(x)}{1-F_H(x)}\right).\end{aligned}$$

A.2 Detailed proof for Proposition 2

We provide the omitted part of the proof for Proposition 2. We verify that $\overline{a}_D \in (0,1)$ given $\underline{\lambda} < \lambda < \widetilde{\lambda}$ and these two semi-separating equilibria are Pareto-dominated by Pool-1 equilibrium when $0 < K < \underline{K}$. We first define the threshold $\widetilde{\lambda}$ as follows:

$$\widetilde{\lambda} = \frac{F_A^2(\phi^{-1}(K))r^2v_o\theta}{F_A^2(\phi^{-1}(K))r^2v_o\theta + (\overline{q}-v_o)F_H^2(\phi^{-1}(K))}.$$

(i) Semi-A region: The type-D creators' mixing probability $\overline{a}_D$ is

$$\overline{a}_D^{semi_A} = \frac{[\lambda(\overline{q}-v_o)-(1-\lambda)r^2v_o]e_D^{semi_A}(0)(1-F_H(x))}{(1-\lambda)r^2v_o[e_D^{semi_A}(1)(1-F_A(x))-e_D^{semi_A}(0)(1-F_H(x))]}.$$

We show $\overline{a}_D^{semi_A} \in (0,1)$ if $\underline{\lambda} < \lambda < \widetilde{\lambda}$. Obviously, $\underline{\lambda} < \lambda$ implies that $\overline{a}_D^{semi_A} > 0$. Using $\lambda < \widetilde{\lambda}$, we have

$$\overline{a}_D^{semi_A} \leq \left(\frac{e_D^{semi_A}(1)(1-F_A(x))}{e_D^{semi_A}(0)(1-F_H(x))}-1\right)^{-1} \times \left(\frac{F_A^2(\phi^{-1}(K))}{F_H^2(\phi^{-1}(K))}\theta - 1\right),$$

Thus, to show $\bar{a}_D^{semi_A} < 1$, we only need to have $\frac{e_D^{semi_A}(1)}{e_D^{semi_A}(0)} \geq \frac{F_A(x^*)}{F_H(x^*)} \geq \frac{F_A(\phi^{-1}(K))}{F_H(\phi^{-1}(K))}$ and $\frac{1-F_A(x)}{1-F_H(x)} > 1 > \frac{F_A(\phi^{-1}(K))}{F_H(\phi^{-1}(K))}$. Then, we show that when $0 < K \leq \underline{K}$, consumer surplus in semi-A is lower than that in the pooling-1 equilibrium.

$$
\begin{aligned}
CS^{semi_A} &= \lambda e_T^{semi_A}(0)(q - v_o) - (1-\lambda) r \left[\bar{a}_D^{semi_A} e_D^{semi_A}(1) + (1 - \bar{a}_D^{semi_B}) e_D^{semi_A}(0)\right] v_o \\
&< e_T^{semi_A}\left[\lambda(q - v_o) - (1-\lambda) r^2 v_o\right] \\
&< e_T^{pool-1}\left[\lambda(q - v_o) - (1-\lambda) r^2 v_o\right] = CS^{pool-1}.
\end{aligned}
$$

(ii) Semi-H region: Creators' optimal efforts are $e_T^{semi_H} = \frac{F_H \delta_H^{semi_H}}{c}$, $e_D^{semi_H}(0) = \frac{r F_H \delta_H^{semi_H}}{c}$, and $e_D^{semi_H}(1) = \frac{r\theta F_A \delta_H^{semi_H}}{c}$. Type-$D$ creators' mixing probability $\bar{a}_D^{semi_H}$ is

$$
\bar{a}_D^{semi_H} = \frac{\lambda(\overline{q} - v_o) - (1-\lambda) r^2 v_o}{r^2 v_o (1-\lambda)} \left[\theta \left(\frac{F_A(x)}{F_H(x)}\right)^2 - 1\right]^{-1},
$$

We show $\bar{a}_D^{semi_H} \in (0,\,1)$ using $\lambda < \widetilde{\lambda}$ for $x^* < x < 1$.

$$
\bar{a}_D^{semi_H} \leq \left[\theta \left(\frac{F_A(x)}{F_H(x)}\right)^2 - 1\right]^{-1} \times \left(\frac{F_A^2(\phi^{-1}(K))}{F_H^2(\phi^{-1}(K))}\theta - 1\right) < 1,
$$

where the $<$ is due to $x > \phi^{-1}(K)$ and thus $\frac{F_A(x)}{F_H(x)} > \frac{F_A(\phi^{-1}(K))}{F_H(\phi^{-1}(K))}$. Moreover, consumer surplus is zero in semi-H and thus lower than that in Pool-1. Creators' profits are also lower due to $\delta_H^{semi_H} < 1$.

A.3 Detail Proof for Lemma 7

Proof of Claim 1

Proof. Given $\frac{K}{r^2} < \frac{\theta-1}{8c}$, we have $\lim\limits_{z \to 1} \frac{dy(z)}{dz} < 0$ and $\lim\limits_{z \to +\infty} \frac{dy(z)}{dz} > 0$. By continuity of $y(z)$ on z, there must exist a value $\hat{z}$ such that $y'(\hat{z}) = 0$. We prove the uniqueness of $\hat{z}$ by contradiction. Suppose there are multiple roots for $\frac{dy(z)}{dz} = 0$, there must exist a horizontal line $\hat{y}(z) = \hat{y}$ that intersects with $y(z)$ for more than three times such that for $z \in \{z_1, z_2, z_3, \cdots\}$, we have $y(z) = \hat{y}$. As $y(z)$ is a solution to Equation (5), we must have that for $z \in \{z_1, z_2, z_3, \cdots\}$,

$$
g_1(z; \hat{y}) = g_2(z; \hat{y}).
$$

where $g_1(z;\hat{y}) = (z-1)^2 \left(\frac{\hat{y}^2}{z^2} - \frac{1}{\theta}\right)$ and $g_2(z;\hat{y}) = \zeta(z - \frac{\hat{y}}{z})^2$ are functions of z parameterized by $\hat{y}$. Next, we only need to prove that $g_1(z;\hat{y})$ and $g_2(z;\hat{y})$ cannot have more than two interactions Let's consider two cases.

- For $\hat{y} \geq 1$, note that for $z \geq \sqrt{\theta}\hat{y}$, $g_2(z;\hat{y}) \leq 0$ and thus cannot intersect with $g_1(z;\hat{y})$. we show that $g_1(z;\hat{y})$ and $g_2(z;\hat{y})$ have only one intersection for $\hat{y} < z < \sqrt{\theta}\hat{y}$. As $\frac{dg_1(z;\hat{y})}{dz} =$

$\frac{2(z-1)}{\theta z^3}(\hat{y}^2\theta - z^3)$ and $\left.\frac{dg_1(z;\hat{y})}{dz}\right|_{z=\sqrt{\theta}\hat{y}} = -\frac{2}{y\theta^{\frac{3}{2}}}(\sqrt{\theta}\hat{y}-1)^2 < 0$, $g_1(z;\hat{y})$ can either first increase and then decrease with z; or always decrease with z. $g_2(z;\hat{y})$ monotonically increases with z because $\frac{dg_2(z;\hat{y})}{dz} = 2\zeta\left(z - \frac{\hat{y}}{z}\right)\left(1 + \frac{\hat{y}}{z^2}\right) > 0$. At $z = \hat{y}$, we have $g_1(\hat{y};\hat{y}) = (\hat{y}-1)^2(1-\frac{1}{\theta}) > \zeta(\hat{y}-1)^2 = g_2(\hat{y};\hat{y})$. At $z = \sqrt{\theta}\hat{y}$, we have $g_1(\sqrt{\theta}\hat{y};\hat{y}) = 0 < \zeta(\sqrt{\theta}\hat{y} - \frac{1}{\theta})^2 = g_2(\sqrt{\theta}\hat{y};\hat{y})$. This means that $g_1(z;\hat{y})$ and $g_2(z;\hat{y})$ have at most one interaction in z for $\hat{y} \geq 1$.

- For $0 < \hat{y} < 1$, we show that $g_1(z;\hat{y})$ and $g_2(z;\hat{y})$ have at most two intersections for $1 < z < \sqrt{\theta}\hat{y}$. As $\frac{dg_1(z;\hat{y})}{dz} \propto (\hat{y}^2\theta - z^3)$ and $(\hat{y}^2\theta - z^3)$ decreases with z, and we have $\left.\frac{dg_1(z;\hat{y})}{dz}\right|_{z=1} \propto (\theta\hat{y}^2 - 1) > 0$, which implies that $g_1(z;\hat{y})$ first increases and then decreases with z. And $g_2(z;\hat{y})$ still monotonically increases with z. At $z = 1$, we have $g_1(1;\hat{y}) = 0 < \zeta(\hat{y}-1)^2 = g_2(1;\hat{y})$. At $z = \sqrt{\theta}\hat{y}$, we have $g_1(\sqrt{\theta}\hat{y};\hat{y}) = 0 < \zeta(\sqrt{\theta}\hat{y} - \frac{1}{\theta})^2 = g_2(\sqrt{\theta}\hat{y};\hat{y})$. This means that $g_1(z;\hat{y})$ and $g_2(z;\hat{y})$ has at most two interactions for $1 < z < \sqrt{\theta}\hat{y}$.

As a result, $y(z)$ first decreases and then increases in z for $z \in (1, \infty)$. This means that for $z \in (1, z(x^*))$, $y(z)$ can either first decrease and then increase in z; or always decrease in z. □

Proof of Claim 2

Proof. Given $\frac{K}{r^2} \geq \frac{\theta-1}{8c}$, we have $\lim_{z\to 1}\frac{dy(z)}{dz} \geq 0$. We can prove that $y(z)$ must have a monotonic relationship with z by contradiction, following the same logic in Proof of Claim 1 above. □

A.4 Analysis for Extreme Belief Case

In our main text, we focus on the intermediate range of the fraction of truthful creators, $\underline{\lambda} < \lambda < \widetilde{\lambda}$, where consumers remain skeptical about the truthfulness of high-quality content and the benchmark outcome is a semi-separating equilibrium in the absence of detection. This is the region where algorithmic labeling is most consequential for shifting market equilibrium behavior. In this Online Appendix, we analyze the extreme belief cases where λ is either very low ($\lambda < \underline{\lambda}$) or very high ($\widetilde{\lambda} < \lambda$).

Case (i). *(*$\lambda < \underline{\lambda}$*)* When λ is very low, there only exists a pooling-$\emptyset$ equilibrium in which creators do not exert effort and content is all low-quality. Consumers' posterior belief upon seeing high-quality content is off-equilibrium-path and can be specified as 0.

Case (ii). *(*$\lambda > \widetilde{\lambda}$*)* When λ is very high, we provide the existence condition for two types of pure strategy equilibria and show that the creator's AI adoption strategies and consumer consumption strategy are invariant in x, in Lemma OA1. Note that $\widetilde{\lambda}$ is the lower bound such that for $\lambda > \widetilde{\lambda}$, two types of pure strategy equilibria are feasible under some detection threshold $x \in (0,1)$. The expression of $\widetilde{\lambda}$ is provided in Section A.2 of the online appendix.

Lemma OA1 (Separating Equilibria)**.** *Two types of separating equilibria exist:*

(i) ***(Sep-All)*** *If* $\underline{K} \leq K \leq \overline{K}$ *and* $\lambda^{sep_A} \leq \lambda < 1$*, consumers fully consume* L_A *content* ($\delta_A = 1$) *and* L_H *content* ($\delta_H = 1$)*. The type-*T *creators avoid AI, and the type-*D *creators always adopt AI.*

(ii) ***(Sep-H)*** *If* $\underline{K} < K \leq \phi(x)$ *and* $\underline{\lambda}^{sep_H} \leq \lambda < \overline{\lambda}^{sep_H}$*, consumers fully consume* L_A *content* ($\delta_A = 1$) *and never consume* L_H *content* ($\delta_H = 0$)*. The type-*T *creators avoid AI, and the type-*D *creators always adopt AI.*

Proof. ***(i) Sep-All equilibrium***

Given consumers' strategy $\delta_c^{sep_A} = (1,1)$, the type-T creators have an incentive to adopt AI and type-D creators do not if and only if $\underline{K} \leq K \leq \overline{K}$. Two types of creators' optimal effort levels can be solved by $e_T^{sep_A}(0) = \frac{1}{c}$ and $e_D^{sep_A}(1) = \frac{r\theta}{c}$. Then, consumers consume AI-labeled content if and only if $\mu_A = \frac{\lambda e_T^{sep_A}(0)(1-F_H(x))}{\lambda e_T^{sep_A}(0)(1-F_H(x))+(1-\lambda) r e_D^{sep_A}(1)(1-F_A(x))} \geq \frac{v_o}{\overline{q}} \Leftrightarrow \lambda \geq \lambda^{sep_A} \equiv \frac{(1-F_A(x))r^2\theta v_o}{(1-F_A(x))r^2\theta v_o+(1-F_H(x))(\overline{q}-v_o)}$, where $\lambda^{sep_A} \in (\widetilde{\lambda}, 1)$ due to $\frac{1-F_A(x)}{1-F_H(x)} > 1 > \frac{F_A(x)}{F_H(x)}$.

(ii) Sep-H equilibrium

Given consumers' strategy $\delta_c^{sep_H} = (1,0)$, the type-T creators avoid AI and the type-D creators adopt AI if and only if

$$\begin{cases} \max\limits_{e_T}\{e_T \cdot F_A(x) - \frac{c}{2\theta} \cdot e_T^2\} - K \leq \max\limits_{e_T}\{e_T \cdot F_H(x) - \frac{c}{2} \cdot e_T^2\} \\ \max\limits_{e_D}\{r \cdot e_D \cdot F_A(x) - \frac{c}{2\theta} \cdot e_D^2\} - K \geq \max\limits_{e_D}\{r \cdot e_D \cdot F_H(x) - \frac{c}{2} \cdot e_D^2\} \end{cases} \Leftrightarrow \frac{\phi(x)}{r^2} \leq K \leq \phi(x).$$

Then, two types of creators' equilibrium efforts are $e_T^{sep_H}(0) = \frac{F_H(x)}{c}$ and $e_D^{sep_H}(1) = \frac{r\theta F_A(x)}{c}$. Therefore, consumers are willing to consume content with L_H, but not content L_A if and only if

$$\begin{cases} \mu_H &= \frac{\lambda e_T^{sep_H}(0)F_H(x)}{\lambda e_T^{sep_H}(0)F_H(x)+(1-\lambda) r e_D^{sep_H}(1)F_A(x)} \geq \frac{v_o}{\overline{q}}. \\ \mu_A &= \frac{\lambda e_T^{sep_H}(0)(1-F_H(x))}{\lambda e_T^{sep_H}(0)(1-F_H(x))+(1-\lambda) r e_D^{sep_H}(1)(1-F_A(x))} < \frac{v_o}{\overline{q}}. \end{cases} \Leftrightarrow \underline{\lambda}^{sep_H} \leq \lambda < \overline{\lambda}^{sep_H},$$

where $\underline{\lambda}^{sep_H} = \frac{F_A^2(x) r^2 v_o \theta}{F_A^2(x) r^2 v_o \theta + F_H^2(x)(\overline{q}-v_o)}$ and $\overline{\lambda}^{sep_H} = \frac{F_A(x)(1-F_A(x)) r^2 v_o \theta}{F_A(x)(1-F_A(x)) r^2 v_o \theta + F_H(x)(1-F_H(x))(\overline{q}-v_o)}$.

Lastly, we show that this equilibrium is Pareto-dominated by the Pool-1 equilibrium when $0 < K \leq \underline{K}$. Creators' profits are lower due to $\delta_A = 0$, and consumer surplus is also lower than that in the Pool-1 equilibrium as follows

$$\begin{aligned} CS^{sep_H} &= \lambda e_T^{sep_H}(0)F_H(x)(\overline{q}-v_o) - (1-\lambda) r e_D^{sep_H}(1)F_A(x)v_o < e_T^{sep_H}(0)F_H(x)\left[\lambda(\overline{q}-v_o) - (1-\lambda)r^2 v_o\right] \\ &< e_T^{pool-1}\left[\lambda(\overline{q}-v_o) - (1-\lambda)r^2 v_o\right] = CS^{pool-1}. \end{aligned}$$

Therefore, the Sep-H equilibrium exists only when $\lambda \in [\underline{\lambda}^{sep_H}, \overline{\lambda}^{sep_H})$ and $K \in (\underline{K}, \phi(x)]$. Notice that the interval of $(\underline{K}, \phi(x)]$ is non-empty if and only if $x > \phi^{-1}(\underline{K})$. In this case, we can have $\underline{\lambda}^{sep_H} > \widetilde{\lambda}$. $\square$