Content Filtering with Inattentive Information Consumers*

Ian Ball[1], James Bono[2], Justin Grana[3], Nicole Immorlica[4], Brendan Lucier[4], Aleksandrs Slivkins[5]

[1]MIT, 77 Massachusetts Ave, Cambridge, MA 02139, USA.
[2]Microsoft, 1 Microsoft Way, Redmond, WA 98052, USA.
[3]Edge & Node, remote only.
[4] Microsoft Research, 1 Memorial Dr, Cambridge, MA 02142, USA.
[5] Microsoft Research, 300 Lafayette St, New York, NY 10012, USA.
justin@edgeandnode.com, slivkins@microsoft.com.

Abstract

I We develop a model of content filtering as a game between the filter and the content consumer, where the latter incurs information costs for examining the content. Motivating examples include censoring misinformation, spam/phish filtering, and recommender systems acting on a stream of content. When the attacker is exogenous, we show that improving the filter's quality is weakly Pareto improving, but has no impact on equilibrium payoffs until the filter becomes sufficiently accurate. Further, if the filter does not internalize the consumer's information costs, its lack of commitment power may render it useless and lead to inefficient outcomes. When the attacker is also strategic, improvements in filter quality may decrease equilibrium payoffs.

1 Introduction

Content filtering is a crucial and widely-applied tool for improving the experience of information consumers. Email filters automatically sort normal, malicious and spam messages, increasing security and saving users from manually sorting mail (Gangavarapu, Jaidhar, and Chanduka 2020; Chae et al. 2017; Bhowmick and Hazarika 2018). Information aggregators and social media platforms have deployed content filters that censor non-credible and potentially deceptive claims (Aldwairi and Alwahedi 2018; Kumar and Geethakumari 2014). Recommender systems learn consumers' preferences to save them from having to sift through unwanted content (Bagher, Hassanpour, and Mashayekhi 2017; Wei, Moreau, and Jennings 2003; Bergemann and Ozmen 2006).[1] Despite major efforts to improve content filters, information consumers remain susceptible to malicious or illegitimate content, *e.g.,* they click on phishing messages (Blythe, Petrie, and Clark 2011; Benenson, Gassmann, and Landwirth 2017) and fall victim to misinformation (Roozenbeek et al. 2020; Pennycook and Rand 2019).

Consumers can take measures to avoid the malicious content. For example, a recipient of a suspicious email could examine the email more carefully, do a quick web search for known malicious patterns, ask an acquaintance's opinion, or even attempt to reach the purported sender by other means. A social media user could carefully check the argumentation in a given post, or consult reputable sources. However, such measures incur substantial costs in time, effort and attention. In particular, the literature on "attention economy" documents that attention in the digital sphere is a scarce resource (Hendricks and Vestergaard 2019). We will refer to these costs as *information costs*[2].

Due to information costs, consumers tend to strategically alter their behavior in response to the (perceived) filter quality. When consumers perceive that a filter is poor, either allowing too much malicious content or censoring too much, they abandon the platform (a risk acknowledged by major platforms for email, social media and news (D'Onfro 2018)). When the filter is exceptional, consumers take content at face value (Sterrett et al. 2019). In the "middle ground," the filter is imperfect and consumers choose whether/how to examine the content to determine its quality.[3]

The considerable investment in improving content filters and consumers' strategic allocation of scarce attention motivates three salient questions:

(Q1) Can the benefits of an increase in filter quality be crowded out by reduced consumer attention in response to the increase in filter quality?

(Q2) If the filter's payoffs do not depend on the consumer's information costs, what inefficiencies (i.e. sub-optimal equilibria) arise and how can they be abated?

(Q3) How does the interaction between the filter and consumer change when the attacker *strategically* crafts its attack in anticipation of this interaction? How does this affect the cost-benefit tradeoff for improving the filter quality?

To answer these questions, we model content filtering as

*All proofs are in the full version, at arXiv:2205.14060. JG and IB were affiliated with Microsoft during this research.

[1]Our model is most relevant to recommender systems that process a stream of items such as new event announcements: *e.g.,* new concerts for a music app, or new properties for a real estate app.

[2]An alternative term, *attention costs*, is also well-established.

[3]An ironic example: a conference serves as a filter for academic publications, and its reputation (*i.e.,* perceived filter quality) is often used to evaluate the merit of a scientific claim (Sangster 2015).

a game between a filter and an information consumer. The filter receives a batch of content, wherein each piece is either legitimate or malicious with some exogenously specified probability. For each piece of content, the filter receives a signal regarding its legitimacy, and either blocks it or forwards it to the consumer. In the latter case, the consumer exerts costly effort to examine the content and then decides whether to accept or ignore it. Both players benefit when the consumer accepts legitimate content, and incur a cost when it does not consume legitimate content or consumes malicious content. In an extension, an *endogenous* attacker sets the mean amount of the malicious content (*attack propensity*) to maximize the expected amount of malicious content the consumer ultimately accepts.

The key novelty is that the consumer strategically chooses the fidelity of its signal and incurs the corresponding information cost. This represents strategic information acquisitions where consumers optimally trade off the physical and cognitive costs of obtaining higher fidelity signals with the benefit associated with the higher fidelity information. We adopt *rational inattention* (Sims 2003), a standard model for consumer's information cost. Specifically, cost is proportional to the expected drop in entropy between the consumer's prior and posterior.[4] The filter may internalize these costs, aiming to maximize consumers' welfare.[5] We also consider a variant in which the filter does *not* internalize the information costs, *e.g.,* when it only cares about detection rates, which may be the case when platforms compete in performance benchmarks. We call these variants, resp., *aligned utilities* and *semi-aligned utilities*.

With this model, we answer our questions as follows:

(A1) With an exogenous attacker and aligned utilities, increasing filter quality is Pareto-improving, but only weakly (Theorem 4.2). There is a "barrier to entry": equilibrium outcomes improve only when the filter is accurate enough.

(A2) A new inefficiency arises when we switch to semi-aligned utilities. Since the filter does not internalize the consumer's information cost, the filter is biased toward forwarding more content. It may not be credible for the filter to block any content, thus introducing a Pareto inefficiency (Theorem 5.1). However, this inefficiency vanishes once the filter is sufficiently accurate (Theorem 5.2), upon which further increases to filter quality are Pareto improving (Theorem 5.3).

(A3) With a strategic attacker, there are two surprising consequences: the consumer does not examine any content in any equilibrium (Theorem 6.1), and improving the filter can make both the filter and the consumer worse off (Theorem 6.2). The attacker raises its attack propensity, and this outweighs the direct benefit of a more accurate filter.

The main practical implication of our results is that rote marginal improvements in filter quality are not unambiguously beneficial. These improvements should either be large enough, or be coupled with other interventions (such as training to decrease information costs), to avoid a damaging reduction in consumer attention.

Conceptually, we identify strategic interaction between content filters and information consumers as a relevant aspect of content filtering. In contrast, prior game-theoretic work on content filtering studies games between filters and attackers (e.g., Lu and Niu 2015; Laszka, Lou, and Vorobeychik 2016), between filters and a mediator (Ben-Porat and Tennenholtz 2018), or between consumers (Acemoglu, Ozdaglar, and Siderius 2021). Adversarial machine learning (Vorobeychik and Kantarcioglu 2018; Joseph et al. 2019) studies attacks on machine learning algorithms (such as content filters). In all this work, consumers naively follow the filter's recommendations. We show that filter-consumer strategic interaction is not captured by attacker-filter games.

While our model may appear similar to models in information design (Kamenica 2019; Candogan and Drakopoulos 2020), and especially information design with rational inattention (Matyskova and Montes 2023), these models are fundamentally different: senders can design arbitrary Blackwell (Blackwell 1951) experiments that generate the receiver's signal. In our model, the filter chooses an *action* that has a direct impact on utility as well as consumer beliefs. This coupling between actions and consumer beliefs is what sets our model apart from those of information design and yields new results.

Our model is similar to (Papanastasiou 2020, P2020 for short) in that they both consider binary environments where a filter and consumers inspect content before choosing an action. However, because consumers in our model choose their signal quality and the filter's signal is noisy (unlike that in P2020), we examine the utility and behavioral impacts in changing filter quality, which is absent in P2020. Additionally, we extend the environment and consider an endogenous attacker, another feature not included in P2020.

2 Our Model and Preliminaries

We consider the *content-filtering game*: a game between two strategic players, an info filter and an info consumer that make decisions about content's legitimacy. We call them the *filter* and the *consumer*, and denote the resp. notation with subscripts $\mathtt{f}$ and $\mathtt{c}$. The game's protocol is as follows:

1. The filter receives a batch of content (*e.g.,* a day's worth of news). The batch consists of malicious content that arrives at a Poisson rate of ρ_0 and legitimate content that arrives at a Poisson rate of ρ_1, per unit time interval. Both rates are common knowledge. W.l.o.g., we normalize $\rho_1 = 1$.

Each piece of content in the batch is identified with a binary random variable X, where $X = 0$ means "malicious" and $X = 1$ means "legitimate." We define

$$q := \Pr[X = 0] = \rho_0/(\rho_0 + 1).$$

[4] Despite alternatives (Milgrom and Weber 1982; Vives 1984; Zhong 2022; Pomatto, Strack, and Tamuz 2023; Caplin, Dean, and Leahy 2022; Gabaix 2019), rational inattention is widely adopted as a standard model for information costs (Martin 2017; Bertoli, Moraga, and Guichard 2020; Ravid 2020; Maćkowiak and Wiederholt 2015; Jiang, Fosgerau, and Lo 2020; Acharya and Wee 2020; Dasgupta and Mondria 2018), in the absence of further behavioral evidence or assumptions (Caplin 2016).

[5] Maximizing users' welfare is a common modeling choice and a reasonable proxy for many online platforms that indirectly profit from user engagement, *e.g.,* via advertising.

2. Each piece of content $X \in \{0,1\}$ is processed by the **filter** as follows. The filter receives a private signal $\Psi_{\mathtt{f}} \in \{0,1\}$ about the content type, representing the output of a classifier so that $\Psi_{\mathtt{f}} = 0$ means "likely malicious" and $\Psi_{\mathtt{f}} = 1$ means "likely legitimate". The signal is drawn independently from a known conditional distribution given X. Denote the resp. true and false positive rates as

$$\pi_x = \Pr[\,\Psi_{\mathtt{f}} = 0 \mid X = x\,], \quad x \in \{0,1\}. \tag{1}$$

W.l.o.g. assume $\pi_0 \geq \pi_1$ (since the filter is free to choose its action conditional on its signal). After receiving the signal, the filter chooses its *action* $a_{\mathtt{f}} \in \{0,1\}$: whether to block the content ($a_{\mathtt{f}} = 0$) or to forward it to the consumer ($a_{\mathtt{f}} = 1$).

3. Each piece of forwarded content is processed by the **consumer** as follows. The consumer chooses how to examine the content. Formally, the consumer controls the distribution of a signal $\Psi_{\mathtt{c}} \in \{0,1\}$, where $\Psi_{\mathtt{c}} = 0$ means "likely malicious" and $\Psi_{\mathtt{c}} = 1$ means "likely legitimate". The signal is drawn independently from some conditional distribution given X, characterized by

$$\tilde{\pi}_x = \Pr[\,\Psi_{\mathtt{c}} = 0 \mid X = x\,], \quad x \in \{0,1\}. \tag{2}$$

These probabilities are chosen by the consumer in advance, at the (information) cost specified below. Then, the consumer chooses its *action* $a_{\mathtt{c}} \in \{0,1\}$: whether to accept the content as legitimate ($a_{\mathtt{c}} = 1$) or to ignore it ($a_{\mathtt{c}} = 0$).

Strategies. The filter and the consumer have pure action strategies $s_{\mathtt{f}}, s_{\mathtt{c}} : \{0,1\} \to \{0,1\}$ so that $a_{\mathtt{f}} = s_{\mathtt{f}}(\Psi_{\mathtt{f}})$ and $a_{\mathtt{c}} = s_{\mathtt{c}}(\Psi_{\mathtt{c}})$. The consumer also chooses probabilities $\mu = (\,\tilde{\pi}_0, \tilde{\pi}_1\,)$ from Eq. (2), called its *information strategy*. Thus, pure strategies are $s_{\mathtt{f}}$ for the filter, and $(s_{\mathtt{c}}, \mu)$ for the consumer. Both players choose their (mixed) strategies before the game starts, and those strategies are applied to the entire batch. (This is justified because the pieces of content are ex-ante equivalent.) We posit that the filter and the consumer choose their (mixed) strategies simultaneously, *i.e.,* without observing one another.

Remark 2.1. *When the filter and consumer have fully aligned utilities (as defined below and discussed in Sections 4, 6), our results carry over to the variant where the players choose their mixed strategies sequentially: the filter moves first, and the consumer best-responds. This is because our results focus on the socially optimal strategy profile (defined in Section 4), which is the same in both variants.*

Remark 2.2. *One pure strategy for the consumer is to* not *examine the content and incur no info cost.*

Notation. A generic mixed strategy profile is denoted $\boldsymbol{\sigma}$. The players' mixed action strategies are, resp., $\boldsymbol{\sigma}_{\mathtt{f}}$ and $\boldsymbol{\sigma}_{\mathtt{c}}$.

We label three filter pure strategies: the *blocking strategy* $s_{\mathtt{blk}}$ which always blocks the content: $s_{\mathtt{blk}}(\cdot) \equiv 0$, the *forwarding strategy* $s_{\mathtt{fwd}}$ which always forwards the content: $s_{\mathtt{fwd}}(\cdot) \equiv 1$, and the *differentiating strategy* $s_{\mathtt{dif}}$ which differentiates between the signals: $s_{\mathtt{dif}}(\psi) \equiv \psi$. We ignore the "unreasonable strategy" in which the filter forwards "likely malicious" content and blocks content that is "likely clean" as it can never be part of a non-trivial equilibrium (see the full version for technical details).

A strategy profile is called *consumer-optimal* if the consumer best-responds to the filter's strategy. Let the *blocking profile* $\sigma_{\mathtt{blk}}$, the *forwarding profile* $\sigma_{\mathtt{fwd}}$, and the *differentiating profile* $\sigma_{\mathtt{dif}}$, be consumer-optimal strategy profiles in which the filter's pure strategy is, resp., $s_{\mathtt{blk}}$, $s_{\mathtt{fwd}}$, and $s_{\mathtt{dif}}$.

	$s_{\mathtt{f}}(1) = 0$	$s_{\mathtt{f}}(1) = 1$
$s_{\mathtt{f}}(0) = 0$	*blocking profile* $\sigma_{\mathtt{blk}}$	*different. profile* $\sigma_{\mathtt{dif}}$
$s_{\mathtt{f}}(0) = 1$	($s_{\mathtt{f}}$ is "unreasonable")	*forwarding profile* $\sigma_{\mathtt{fwd}}$

Utilities. The consumer's utility per piece of content is the difference between the *action payoff* $u(a_{\mathtt{f}} \cdot a_{\mathtt{c}}, X)$, determined by how the actions match the content type, and the *information cost* for examining the content. We interpret the product $a_{\mathtt{f}} \cdot a_{\mathtt{c}} \in \{0,1\}$ as an aggregate action: indeed, the content is accepted if $a_{\mathtt{f}} \cdot a_{\mathtt{c}} = 1$, and ignored otherwise. The consumer receives a reward when legitimate content is accepted ($a_{\mathtt{f}} \cdot a_{\mathtt{c}} = X = 1$), and penalties if the content is misclassified ($a_{\mathtt{f}} \cdot a_{\mathtt{c}} \neq X$). We normalize action payoffs to 0 if malicious content is ignored ($a_{\mathtt{f}} \cdot a_{\mathtt{c}} = X = 0$). Thus, action payoffs $u(a_{\mathtt{f}} \cdot a_{\mathtt{c}}, X)$ are summarized by a 2×2 table below, with $b, c_1, c_2 \geq 0$.

	$X = 0$	$X = 1$
$a_{\mathtt{f}} \cdot a_{\mathtt{c}} = 0$	0	$-c_1$
$a_{\mathtt{f}} \cdot a_{\mathtt{c}} = 1$	$-c_2$	b

The information cost is the cost of obtaining signal $\Psi_{\mathtt{c}}$ about content type X. It is proportional to how far the consumer's beliefs shift away from its prior, and only accrues when the filter does not block content. More abstractly, we define the information cost for obtaining some randomized signal Ψ about some hidden state X given some event $\mathcal{E}$, denoted $\mathtt{C}[\,\Psi; X \mid \mathcal{E}\,]$ and determined by the conditional joint distribution of (Ψ, X) given $\mathcal{E}$. We adopt the (widely accepted) definition from Sims (2003).

$$\mathtt{C}[\,\Psi; X \mid \mathcal{E}\,] = \lambda \cdot I\,(\,\Psi; X \mid \mathcal{E}\,)\,, \tag{3}$$

where $I\,(\,\Psi; X \mid \mathcal{E}\,) \geq 0$ is the mutual information conditional on the event $\mathcal{E}$ and $\lambda > 0$ is a known parameter. Thus, the information cost for examining the content is defined via (3) as $\mathtt{C}[\,\Psi_{\mathtt{c}}; X \mid a_{\mathtt{f}} = 1\,]$. Note that the cost indirectly depends on filter's mixed action strategy since information costs are a function of the consumers prior upon receiving content, which depends on the filter's strategy.

The consumer's expected payoff per a random piece of content X under mixed strategy profile $\boldsymbol{\sigma}$ is therefore

$$v_{\mathtt{c}}(\boldsymbol{\sigma}) = \mathbb{E}\left[\,u(a_{\mathtt{c}} \cdot a_{\mathtt{f}}, X) - a_{\mathtt{f}} \cdot \mathtt{C}[\,\Psi_{\mathtt{c}}; X \mid a_{\mathtt{f}} = 1\,]\,\right],$$

where the expectation is over $X, \Psi_{\mathtt{f}}, \Psi_{\mathtt{c}}, \boldsymbol{\sigma}$. As a shorthand, let $u(\boldsymbol{\sigma}) = \mathbb{E}[\,u(a_{\mathtt{c}} \cdot a_{\mathtt{f}}, X)\,]$ and $\mathtt{C}(\boldsymbol{\sigma}) = \mathtt{C}[\,\Psi_{\mathtt{c}}; X \mid a_{\mathtt{f}} = 1\,]$ be the corresponding expected action payoff and information cost.

The consumer's total expected utility over the batch is

$$V_{\mathtt{c}}(\boldsymbol{\sigma}) = (1 + \rho_0)\; v_{\mathtt{c}}(\boldsymbol{\sigma}) = v_{\mathtt{c}}(\boldsymbol{\sigma})/(1-q). \tag{4}$$

where $1 + \rho_0$ represents the expected batch size.

To define the filter's utility, we consider two variants. The main variant (*aligned utilities*) is that the filter's utility equals the consumer's. We also consider another variant (*semi-aligned utilities*) when the filter internalizes the action costs but not the information costs. Let $V_{\mathtt{f}}(\boldsymbol{\sigma})$ be filter's total expected utility under profile $\boldsymbol{\sigma}$. Then $V_{\mathtt{f}}(\boldsymbol{\sigma}) = V_{\mathtt{c}}(\boldsymbol{\sigma})$ for aligned utilities, and $V_{\mathtt{f}}(\boldsymbol{\sigma}) = u(\boldsymbol{\sigma})/(1-q)$ for semi-aligned utilities.

Value of Technological Change. We are particularly interested in how improving the technology impacts equilibrium outcomes. Specifically, we consider improving the quality of the filter, in terms of raising π_0 and/or lowering π_1.[6] We adopt Perfect Bayesian Equilibrium (PBE) as a solution concept (Mas-Colell et al. 1995).

For concreteness, fix some equilibrium selection rule, f, (Matsui and Matsuyama 1995) and filter quality parameters, π_0 and π_1. For each player $i \in \{\mathtt{f}, \mathtt{c}\}$, let $V_i^f(\pi_0, \pi_1)$ be i's equilibrium payoff under this rule. We are interested in the difference in equilibrium payoffs between a high- and low-quality filter:

$$V_i^f(\pi_0', \pi_1') - V_i^f(\pi_0, \pi_1): \quad i \in \{\mathtt{f}, \mathtt{c}\}, \tag{5}$$

where $\pi_0' \geq \pi_0$ and $\pi_1' \leq \pi_1$. We call (5) the *value of technological change* (VoTC). We say that VoTC is positive (resp., negative) if Eq. (5) is that way for both players, *i.e.,* if improving the filter Pareto-increases (resp., Pareto-decreases) equilibrium payoffs.

Consider VoTC under infinitesimal filter improvement:

$$\frac{\partial}{\partial \pi_0} V_i^f(\pi_0, \pi_1),\ -\frac{\partial}{\partial \pi_1} V_i^f(\pi_0, \pi_1):\ i \in \{\mathtt{f}, \mathtt{c}\}, \tag{6}$$

assuming the partial derivatives in (6) are well-defined. We call (6) the *Marginal Value of Technology* (MVoT). The MVoT specifies how much a rational filter would pay to improve its quality. A zero (resp., negative) MVoT means the filter would not pay anything (resp., would have to *be paid*).

3 Consumer Beliefs

This section presents a preliminary analysis of consumer behavior, which applies to both aligned and semi-aligned utilities, and serves as scaffolding for what follows.

An important quantity is the consumer's belief that the forwarded content is malicious, given that the filter's mixed action strategy is $\boldsymbol{\sigma}_{\mathtt{f}}$. We define this quantity as:

$$q(\boldsymbol{\sigma}_{\mathtt{f}}) := \Pr[X = 0 \mid \boldsymbol{\sigma}_{\mathtt{f}}, a_{\mathtt{f}} = 1]. \tag{7}$$

Note that $q(s_{\mathtt{fwd}})$ is simply $q := \Pr[X = 0]$.

The following lemma shows that the consumer's behavior is uniquely determined by $q(\boldsymbol{\sigma}_{\mathtt{f}})$:

Lemma 3.1. *Given any filter mixed strategy $\boldsymbol{\sigma}_{\mathtt{f}} \neq s_{\mathtt{blk}}$ the consumer's best response to $\boldsymbol{\sigma}_{\mathtt{f}}$ is determined by $q(\boldsymbol{\sigma}_{\mathtt{f}})$.*

As per Remark 2.2, the consumer can choose to *not* examine the content and incur no information costs. Below we establish a regime where that is indeed optimal. Define:

$$\begin{aligned} 1 > q_{\mathtt{H}} &:= \frac{\exp(b/\lambda) - \exp(-c_1/\lambda)}{\exp(b/\lambda) - \exp(-(c_1 + c_2)/\lambda)} \\ &> q_{\mathtt{L}} := q_{\mathtt{H}} \cdot \exp(-c_2/\lambda) > 0. \end{aligned} \tag{8}$$

[6] Filter's quality takes two numbers to describe.

Proposition 3.2. *Let $\boldsymbol{\sigma}$ be a consumer-optimal mixed strategy profile with filter's mixed action strategy $\boldsymbol{\sigma}_{\mathtt{f}} \neq s_{\mathtt{blk}}$. Then $\mathtt{C}(\boldsymbol{\sigma}) = 0$ if and only if $q(\boldsymbol{\sigma}_{\mathtt{f}}) \notin (q_{\mathtt{L}}, q_{\mathtt{H}})$. Furthermore, if $q(\boldsymbol{\sigma}_{\mathtt{f}}) \leq q_{\mathtt{L}}$ the consumer's optimal strategy is to accept all content. If $q(\boldsymbol{\sigma}_{\mathtt{f}}) > q_{\mathtt{H}}$ the consumer's optimal strategy is to ignore all content.*

In words, if unblocked content is too likely to be malicious (resp., legitimate) for a given $\boldsymbol{\sigma}_{\mathtt{f}}$, the consumer's best-response is to ignore (resp., accept) it without examination.

Remark 3.3. *The quantities $q(\boldsymbol{\sigma}_{\mathtt{f}})$, $q_{\mathtt{H}}$, $q_{\mathtt{L}}$ are meaningful as per Proposition 3.2. They usefully encapsulate the numerous parameters in our model, and are essential in our subsequent results. Note that (7) is determined by the joint distribution of X and the filter's signal $\Psi_{\mathtt{f}}$, whereas (8) is determined by all parameters related to the costs.*

We now derive the MVoT under some consumer-optimal profiles in some parameter regimes. A key quantity here is

$$q_{\mathtt{dif}} := q(s_{\mathtt{dif}}) = \Pr[X = 0 \mid s_{\mathtt{dif}}, \Psi_{\mathtt{f}} = 1] < q, \tag{9}$$

where the inequality follows because $\pi_0 \geq \pi_1$.

Proposition 3.4. *For $i \in \{\mathtt{f}, \mathtt{c}\}$ and $x \in \{0, 1\}$:*

(a) ***Zero MVoT.*** $\frac{\partial}{\partial \pi_x} V_i(\sigma_{\mathtt{fwd}}) = \frac{\partial}{\partial \pi_x} V_i(\sigma_{\mathtt{blk}}) = 0$. *If $q_{\mathtt{dif}} > q_{\mathtt{H}}$ then $\frac{\partial}{\partial \pi_x} V_i(\sigma_{\mathtt{dif}}) = 0$.*

(b) ***Constant MVoT.*** *If $q_{\mathtt{dif}} < q_{\mathtt{L}}$ then*

$$\begin{aligned} \frac{\partial}{\partial \pi_0} V_i(\sigma_{\mathtt{dif}}) &= \frac{q}{1-q} c_2 > 0 \\ \frac{\partial}{\partial \pi_1} V_i(\sigma_{\mathtt{dif}}) &= -(c_1 + b) < 0. \end{aligned}$$

In words, there is no benefit to improving the filter if the filter's action does not depend on its signal, or the consumer's best response is simply to ignore all content. On the other hand, if the consumer accepts the filter's recommendation, then MVoT is constant. To fully characterize the MVoT, subsequent analysis will focus on deriving the MVoT when $q_{\mathtt{L}} < q(\boldsymbol{\sigma}) < q_{\mathtt{H}}$ *and* establishing which profile constitutes an equilibrium.

4 Aligned Utilities ($V_{\mathtt{f}} = V_{\mathtt{c}}$)

In this section, we consider *aligned utilities*. Let $V := V_{\mathtt{f}} = V_{\mathtt{c}}$. We focus on socially optimal profiles (ones that maximize V), noting that any such profile is an equilibrium. Let $V^* = V_i^f(\pi_0, \pi_1)$, where f chooses the equilibrium that maximizes V among all equilibria. Our first result is that V^* has a simple characterization in terms of two pure profiles defined in Section 2, the differentiating profile $\sigma_{\mathtt{dif}}$ and forwarding profile $\sigma_{\mathtt{fwd}}$:

Proposition 4.1. $V^* = \max(V(\sigma_{\mathtt{dif}}), V(\sigma_{\mathtt{fwd}}))$.

While it is straightforward to algebraically demonstrate which of these two profiles are the best among the *pure* strategy profiles, it is more difficult to prove that there is no benefit from the filter using a mixed strategy. Indeed, in our game the payoffs at a mixed equilibrium are not (necessarily) linear in the mixing probabilities, because the latter enter non-linearly in the information costs. Consequently, it is no longer trivially guaranteed that some pure strategy profile is socially optimal.

The main result here fully characterizes the marginal value of technological change (MVoT) in terms of V^*.

Theorem 4.2.

(a) ***Zero MVoT.*** *Suppose* $q_{\mathtt{dif}} > q_{\mathrm{H}}$ *or* $V(\sigma_{\mathtt{dif}}) < V(\sigma_{\mathtt{fwd}})$. *Then* $\partial V^*/\partial\pi_0 = \partial V^*/\partial\pi_1 = 0$.

(b) ***Constant MVoT.*** *If* $q_{\mathtt{dif}} < q_{\mathrm{L}}$ *and* $V(\sigma_{\mathtt{dif}}) > V(\sigma_{\mathtt{fwd}})$,

$$\frac{\partial V^*}{\partial \pi_0} = \frac{q}{1-q} c_2 > 0 \quad \textit{and} \quad \frac{\partial V^*}{\partial \pi_1} = -(c_1 + b) < 0.$$

(c) ***Non-constant MVoT.*** *Suppose* $q_{\mathtt{dif}} \in (q_{\mathrm{L}}, q_{\mathrm{H}})$ *and* $V(\sigma_{\mathtt{dif}}) > V(\sigma_{\mathtt{fwd}})$. *Then*

$$\frac{\partial V^*}{\partial \pi_0} = \frac{q}{1-q} \lambda \cdot \log\left(\frac{q_{\mathrm{H}}}{q_{\mathtt{dif}}}\right) > 0 \tag{10}$$

$$\frac{\partial V^*}{\partial \pi_1} = \lambda \cdot \log\left(\frac{1-q_{\mathrm{H}}}{1-q_{\mathtt{dif}}}\right) < 0.$$

The main insight of Theorem 4.2 is that MVoT is weakly *but not strictly* positive. That is, when incentives are aligned, improving the filter quality can never hurt the players, though in some cases it may have no impact. Moreover, we fully characterize MVoT behavior based on how $q_{\mathtt{dif}}$ compares with $(q_{\mathrm{L}}, q_{\mathrm{H}})$, and whether $V(\sigma_{\mathtt{dif}}) < V(\sigma_{\mathtt{fwd}})$. This is summarized in the table below. (In this table, VoTC is positive in both cells in which it is not zero.)

	$q_{\mathtt{dif}} < q_{\mathrm{L}}$	$q_{\mathtt{dif}} \in (q_{\mathrm{L}}, q_{\mathrm{H}})$	$q_{\mathtt{dif}} > q_{\mathrm{H}}$
$V^*(\sigma_{\mathtt{dif}}) > V^*(\sigma_{\mathtt{fwd}})$	Constant	Non-linear	Zero
$V^*(\sigma_{\mathtt{dif}}) < V^*(\sigma_{\mathtt{fwd}})$		Zero VoTC	

We have two *barriers to entry* in filter technology. First, recall that we have zero MVoT when $q_{\mathtt{dif}} > q_{\mathrm{H}}$, and note that the filter quality is higher for lower values of $q_{\mathtt{dif}}$. Therefore, the filter must be of sufficiently high quality for improvements to make a difference.

Second, if $V(\sigma_{\mathtt{fwd}}) > V(\sigma_{\mathtt{dif}})$ then improving the filter does not help, either. In particular, the forwarding profile $\sigma_{\mathtt{fwd}}$ is now socially optimal, and so the filter is better off forwarding all content regardless of its signal. The next proposition shows that there exists parameter regimes where the socially optimal equilibrium is one in which the MVoT is 0. To this end, we characterize this regime precisely in terms of the model fundamentals.

Proposition 4.3. *Let* $D_{\mathrm{KL}}(p \,\|\, q)$ *be the Kullback-Leibler divergence between Bernoulli distributions with success probabilities* p *and* q. *Then* $V(\sigma_{\mathtt{dif}}) \geq V(\sigma_{\mathtt{fwd}})$ *if and only if one of the following conditions hold:*

(a) $q \geq q_{\mathrm{H}}$,

(b) $q_{\mathrm{L}} < q_{\mathtt{dif}} < q < q_{\mathrm{H}}$ *and*

$$\triangle U > \lambda\left[D_{\mathrm{KL}}(q \,\|\, q_{\mathrm{L}}) - \beta \cdot D_{\mathrm{KL}}(q_{\mathtt{dif}} \,\|\, q_{\mathrm{L}})\right],$$

(c) $q_{\mathtt{dif}} \leq q_{\mathrm{L}} < q < q_{\mathrm{H}}$ *and* $\triangle U > \lambda \cdot D_{\mathrm{KL}}(q \,\|\, q_{\mathrm{L}})$,

(d) $q_{\mathtt{dif}} < q \leq q_{\mathrm{L}} < q_{\mathrm{H}}$ *and* $\triangle U > 0$,

where we used the following shorthand

$$\triangle U := \mathbb{E}\left[u\left(s_{\mathtt{dif}}(\Psi_{\mathtt{f}}), X\right)\right] - \mathbb{E}\left[u\left(s_{\mathtt{fwd}}(\Psi_{\mathtt{f}}), X\right)\right],$$

for the expected increase in action payoffs with an always-accepting consumer when the filter's strategy switches from $s_{\mathtt{fwd}}$ *to* $s_{\mathtt{dif}}$*; and* $\beta := \Pr\left[a_{\mathtt{f}} = 1 \mid s_{\mathtt{f}} = s_{\mathtt{dif}}\right]$ *is the* ex ante *probability that a differentiating filter forwards the content.*[7]

It is straighforward to show that both barriers are cleared once the filter quality is high enough:

Corollary 4.4. *There exist thresholds* $\pi_0' < 1$ *and* $\pi_1' > 0$ *such that* $q_{\mathtt{dif}} < q_{\mathrm{H}}$ *and* $V(\sigma_{\mathtt{dif}}) > V(\sigma_{\mathtt{fwd}})$ *for any* $\pi_0 > \pi_0'$ *and* $\pi_1 < \pi_1'$.

Finally, Proposition 4.3 implies that the non-linear VoTC regime from Theorem 4.2 is feasible. Indeed, this regime corresponds to case (b) of the proposition.

5 Semi-Aligned Utilities ($V_{\mathtt{f}} = u$)

This section considers *semi-aligned utilities*: $V_{\mathtt{f}}(\sigma) = u(\sigma)$. Our results concern Pareto-efficiency. We show that all equilibria may be Pareto-inefficient (in stark contrast with the aligned utilities), but this inefficiency vanishes if the filter quality is sufficiently high. Put differently, improving the filter has an important side benefit of guaranteeing Pareto-efficient equilibria.

For clarity, we focus on the regime where $q_{\mathrm{L}} < q_{\mathtt{dif}} < q_H < q$ (Similar results holds for other regimes, but have a higher notation burden). In this regime, the inefficiency arises when one measure of filter quality is sufficiently low. Specifically, we summarize filter quality as one number that is strictly pointwise-increasing in π_0 and $-\pi_1$,

$$\mathcal{Q}(\pi_0, \pi_1) = \frac{\pi_0}{\pi_1} \frac{1-\pi_1}{1-\pi_0}. \tag{11}$$

We compare (11) to a threshold driven by cost parameters:

$$\Lambda = \frac{b + c_1}{c_2} \frac{1-q_{\mathrm{L}}}{q_{\mathrm{L}}}. \tag{12}$$

Theorem 5.1 (inefficiency)**.** *Assume* $q_{\mathrm{L}} < q_{\mathtt{dif}} < q_{\mathrm{H}} < q$. *If furthermore* $\mathcal{Q}(\pi_0, \pi_1) < \Lambda$, *then profile* $\sigma_{\mathtt{dif}}$ *strictly Pareto-dominates any equilibrium but is not an equilibrium itself. In particular, any equilibrium is Pareto-inefficient.*

The key insight behind Theorem 5.1 is that a low quality filter cannot commit to $\sigma_{\mathtt{dif}}$ because it has an incentive to trick the consumer into incurring information costs that are higher than optimal for the consumer. Under $\sigma_{\mathtt{dif}}$, the filter incurs a cost of $(1-q)\pi_1 c_1$ for blocking clean content but incurs some benefit from the consumer's content inspection. If the filter could *convince* the consumer it would choose $s_{\mathtt{dif}}$, the filter would be better off by instead forwarding all content, not incurring the cost of $(1-q)\pi_1 c_1$ and still enjoying the benefit of the consumer inspecting the content. Knowing this, the filter can *not* convince the consumer that it would play $s_{\mathtt{dif}}$ and thus $\sigma_{\mathtt{dif}}$ is not an equilibrium.

To escape this inefficiency, one can improve the filter, ensuring that $\mathcal{Q} > \Lambda$. The VoTC would be strictly positive.

[7] Unwrapping, $\triangle U = \pi_0 q c_2 - \pi_1(1-q)(b+c_1)$ and $\beta = (1-\pi_0)q + (1-\pi_1)(1-q)$.

Theorem 5.2 (escaping the inefficiency)**.** *Assume* $q_{\mathtt{H}} < q$ *and suppose* $\pi_0' \leq \pi_0$*,* $\pi_1' > \pi_1$ *and* $\mathcal{Q}(\pi_0, \pi_1) > \Lambda > \mathcal{Q}(\pi_0', \pi_1')$*. Then:*

(a) *The differentiating profile* $\sigma_{\mathtt{dif}}$ *is a Pareto-efficient equilibrium, and it Pareto-dominates any other equilibrium.*

(b) *The VoTC by switching from any equilibrium with filter quality* (π_0', π_1') *to* $\sigma_{\mathtt{dif}}$ *with filter quality* (π_0, π_1) *is strictly positive.*

The intuition behind Theorem 5.2 is as follows. As soon as the filter is of sufficiently high quality, $\sigma_{\mathtt{dif}}$ becomes an equilibrium, is Pareto efficient and furthermore, is the equilibrium preferred by both players. Behaviorally, when the filter is of sufficiently high quality, it is credible for the filter to use strategy $s_{\mathtt{dif}}$. Content with a strong bad signal is so likely to be malicious that the filter prefers not to forward it. As a result, the filter can credibly commit to playing $\sigma_{\mathtt{dif}}$. This characterization again highlights the non-linear nature of filter improvements and the importance of the filter meeting a baseline level of quality. However, unlike in the aligned section where it was the *consumer* that would abandon platforms with low quality filters, with semi-aligned incentives it is the *filter*'s incentive to forward too much content that leads to inefficient outcomes with low quality filters.

Consider the regime of Theorem 5.2(a), *i.e.,* $q_{\mathtt{H}} < q$ and $\mathcal{Q}(\pi_0, \pi_1) > \Lambda$. Once the filter and consumer enter this regime, further improving the filter would keep them in that regime. Theorem 5.3 shows that such improvements would benefit both players, and characterizes the resulting VoTC.

Theorem 5.3. *Assume* $q_{\mathtt{L}} < q_{\mathtt{dif}} < q_{\mathtt{H}} < q_{\mathtt{H}}$ *and* $\mathcal{Q}(\pi_0, \pi_1) > \Lambda$*. Under equilibrium* $\sigma_{\mathtt{dif}}$*, the MVoT is positive for both players, constant for the filter, and non-constant for the consumer.*

6 Endogenous Attacker

In this section, we extend our model to include the *attacker*: a third strategic player who is responsible for choosing the rate of malicious content, ρ_0. We focus on *aligned utilities*, to better isolate the novelty brought by endogenizing the attacker. We find two surprising consequences: the consumer does not incur information costs in equilibrium, and that improving the filter can make both the filter and consumer worse off.

Modeling choices and notation. We restrict the attacker to pure strategies, *i.e.,* to choose its rate ρ_0 deterministically for the entire batch. One interpretation is that the attacker is not sophisticated enough to implement mixed strategies in this context.[8]

As in the original model, all three players choose a strategy to use on the entire batch. The attacker and filter move first and simultaneously. The consumer observes the attacker's choice of ρ_0 but not the filter's chosen strategy, and moves next. Therefore, for a fixed and known value of ρ_0, the game reduces to the content-filtering game defined in Section 2. Importantly, as per Remark 2.1, our results carry over to the variant where the consumer observes the strategies of both the attacker and the filter. Furthermore, the results carry over to the case where the filter also observes ρ_0.

The attacker's expected utility, denoted $V_{\mathtt{a}}$, is the expected number of malicious pieces of content that are accepted by the consumer.[9] Fixing the strategies of all players and letting Y be the number of malicious messages in a batch, we have

$$V_{\mathtt{a}} = \mathbb{E}[Y] \ \Pr\left[a_{\mathtt{f}} = a_{\mathtt{c}} = 1 \mid X = 0 \right], \qquad (13)$$

where $\mathbb{E}[Y] = \rho_0 = q/(1-q)$.

Denote strategy profiles as $(\rho_0, \boldsymbol{\sigma})$. We denote the players' utilities by $V_{\mathtt{a}}$ and $V = V_{\mathtt{f}} = V_{\mathtt{c}}$. In general, we expand any quantities that take as an input $\boldsymbol{\sigma}$ to also take as an input ρ_0. For example, we write $V_{\mathtt{a}} = V_{\mathtt{a}}(\rho_0, \boldsymbol{\sigma})$ and $V = V(\rho_0, \boldsymbol{\sigma})$. Likewise, we write $q(\boldsymbol{\sigma}_{\mathtt{f}}) = q(\rho_0, \boldsymbol{\sigma}_{\mathtt{f}})$ in Eq. (7).

Note that the rate ρ_0 only enters the model through its impact on $q := \Pr[X = 0] = \rho_0/(\rho_0 + 1)$, which can take an arbitrary value in the interval $(0, 1)$. Therefore, one could equivalently reparameterize the model so that the attacker sets $q \in (0, 1)$ directly.

Equilibrium information costs. Our first result is that the consumer never incurs information costs in an equilibrium.

Theorem 6.1. $\mathtt{C}(\rho_0^*, \boldsymbol{\sigma}^*) = 0$ *for any equilibrium* $(\rho_0^*, \boldsymbol{\sigma}^*)$.

The key driver of Theorem 6.1 is that for a fixed filter's strategy, the attacker's expected payoff under the consumer's best response is *decreasing* in ρ_0 when $q(\rho_0, \boldsymbol{\sigma}_{\mathtt{f}}) \in (q_{\mathtt{L}}, q_{\mathtt{H}})$. Behaviorally, as the relative proportion of malicious content rises, a combination of the consumer's increased information costs and required certainty to accept content *reduces* the total amount of malicious content that is ultimately accepted (of course, this comes at a higher cost due to ignoring clean content). On the other hand, for $q(\rho_0, \boldsymbol{\sigma}_{\mathtt{f}}) < q_{\mathtt{L}}$, the attacker's payoff is increasing in ρ_0 since the consumer's best response is to accept all content. As a result, for a fixed filter strategy, the attacker's optimal strategy is to set ρ_0 such that $q(\rho_0, \boldsymbol{\sigma}_{\mathtt{f}}) = q_{\mathtt{L}}$. In this sense, the consumer's attention serves as a deterrent to attack: the amount of malicious content will not exceed the amount such that the consumer incurs information costs in deciding whether content is legitimate.

Negative VoTC. We find that improving the filter can *reduce* the equilibrium utility of the filter and the consumer.

As in Section 4, we focus on equilibria $(\rho_0^*, \boldsymbol{\sigma}^*)$ that maximizes the utility for the filter and the consumer, *i.e.,* satisfy

$$V(\rho_0^*, \boldsymbol{\sigma}^*) \geq V(\rho_0, \boldsymbol{\sigma}) \text{ for any equilibrium } (\rho_0, \boldsymbol{\sigma}).$$

and label this equilibrium payoff V^*. We are interested in VoTC in terms of V^*.

Our negative VoTC result can now be succinctly formulated using the ratio $\frac{\pi_0}{\pi_1}$ and the threshold Λ from Eq. (12).

[8] A mixed strategy chooses ρ_0 at random once and keeps it fixed.

[9] We do not impose *production costs* on the attacker for generating malicious content. These costs are often small in practice: *e.g.,* a generative AI model can produce many deep-fakes, an inexpensive phish-kit can generate many fake emails (Volkov 2020). Our results generalize to allow for small but positive production costs.

Theorem 6.2 (Negative VoTC)**.** *Suppose $\pi_0/\pi_1 < \Lambda$. Then sufficiently improving both π_0 and π_1 strictly decreases the equilibrium utility V^*. More formally: there exist thresholds $\hat{\pi}_0 \in (\pi_0, 1)$ and $\hat{\pi}_1 \in (0, \pi_1)$ such that for any $\pi'_0 \in (\hat{\pi}_0, 1)$ and $\pi'_1 \in (0, \hat{\pi}_1)$ improving the filter quality to (π'_0, π'_1) strictly decreases V^*.*

What drives this result is that improvements in filter technology can be completely crowded out by an increase in the attack propensity. One key reason is that the socially optimal equilibrium switches from $\sigma_{\texttt{fwd}}$ to $\sigma_{\texttt{dif}}$ as the filter technology improves. Specifically, when the filter is poor quality, the socially optimal equilibrium is $\sigma_{\texttt{fwd}}$. Then, the attacker sets ρ_0 such that $q(\rho_0, \sigma_{\texttt{fwd}}) = q_{\texttt{L}}$ and the consumer accepts all content. However, for a high quality filter, the socially optimal equilibrium is $\sigma_{\texttt{dif}}$. In that case, the attacker sets ρ_0 such that $q(\rho_0, \sigma_{\texttt{dif}}) = q_{\texttt{L}}$. Consequently, the expected fraction of malicious content that reaches the consumer is the same in both equilibria and therefore, the equilibrium expected utility for the filter and consumer *conditional on content reaching the consumer* is the same. However, since under $\sigma_{\texttt{dif}}$ the filter blocks some clean content, the filter's and consumer's expected utility under the $\sigma_{\texttt{dif}}$ is strictly lower than the expected utility under $\sigma_{\texttt{fwd}}$. Although under $\sigma_{\texttt{dif}}$ the filter blocks some malicious content, that benefit is not justified by the increase in attack intensity.

Another key feature driving this result is the filter's inability to commit to $s_{\texttt{fwd}}$. If the filter were able to commit to $s_{\texttt{fwd}}$, then equilibrium expected utilities would not depend on π_0 and π_1 and thus payoffs would not change as the filter improved in quality. However, because the filter and attacker act simultaneously, once the filter is of sufficiently high quality, the filter has an incentive to switch to $s_{\texttt{dif}}$. However, under $s_{\texttt{dif}}$, the attacker increases ρ_0, ultimately lowering equilibrium expected payoffs for the filter and consumer.

7 Conclusions and Open Questions

We develop a model of strategic interactions between a content filter and inattentive content consumers; such interactions are a common feature in many applications. Our equilibrium analysis undermines the common notions that improving filter quality is unambiguously beneficial and that the improvements are necessarily linear in the natural parameters (such as the true/false positive rates). We conclude that consumers' strategic inattention is essential for the analysis of content filtering.

The main policy implication is that content filtering does not reduce to a classification problem in machine learning. In addition to rote improvements in filter quality, one should consider interventions to reduce consumers' information costs and increase vigilance.[10] Our analysis illuminates non-obvious positive consequences of these interventions that arise due to strategic interactions: *e.g.,* increasing the marginal benefits of improvements in filter quality, or disincentivizing the attacker from inserting more malicious content. Detailing whether and which interventions are desirable remains an intriguing open question.

We focus on a homogeneous and stationary world in which the homogeneous players' strategies are non-adaptive and fixed throughout. Effectively, we consider a "single-round" game that concerns a single piece of content. This stationary world is, of course, an idealization of a dynamic world in which the players continuously adapt to one another. Such dynamic worlds are notoriously difficult to analyze, and are not well-understood even in simple scenarios.[11] Focusing on equilibria of a "single-round" game is a common route towards tractability. Nevertheless, adding dynamics with heterogeneous consumers is a viable extension.

A key simplification in our model is that all legitimacy-related quantities are *binary*: the legitimacy itself, the filter's signal and action and the consumer's signal and action. Indeed, the filter's and the consumer's signal could be fractional, reflecting the likelihood of the content piece being malicious. Filter's actions could also include, *e.g.,* putting the content piece into a spam folder or attaching a warning. Furthermore, the content piece itself may sometimes be a mix of genuine and malicious, *e.g.,* a genuine social media post may be contaminated by propaganda. Accordingly, a consumer might choose an 'intermediate' action, *e.g.,* accept the content piece with some reservations. Relaxing these binary choices could potentially lead to more refined conclusions, but might also lose the appealing simplicity and tractability of the "binary" model.

Our model of information costs, while suitable (and standard) for idealized models, could potentially be refined to reflect more realistic scenarios of information discovery. First, the *process* of information discovery could be modeled more explicitly, perhaps via an analogy to machine learning algorithms for similar problems. Second, the *information sources* available to a human user may differ from the one readily available to the filter. For example, a human receiving an email might intuitively pick up on a suspicious tone or an unusual visual layout, whereas a spam filter would be restricted to specific pre-trained characteristics of the email. Moreover, a human user might do a quick web search to resolve a suspicion (*e.g.,* of spam, phishing, or misinformation), or even ask a friend, whereas a spam/content filter might consult its internal database. On the other hand, such refinements might be application-specific and/or involve some unobvious modeling choices.

Another approach towards modeling information costs is to handle a large, abstract *class* thereof, without attempting to micro-found any particular function shape in this class. In the full version, we obtain an initial result in this direction, generalizing the conclusions in Section 4 to arbitrary information costs under some generic conditions.

References

Acemoglu, D.; Ozdaglar, A.; and Siderius, J. 2021. A model of online misinformation. Technical report, National Bureau of Economic Research.

[10] Such interventions are not uncommon in practice. Mandatory corporate trainings are now wide-spread. Some IT departments even implement "secret exercises", *e.g.,* send out phishing emails to all employees and reprimand those who fall for these emails.

[11] They are studied in (decentralized) multi-agent learning, *e.g.,* Ch. 9.5 in (Slivkins 2019) for introductory background.

Acharya, S.; and Wee, S. L. 2020. Rational inattention in hiring decisions. *American Economic Journal: Macroeconomics*, 12(1): 1–40.

Aldwairi, M.; and Alwahedi, A. 2018. Detecting fake news in social media networks. *Procedia Computer Science*, 141: 215–222.

Bagher, R. C.; Hassanpour, H.; and Mashayekhi, H. 2017. User trends modeling for a content-based recommender system. *Expert Systems with Applications*, 87: 209–219.

Ben-Porat, O.; and Tennenholtz, M. 2018. A game-theoretic approach to recommendation systems with strategic content providers. *arXiv preprint arXiv:1806.00955*.

Benenson, Z.; Gassmann, F.; and Landwirth, R. 2017. Unpacking spear phishing susceptibility. In *International Conference on Financial Cryptography and Data Security*, 610–627. Springer.

Bergemann, D.; and Ozmen, D. 2006. Optimal pricing with recommender systems. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, 43–51.

Bertoli, S.; Moraga, J. F.-H.; and Guichard, L. 2020. Rational inattention and migration decisions. *Journal of International Economics*, 126: 103364.

Bhowmick, A.; and Hazarika, S. M. 2018. E-mail spam filtering: a review of techniques and trends. *Advances in Electronics, Communication and Computing*, 583–590.

Blackwell, D. 1951. Comparison of experiments. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, volume 2, 93–103. University of California Press.

Blythe, M.; Petrie, H.; and Clark, J. A. 2011. F for fake: four studies on how we fall for phish. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 3469–3478.

Candogan, O.; and Drakopoulos, K. 2020. Optimal signaling of content accuracy: Engagement vs. misinformation. *Operations Research*, 68(2): 497–515.

Caplin, A. 2016. Measuring and modeling attention. *Annual Review of Economics*, 8: 379–403.

Caplin, A.; Dean, M.; and Leahy, J. 2022. Rationally inattentive behavior: Characterizing and generalizing Shannon entropy. *Journal of Political Economy*, 130(6): 1676–1715.

Chae, M.; Alsadoon, A.; Prasad, P.; and Sreedharan, S. 2017. Spam filtering email classification (SFECM) using gain and graph mining algorithm. In *2017 2nd International Conference on Anti-Cyber Crimes (ICACC)*, 217–222. IEEE.

Dasgupta, K.; and Mondria, J. 2018. Inattentive importers. *Journal of International Economics*, 112: 150–165.

D'Onfro, J. 2018. Google now lists fake news and 'objectionable content' as risks to its business. https://www.cnbc.com/2018/02/06/alphabet-adds-third-party-content-misleading-information-as-risks.html. Accessed: 2023-01-20.

Gabaix, X. 2019. Behavioral inattention. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 2, 261–343. Elsevier.

Gangavarapu, T.; Jaidhar, C.; and Chanduka, B. 2020. Applicability of machine learning in spam and phishing email filtering: review and approaches. *Artificial Intelligence Review*, 1–63.

Hendricks, V. F.; and Vestergaard, M. 2019. *Reality lost: Markets of attention, misinformation and manipulation*. Springer Nature.

Jiang, G.; Fosgerau, M.; and Lo, H. K. 2020. Route choice, travel time variability, and rational inattention. *Transportation Research Part B: Methodological*, 132: 188–207.

Joseph, A. D.; Nelson, B.; Rubinstein, B. I. P.; and Tygar, J. D. 2019. *Adversarial Machine Learning*. Cambridge University Press.

Kamenica, E. 2019. Bayesian persuasion and information design. *Annual Review of Economics*, 11: 249–272.

Kumar, K. K.; and Geethakumari, G. 2014. Detecting misinformation in online social networks using cognitive psychology. *Human-centric Computing and Information Sciences*, 4(1): 1–22.

Laszka, A.; Lou, J.; and Vorobeychik, Y. 2016. Multi-defender strategic filtering against spear-phishing attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Lu, J.; and Niu, R. 2015. A state estimation and malicious attack game in multi-sensor dynamic systems. In *2015 18th International Conference on Information Fusion (Fusion)*, 932–936. IEEE.

Maćkowiak, B.; and Wiederholt, M. 2015. Business cycle dynamics under rational inattention. *The Review of Economic Studies*, 82(4): 1502–1532.

Martin, D. 2017. Strategic pricing with rational inattention to quality. *Games and Economic Behavior*, 104: 131–145.

Mas-Colell, A.; Whinston, M. D.; Green, J. R.; et al. 1995. *Microeconomic theory*, volume 1. Oxford university press New York.

Matějka, F.; and McKay, A. 2015. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1): 272–98.

Matsui, A.; and Matsuyama, K. 1995. An approach to equilibrium selection. *Journal of Economic Theory*, 65(2): 415–434.

Matyskova, L.; and Montes, A. 2023. Bayesian persuasion with costly information acquisition. *Journal of Economic Theory*, 105678.

Milgrom, P.; and Weber, R. J. 1982. The value of information in a sealed-bid auction. *Journal of Mathematical Economics*, 10(1): 105–114.

Papanastasiou, Y. 2020. Fake news propagation and detection: A sequential model. *Management Science*, 66(5): 1826–1846.

Pennycook, G.; and Rand, D. G. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188: 39–50.

Pomatto, L.; Strack, P.; and Tamuz, O. 2023. The cost of information: The case of constant marginal costs. *American Economic Review*, 113(5): 1360–1393.

Ravid, D. 2020. Ultimatum bargaining with rational inattention. *American Economic Review*, 110(9): 2948–63.

Roozenbeek, J.; Schneider, C. R.; Dryhurst, S.; Kerr, J.; Freeman, A. L.; Recchia, G.; Van Der Bles, A. M.; and Van Der Linden, S. 2020. Susceptibility to misinformation about COVID-19 around the world. *Royal Society open science*, 7(10): 201199.

Sangster, A. 2015. You cannot judge a book by its cover: The problems with journal rankings. *Accounting Education*, 24(3): 175–186.

Sims, C. A. 2003. Implications of rational inattention. *Journal of monetary Economics*, 50(3): 665–690.

Slivkins, A. 2019. Introduction to Multi-Armed Bandits. *Foundations and Trends® in Machine Learning*, 12(1-2): 1–286. Published with *Now Publishers* (Boston, MA, USA). Also available at https://arxiv.org/abs/1904.07272. Latest online revision: Jan 2022.

Sterrett, D.; Malato, D.; Benz, J.; Kantor, L.; Tompson, T.; Rosenstiel, T.; Sonderman, J.; and Loker, K. 2019. Who shared it?: Deciding what news to trust on social media. *Digital journalism*, 7(6): 783–801.

Vives, X. 1984. Duopoly information equilibrium: Cournot and Bertrand. *Journal of economic theory*, 34(1): 71–94.

Volkov, D. 2020. How much is the phish? Underground market of phishing kits is booming — Group-IB. https://www.group-ib.com/media/how-much-is-the-phish/. Accessed: 2022-02-10.

Vorobeychik, Y.; and Kantarcioglu, M. 2018. Adversarial machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(3): 1–169.

Wei, Y. Z.; Moreau, L.; and Jennings, N. R. 2003. Recommender systems: A market-based design. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, 600–607.

Zhong, W. 2022. Optimal dynamic information acquisition. *Econometrica*, 90(4): 1537–1582.