The Implied Social Contract in Social Media

Sarah H. Cen
shcen@mit.edu
MIT
Cambridge, MA, USA

Aleksander Mądry
madry@mit.edu
MIT
Cambridge, MA, USA

Devavrat Shah
madry@mit.edu
MIT
Cambridge, MA, USA

ABSTRACT

By moderating the information that users see, social media platforms have an outsized impact on users' beliefs and behaviors. Consequences of this influence, such as the rise of misinformation and declining mental health of teenage users, have prompted calls for regulation. However, many of the proposals to directly regulate content are impeded by existing laws and protections.

In this work, we take a different approach. In particular, we study how the interactions between users and platforms give rise to an *implied social contract*, then discuss how the terms of this contract can be surfaced using implied contract theory. Under this contract, the role of the state is to ensure that both parties—especially, the platform—honor its terms.

To illustrate the implications of this framework, we specialize and define the social contract in the context of *algorithmic filtering* (i.e., how platforms curate the content that appears on users' feeds). We then present an auditing procedure that checks whether a platform honors its social contract on algorithmic filtering and, notably, requires no more than black-box access to the filtering algorithm. We provide theoretical guarantees on the audit's performance and simulations demonstrating its impact on the users' content.

CCS CONCEPTS

• **Social and professional topics → Governmental regulations**; • **Information systems → World Wide Web**.

KEYWORDS

social media, social contract, implied contract, algorithmic filtering, auditing, regulation, governance, decision robustness.

1 INTRODUCTION

Social media has become our one-stop-shop for online content. These days, we turn to social media not only to keep up with friends and family, but also to see the latest in news, entertainment, sports, fashion, and more.

Social media platforms play a crucial role in this information exchange. In particular, they *curate* the content that users see using a process known as *algorithmic filtering*: for each user, the platform gathers data—such as the articles the user has clicked on and the information the user volunteers on their profile—then uses this data to algorithmically generate a personalized feed.

By moderating what users see, filtering algorithms have an outsized impact on user and their beliefs. Videos can shape a user's interests and hobbies [5, 30]. Advertisements can affect what a user chooses to buy and even where they choose to live [3, 16]. Articles can inform a user's opinions on recent political issues [26, 31]. In this way, how platforms curate content can have significant downstream effects. Indeed, there is evidence that Facebook's algorithms amplified political misinformation during the 2016 US elections [2] and that Instagram promoted content that had measurable, negative effects on the mental health of its teenage users [33].

As the popularity of social media continues to grow, so does the reach and power of filtering algorithms, giving rise to the question:

When and how should algorithmic filtering be regulated?

Answering this question is no easy task. In ths US, regulating social media is challenging due to *Section 230 of the Communications Decency Act*, the strength of *free speech protections*, and *trade secret law*. Under Section 230, regulators cannot hold platforms legally responsible for the information contained in a piece of content as long as the platform does not *create* the content [9]. Beyond Section 230, platforms are protected under the First Amendment. Because the right to free speech in the US extends to corporations, how platforms choose to promote and demote content is considered the "speech" of the platform and is therefore protected [20]. Even if lawmakers agree on what content should and should not be amplified, regulations that *directly* limit harmful content are unlikely to hold up in court due to their tendency to infringe on the free speech of users and content creators [1, 20]. On top of all this, filtering algorithms are hidden behind trade secret law, making it difficult to investigate how they work and when they are (or are not) to blame.

A different perspective. These factors make it difficult to directly regulate the content or content policies of social media platforms (e.g., to define and remove misinformation).

As an alternative to content-focused regulations, we propose to regulate the *implied social contract* between users and their social media platforms. Specifically,

- We explain how an implied social contract arises in the context of social media. We then explore how the existence of such a contract requires that the state step in when platforms fail to uphold their end of the agreement.
- We examine the implied social contract in the context of algorithmic filtering. This analysis gives rise to a requirement on a platform's filtering algorithm that, intuitively, gives users more agency over the content that they see.
- We provide a concrete auditing procedure, which ensures that, with high probability, a platform's filtering algorithm upholds its social contract with users. We conclude by simulating the impact of the audit.

The Web Conference, Under review, 2022

Crucially, regulating the implied social contract plays nicely with existing laws—such as Section 230 and free speech protections. At its core, it ensures that platforms are held *accountable to their users*. Additionally, the audit we propose does not require more than black-box access to the filtering algorithm (and therefore respects trade secret law), and it can be run on synthetic user data (and therefore respects user privacy). We note that, although this work is motivated by the regulatory landscape of the US, the framework and procedure may be applicable in other contexts.

Our main contributions are summarized in greater detail below.

The implied social contract in social media. We investigate how an implied social contract arises between users and their social media platforms. Specifically, a social contract arises when individuals (e.g., users) collectively surrender a part of their autonomy to an authority (e.g., a platform) in exchange for services that improve the individuals' well being in a way that the individuals could not achieve on their own [8, 13, 17, 28]. Although some social contracts are written expressly (e.g., the US Constitution), others are implied. Importantly, implied contracts are also legally binding [15, 29].

In the context of social media, users join a platform in exchange for social media services, including access to their social network and the latest content. In the process, users give up their autonomy in multiple ways (e.g., agency over their personal data and the ability to directly control what they see on their feeds), thus creating a social contract. If one party fails to uphold their end of a social contract, the other is, in theory, free to terminate the relationship. However, while users uphold their end of the social contract through their continued participation, what binds platforms to users is often unwritten and unenforced because users cannot easily leave platforms (e.g., due to the lack of competition). As such, the role of the state would be to correct this market failure by requiring that platforms uphold their contract with users.

The social contract on algorithmic filtering. We then examine the implications of an implied social contract on algorithmic filtering. In particular, we translate the social contract into an auditable requirement on the platform's filtering algorithm. To do so, we use the notion of a *baseline feed*—a feed generated by drawing content "naturally" from a user's social network as seen in [11]—to ground the contract. Upholding its contract with a user would *not* require the platform to give the user their baseline feed, as such a requirement would remove the platform's ability to curate content to the detriment of the platform—who would lose their main source of revenue—and the user—who would no longer receive the algorithmic curation they enjoy. Rather, it would require that a user's filtered feed is sufficiently similar to the baseline feed.

One way to view this requirement is through the lens of "consent", which appears in both social contract theory and the social media literature [6, 14, 24]. In particular, the baseline feed represents the content to which the user has consented (e.g., by friending or following others). Although what constitutes consent is debated, we adopt the baseline feed as our reference, noting that our proposal is easily adaptable to one's preferred notion of "consent".

A concrete auditing procedure. We adapt a recent method proposed by Cen and Shah [11] in order to audit the platform. The audit checks whether the platform upholds the social contract on algorithmic filtering by monitoring the algorithm's behavior. We use this method because it offers several benefits; in particular, it needs only black-box access to the filtering algorithm and does not require real user data. For a given set of inputs, the auditor first runs the black-box filtering algorithm to obtain the filtered feed. The audit then verifies whether the filtered feed satisfies a property known as *decision robustness*—a measure of the similarity between two feeds in a way that is designed specifically for the social media context [11]—with respect to the baseline feed.

Implications of regulating the social contract. We then study how enforcing the social contract on algorithmic filtering affects the users' content. In particular, although the audit prevents the platform from filtering any possible feed, the platform still has a reasonable amount of flexibility in how it filters. Through an example, we show that the platform can achieve good performance (e.g., high revenue) while upholding the social contract. In this same example, platforms are also incentivized to ensure that the filtered content is sufficiently diverse. As such, regulating of the social contract (in this setting) results in feeds that (i) are "close" to users' baseline feeds with small deviations and (ii) contain a little more content diversity than the users' baseline feeds.

2 THE IMPLIED SOCIAL CONTRACT

In this section, we review social and implied contracts, then unpack the (implied) social contract between users and social media platforms. To better understand its implications, we specialize this analysis to algorithmic filtering—how platforms curate user content—discussing how upholding its social contract with users would require the platform to give users more agency over the content that they see. We conclude by discussing the benefits and limitations of regulating the social contract in social media.

2.1 Contract theory

Under *social contract theory*, individuals consent to collectively surrender a part of their autonomy to an authority. In exchange, the authority provides services that improve the collective's well being in a way that the individuals could not have achieved on their own [8, 13, 17, 28]. Social contract theory relies on the notion of *consent* in that the authority's power is derived from the individuals it oversees. The authority is therefore only legitimate if it upholds its end of the agreement.

The social contract has historically been used to explain the role of the state. The reasoning goes: citizens give up rights and freedoms that they would have had under the "State of Nature" in return for common laws that govern all citizens and a state that enforces these laws [19]. The "Veil of Ignorance" has been used to envision what laws a state should adopt for the good of the collective [25]. It poses the following thought experiment: individuals should design laws as if they did not yet know their circumstances in life (e.g., whether they would be born rich or poor, short or tall). Although the concept of social contracts was originally developed to explain the relationship between citizens and their state, it has been applied more broadly to contexts in which individuals surrender (part of) their autonomy to an authority.

Table 1: Key concepts

Term	Description	Section
Social contract	Agreement that arises when individuals collectively surrender a degree of their autonomy to an authority in exchange for services that improve the collective's well being.	2.1
Implied contract	Agreement between two parties that exchange goods & services that may be non-verbal or unwritten.	2.1
Algorithmic filtering	The process by which social media platforms curate the content (e.g., posts, articles, ads) users see.	2.3.
Baseline feed	Feed obtained by drawing content uniformly at random from a user's social network.	2.3 & 3.2
Decision robustness	Notion of robustness specialized to context of algorithmic filtering. Given two feeds (e.g., the baseline and filtered feeds), ensures that their downstream effects are effectively indistinguishable.	3.3
Audit	Checks if the platform upholds its social contract on algorithmic filtering by testing for decision robustness. Requires only black-box access to filtering algorithm.	3.4 & 4

In some cases, the social contract is explicit, e.g., expressed in a Constitution. In others, it is implied. Under *implied contract theory*, an implied contract may exist when two parties exchange goods or services even if the terms are non-verbal or unwritten [23]. Contracts rely on the notion of consent, and implied contracts arise when consent can be reasonably "inferred from the parties' conduct or from the circumstances surrounding their relationship" [12]. Implied contracts are, by definition, unwritten and typically reflect unspoken norms. For instance, if previous interactions between two parties suggest a norm (e.g., users willingly provide certain data under a platform's existing data privacy policy) and one party suddenly deviates from this norm while the other continues (e.g., the platform changes the privacy policy but users do not have sufficient opportunity or means to adjust their behavior), the former may be found in violation of the implied contract. Notably, implied contacts hold the same legal force as express ones [15, 29].

In both social and implied contract theory, the natural (i.e., non-legal) mechanism that holds one party responsible to the other is sometimes absent. In particular, when one party fails to uphold their end of the contract, the other is, in theory, free to terminate the relationship. In reality, this freedom does not exist when the cost of leaving the relationship is much higher for one party than for the other. For instance, when there is a lack of competition, one party is often able to get away with shirking their responsibilities.

2.2 The contract between a user and their social media platform

In many ways, the relationship between users and social media platforms—especially, in terms of the platform's obligations to users—is governed by unwritten norms. Although platforms typically have terms of service or community standards, these primarily govern what users—rather than platforms—can and cannot do. They lay out (i) the services a platform provides and (ii) how users must behave in order to receive these services. Should the user not agree to the platform's terms, their main means of expressing disagreement is leaving the platform. Facebook, for instance, states in its terms of service: if "[you] no longer want to be a part of the Facebook community, you can delete your account at any time"[1].

The problem is: in social media, leaving a platform is not easy. For one, there is little competition in the social media space (e.g., due to platforms buying out their competitors, as Facebook did with Instagram [7]). Users that wish to leave a platform may not do so due to the difficulty of finding another that provides similar services. Even if users are able to find one, there are also *network effects.* Namely, because each user invests time building their profile and social network on each platform, switching platforms is not trivial, especially if many of the people that the user would like to connect with are not on competing platforms. On top of all this, platforms generally reserve the right to update their terms of service—which would allow them to change what they require of users or relax what they do promise users—at any time, so long as users are given notice [27]. As such, *terms of service and community standards primarily bind users to platforms, not the other way around.*

Instead, what binds platforms to users is an implied social contract. Specifically, users join a platform in exchange for information goods and services. In the process, users surrender their autonomy in multiple ways, thus creating a social contract. For example, users give up their agency over much of their personal information, allowing platforms to collect, store, and sell their data. While users uphold their end of the social contract through their continued participation, what binds platforms to users is, for the most part, unwritten—it is *implied.* The question remains: *What are the terms of a contract that is unwritten?*

To this end, we can return to implied contract theory. Recall from implied contract theory that an implied contract exists when it can be "inferred from the parties' conduct or from circumstance surrounding the relationship" [12]. In other words, the terms of an implied contract are determined by what one would reasonably (or, at least, minimally) expect from each party.[2] In social media, users give up their autonomy over their personal data and their ability to choose their information sources in order to receive services from the platform. The implied social contract is therefore determined by how users would collectively and reasonably expect the platform to behave when providing these services. *The users' collective expectations of how the platform targets advertisements, treats their*

[1] https://www.facebook.com/legal/terms

[2] As an example, suppose that, one year, a homeowner mows her lawn and the lawn of her neighbor. In return, the neighbor gives the homeowner X dollars. Suppose that the next three years, this interaction repeats but, on the fifth year, the homeowner mows both lawns, and the neighbor refuses to pay. Due to the history of interaction, the neighbor is breaking an implied contract with the homeowner.

There are also implied contracts on products (e.g., that a product will work as intended) and in medical care (e.g., that a doctor will deliver medical care to the best of their abilities in return for payment of medical fees).

personal data, or monitors their chats—anything that is not expressly written—form the implied social contract.

Although the user can, in theory, terminate the contract if this baseline is not reasonably met, users face barriers when leaving a platform, as discussed above. Historically, these barriers have allowed platforms to operate with little accountability, assured that their user base will remain intact. The role of regulation would therefore be to correct this market failure by ensuring that platforms honor the implied social contract it has with users. In this work, we restrict our attention to the implied social contract as it pertains to algorithmic filtering.

2.3 Regulating algorithmic filtering

In this section, we unpack the implied social contract as it pertains to *algorithmic filtering*. Recall that algorithmic filtering is the mechanism that platforms use to select the content that appears on a user's feed. Filtering algorithms govern anything from the advertisements a user sees to the comments that automatically appear each post. By determining what information users receive and how they receive it, filtering algorithms wield a great deal of influence.

Users comply with algorithmic filtering because it provides a personalized social media experience and saves users having to search for relevant, interesting content themselves. In return for this service, platforms receive users' time, attention, and data. In other words, users give up their ability to directly control the information to which they are exposed in exchange for content curation, and the way that platforms honor this transfer of agency is, in many ways, shaped by unwritten norms. Algorithmic filtering is therefore governed by an implied social contract. Determining the terms of this contract comes down to determining what behavior users would *collectively and reasonably expect from the platform*.

To this end, we return to the Veil of Ignorance—a thought experiment coined by Rawls [25] and further developed in the 20th century [18, 32]—for inspiration. In his writings on the social contract between citizens and their state, Rawls proposed that individuals should collectively decide the laws that govern them by imagining that they are under a Veil of Ignorance—that they have yet to be born and do not know the circumstances into which they will be born [25]. Rawls argued that this mental exercise would yield the appropriate laws that a state should adopt. In other words, Rawls believed the Veil of Ignorance was the correct way to determine a reasonable *baseline* on which to build the social contract.

In the context of algorithmic filtering, we posit that the appropriate baseline—what users collectively and reasonably expect from their platforms—should be determined analogously. Imagine that users were ignorant of their social media interests as well as the platforms' incentives. In this world, what would such users wish to see on their feeds? We argue that users would expect to see a feed containing content similar to that for which the user has given consent. In particular, we define a user's *baseline feed* as a feed curated solely from the posts of a user's friends, pages that the user follows, and so on. The baseline content can be viewed as what *all* users agree that a platform should filter. In reality, platforms should, of course, be allowed to deviate from this baseline feed. After all, users enjoy the unexpected content that come with personalization, and platforms rely on their ability to inject content for revenue. The goal of regulation, then, would be to ensure that, at every item step, a user's filtered feed is "similar" to their baseline feed.

Although not all implied contracts necessitate regulation, the downstream effects from algorithmic filtering—from the spread of misinformation to discriminatory advertising—warrant state oversight. So, what does it mean for two feeds to be "similar" and how do auditors ensure that platforms uphold the implied social contract? We formalize these two questions in Sections 3-4.

2.4 Implications

In this section, we discuss the benefits and limitations of regulating the implied social contract between users and their platforms. Note that we do not recommend this approach as a be-all, end-all. Regulating social media will require multiple, complementary measures.

User-driven. Enforcing the implied social contract is a user-driven approach that increases the platform's accountability to users. Unlike regulations that seek to define and regulate "harmful" content (a task that is notoriously difficult both technologically and legally [20]), a user-driven approach seeks to empower users by giving them more agency over the content they are shown. A user-driven approach stands in contrast to regulations that draw *bright lines*, i.e., establish a global standard with which platforms must comply. There are two drawbacks to bright lines. For one, they are rigid, and this lack of flexibility has its downsides, as discussed below. For another, bright lines are drawn by policymakers. Instead of empowering users, bright lines simply transfer the authority to define "appropriate" content from the hands of social media giants into the hands of lawmakers.

Flexible. When designing regulations, policymakers must balance specificity against generality. Regulations that are too specific risk being applicable only in limited contexts or becoming obsolete. The implied social contract, on the other hand, is flexible and adaptable. Because the baseline feed is different for each user, it does not impose a global standard across all users. This approach is also flexible in that it is algorithm-agnostic. The implied social contract (as well as the audit we propose below) does not require the filtering algorithm to be of a specific form. On the other hand, regulations that work only for specific algorithms may not be applicable across platforms or time. In this way, the implied social contract is adaptable because it creates a user-driven baseline.

Compatible with existing laws. Enforcing the implied social contract does not conflict with Section 230 of the CDA, free speech protections, or trade secret law. In particular, regardless of whether or not platforms are considered the "publishers" of the content that they distribute or whether they remove content in "good faith" (as allowed under Section 230), the implied social contract holds platforms accountable to users by requiring the platforms provide users with the social media service that they seek in exchange for their time, attention, and data.

In addition, the implied social contract respects free speech protections. In general, free speech protections have made it difficult for lawmakers to regulate social media content directly. Even if lawmakers agree on a definition of "harmful" content (e.g., misinformation), criminalizing its distribution often results in the *chilling* *effect*: to play it safe, platforms over-police content and end up removing innocent posts (e.g., articles are mistakenly labeled misinformation). If a regulation leads to this outcome, courts will likely strike it down for infringing on the free speech of content creators whose innocent posts are removed. The implied social contract, on the other hand, does not define universally "good" or "bad" content. Rather, it places agency into the hands of users, allowing them to control what they see.

Lastly, as mentioned above, the implied social contract as well as the method we propose in this work requires only black-box access to a platform's filtering algorithm. As such, it allows platforms to keep their algorithms hidden as trade secrets.

Limitations. One limitation of user-driven regulations is that what users want (or indicate that they want) may not lead to desirable outcomes. For instance, a user may enjoy conspiracy theories, which would be reflected in their baseline feed. Enforcing the social contract would therefore not prevent platforms from showing conspiracy theories on that user's feed. As such, the social contract is not a be-all, end-all solution, and we recommend that it stand alongside other complementary regulations. For instance, if there is an outcome that lawmakers universally agree is undesirable, a user-driven approach alone is unlikely to prevent it. Even so, Keller supports user-driven regulations [20], stating, "If our behavior—our "revealed preferences," in economic parlance—says we want trashy but legal content, should laws prevent platforms from giving it to us?" Although allowing users to select their own content can lead to undesirable outcomes, restoring user agency is an important step towards correcting the power imbalance between platforms and users (cf. [20] for a more comprehensive discussion).

3 PROBLEM STATEMENT

In this section, we translate the implied social contract on algorithmic filtering into a formal requirement on a platform's filtering algorithm. Specifically, the contract requires that the filtering algorithm is *decision robust*, and the auditor's role is to test for decision robustness. Intuitively, decision robustness requires that a platform respect the agency of users by ensuring that its filtering algorithm does not inject content to manipulate the user in ways that the user has not consented to under the social contract.

3.1 Setup

Consider a system with three agents: a platform, a user, and an auditor. In order to curate feeds for its users, the **platform** employs a filtering algorithm $\mathcal{F} : \mathcal{X} \to \mathcal{Z}^m$. The filtering algorithm takes in inputs $\mathbf{x} \in \mathcal{X}$, which could encode anything from the history of user behaviors (e.g., what they have clicked on or watched in the past) to the set of available content.[3] The filtering algorithm then produces a feed $Z = \mathcal{F}(\mathbf{x})$ for said **user**, where $Z = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m)$ is a feed containing m pieces of content, and $\mathbf{z}_i \in \mathcal{Z}$ denotes the i-th element in the user's feed.

Assumption 1. *Let* $Z = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m) = \mathcal{F}(\mathbf{x})$ *for some* $\mathbf{x} \in X$. *Then,* $z_i \overset{i.i.d.}{\sim} p_{\mathbf{z}}(\cdot; \theta)$ *for* $\theta \in \Theta \subset \mathbb{R}^r$.

The **auditor**'s goal is to check whether the platform's filtering algorithm upholds the social contract it has with users given no more than black-box access to $\mathcal{F}$. By "black-box access", we mean that the platform can run $\mathcal{F}$ on a set of n inputs $X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ and observe the outputs $(Z_1, Z_2, \dots, Z_n)$, but the auditor does not need deeper access to the algorithm.[4]

The auditor would also like to respect the privacy of users. As a result, $\mathbf{x}_i$ need not correspond to real users. They could, instead, correspond to hypothetical users or user features after being run through a privacy-preserving pre-processing step.

3.2 Formalizing the social contract on algorithmic filtering

Recall from Section 2.3 that the social contract on algorithmic filtering implies that, for every input $\mathbf{x} \in \mathcal{X}$, the filtered feed must be sufficiently close to the baseline feed.

Formally, consider a (hypothetical) user with corresponding inputs $\mathbf{x} \in \mathcal{X}$. The platform constructs this user's feed from two pools of content. We refer to the first pool, $\mathcal{Z}_B(\mathbf{x})$, as the user's *baseline pool*: content to which the user has given consent, such as updates from the user's friends, articles on pages that the user subscribes to, posts by influencers that the user follows, and so on. We refer to the second pool, $\mathcal{Z}_I(\mathbf{x})$, as the user's *injected pool*: content that the platform may inject into the user's feed, such as advertisements, suggested posts, recommended products, and more.

As such, the platform constructs the user's filtered feed $Z = \mathcal{F}(\mathbf{x}) \subset \mathcal{Z}_B(\mathbf{x}) \cup \mathcal{Z}_I(\mathbf{x})$ by selecting some content from $\mathcal{Z}_B(\mathbf{x})$ as well as injecting content from $\mathcal{Z}_I(\mathbf{x})$.

Assumption 2. *We assume that the platform provides the auditor black-box access to a baseline filtering algorithm* $\mathcal{B} : \mathcal{X} \to \mathcal{Z}^m$, *where for given inputs (i.e., a hypothetical user)* $\mathbf{x}$, *the baseline algorithm draws the content uniformly at random from* $\mathcal{Z}_B(\mathbf{x})$, *i.e.,* $\mathbf{b}_i \overset{i.i.d.}{\sim} UAR(\mathcal{Z}_B(\mathbf{x}))$ *for* $(\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_m) = \mathcal{B}(\mathbf{x})$.

This assumption is not strong. Providing access to a baseline filtering algorithm adds little extra burden to a platform that is already providing access to their filtering algorithm $\mathcal{F}$. Under this setup, the auditor's goal from Section 2.3 can be expressed as follows:

> *Given a set of inputs* $X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, *the auditor requires that the filtered feed* $\mathcal{F}(\mathbf{x}_i)$ *is "close" to the baseline feed* $\mathcal{B}(\mathbf{x}_i)$ *for all* $i \in [n]$.

It remains to determine an appropriate notion of "closeness". For this, we turn to the concept of *decision robustness*, as proposed by Cen and Shah [11].

3.3 Decision robustness

Decision robustness is a concept developed by Cen and Shah [11] that specializes the notion of robustness to the context of algorithmic filtering. Intuitively, decision robustness is built on the idea that the "closeness" between two feeds should be measured with respect to their downstream impact on users, and it uses learning and decision

[3] The set of available content may change with time. As such, we could write $\mathcal{Z}$ as $\mathcal{Z}_t$. For simplicity, we use $\mathcal{Z}$, which can be easily generalized to the time-varying setting.

[4] The number of times n that the auditor can query $\mathcal{F}$ is usually limited by the fact that, if the auditor had infinite queries, it would be able to reconstruct $\mathcal{F}$, which would violate trade secret protections. We do not study this aspect of the audit in this work. We assume that n is given.

theory to formalize a definition of closeness. We slightly modify the definition of decision robustness from [11] as follows.

Assumption 3. *Once the auditor is given $Z = \mathcal{F}(\mathbf{x})$ and $Z_B = \mathcal{B}(\mathbf{x})$, the feeds are randomly shuffled such that the auditor does not know which feed is the baseline feed and which is the filtered feed. Let the shuffled (unidentified) feeds be denoted by Z' and Z'', i.e., $(Z', Z'') = (Z, Z_B)$ with probability $1/2$ and $(Z', Z'') = (Z_B, Z)$, otherwise. By Assumption 1, let the auditor believe that $\mathbf{z}'_i \overset{\text{i.i.d.}}{\sim} p_{\mathbf{z}}(\cdot; \theta')$ and $\mathbf{z}''_i \overset{\text{i.i.d.}}{\sim} p_{\mathbf{z}}(\cdot; \theta'')$ for all $i \in [m]$ and some $\theta', \theta'' \in \Theta \subset \mathbb{R}^r$.*

Definition 1 (Decision robustness for an input). *Suppose Assumptions 1-3 hold. For a given $\mathbf{x} \in \mathcal{X}$, let Z' and Z'' be as defined in Assumption 3. $\mathcal{F}$ is $(\mathbf{x}, \alpha, \Theta)$-decision robust if and only if the uniformly most powerful unbiased (UMPU) hypothesis test with significance α cannot reject the hypothesis $H_0 : \theta' = \theta''$.*

Definition 2 (Decision robustness for a set of inputs). *Consider a set of inputs $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$. $\mathcal{F}$ is (X, α, Θ)-decision robust if and only if $\mathcal{F}$ is $(\mathbf{x}, \alpha, \Theta)$-decision robust for all $\mathbf{x} \in X$.*

See the Appendix for a precise definition of the UMPU test. Intuitively, decision robustness implies that no reasonable auditor cannot confidently reject the hypothesis that both feeds are generated by the same $p_{\mathbf{z}}(\cdot; \theta') = p_{\mathbf{z}}(\cdot; \theta'')$ for $\theta', \theta'' \in \Theta$. In other words, $\mathcal{F}$ is decision robust if Z and Z_B are *effectively indistinguishable in terms of their downstream effects.*

Interpretation. Cen and Shah [11] establish a strong intuition for the notion of decision robustness that we summarize below. We refer the reader to their work for formal results.

Consider a (hypothetical) user who is shown the filtered feed $Z = \mathcal{F}(\mathbf{x})$, then faces a set of decision points Q, such as where the user decides to eat that evening, what shoes to buy, and even which presidential candidate to support. We consider any possible Q, as long as each decision point has a finite number of options.

Suppose that, in a parallel universe that is otherwise identical, the user is shown $Z_B = \mathcal{B}(\mathbf{x})$ instead. Since both universes are otherwise identical, the user faces the same decision points Q after seeing Z_B. Let the decisions that the user makes in the first universe be denoted by D and the latter by D_B.

Suppose we had access to D and D_B. (Recall that this thought experiment is hypothetical, and the audit we propose does *not* need observations of the user's decisions, which are unethical to obtain.) Suppose we are*not* told whether D represents the decisions from the filtered or the baseline universe, and similarly for D_B.

Then, decision robustness has the following intuition (cf. Section 2 in [11] for details).

> *If $\mathcal{F}$ is $(\mathbf{x}, \alpha, \Theta)$-decision robust, then (under mild conditions) the auditor cannot say with $(1-\alpha)$-confidence that D is from the filtering universe but D_B is not* for any Q.

Stated differently, decision robustness ensures that the downstream effect of the filtered feed on a user's (hypothetical) decisions—no matter what decision points Q they face—is similar to the downstream effects that the baseline feed would have had.

Notably, decision robustness does *not* require that the filtered feed $\mathcal{F}(\mathbf{x})$ is identical to the baseline feed $\mathcal{B}(\mathbf{x})$. Rather, it only requires that they are similar in their downstream effects. This flexibility is critical, as it allows the platform some freedom in how they filter—the source of their revenue—and it allows platforms to inject personalized content to the benefit of users.

3.4 Auditor's objective

Given the definition of decision robustness, checking whether the platform upholds its end of the social contract with users can be translated as follows:

$$\begin{gathered}\mathcal{F} \text{ upholds the social contract on algorithmic filtering} \\ \Longleftrightarrow \\ \mathcal{F} \text{ is } (X, \alpha, \Theta)\text{-decision robust}\end{gathered}$$

As such, given a set of inputs X, false positive rate α, and model family Θ, the auditor's goal is to determine whether $\mathcal{F}$ is (X, α, Θ)-decision robust with (i) no more than black-box access to $\mathcal{F}$ and (ii) without access to users or their decisions. We assume that X, α, and Θ are given. We provide intuition for X, α, and Θ in the next section; how to select them is beyond the scope of this work.

4 AUDITING THE ALGORITHM

In this section, we present an auditing procedure. We first introduce notation and definitions, then unpack the auditing procedure in detail. We conclude with a result, which states that the filtering algorithm $\mathcal{F}$ is asymptotically, approximately decision robust (i.e., the platform upholds the social contract on algorithmic filtering) if $\mathcal{F}$ passes the audit.

4.1 Notation and definitions

Let χ^2_r denote the chi-squared distribution with r degrees of freedom. Let $\chi^2_r(a)$ be defined such that $P(u \leq \chi^2_r(a)) = a$, where $u \sim \chi^2_r$. Let $I(\theta) \in \mathbb{R}^{r \times r}$ denote the Fisher information matrix at θ (see the Appendix for a precise definition). Let $Y = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_q) \in \mathcal{Y}^q$. Suppose that $\mathbf{y}_i \overset{\text{i.i.d.}}{\sim} p_{\mathbf{y}}(\cdot; \phi)$ for all $i \in [q]$, where $\phi \in \Phi$ is unknown. An *estimator* $\mathcal{E} : \mathcal{Y}^* \rightarrow \Phi$ produces an estimate $\mathcal{E}(Y)$ of the parameters ϕ that generated Y [21].

Definition 3 (Maximum likelihood estimator). *Suppose $\mathbf{y}_i \overset{\text{i.i.d.}}{\sim} p_{\mathbf{y}}(\cdot\,; \phi)$ for all $i \in [q]$, as described above. When it exists, the* maximum likelihood estimator *(MLE) $\mathcal{E}^+ : \mathcal{Y}^* \rightarrow \Phi$ satisfies*

$$\prod_{i=1}^{q} p_{\mathbf{y}}(\mathbf{y}_i; \mathcal{E}^+(Y)) \geq \prod_{i=1}^{q} p_{\mathbf{y}}(\mathbf{y}_i; \phi)$$

for all $\phi \in \Phi$.

4.2 Auditing procedure

The auditing procedure is given in Algorithm 1. This procedure—which is adapted from the algorithm in [11]—repeatedly runs three steps. For every input $\mathbf{x} \in X$, the auditor runs $\mathcal{F}$ and $\mathcal{B}$ to obtain the feeds Z and Z_B. The auditor randomly assigns one of them to Z' and the other to Z''. The auditor then computes the MLE $\hat{\theta}'$ of some parameter $\theta' \in \Theta$ given the samples $Z' = (\mathbf{z}'_1, \mathbf{z}'_2, \ldots, \mathbf{z}'_m)$. Similarly, the auditor computes the MLE $\hat{\theta}''$ of some parameter $\theta'' \in \Theta$ given the samples $Z'' = (\mathbf{z}''_1, \mathbf{z}''_2, \ldots, \mathbf{z}''_m)$. Lastly, the auditor computes two statistics then checks whether they exceed the threshold $\frac{2}{m}\chi^2_r(1-\alpha)$ (Line 8). If either exceeds the threshold for

Algorithm 1: Auditing procedure

```
Input: Inputs $X \in \mathcal{X}^n$; black-box access to the filtering
       algorithm $\mathcal{F} : \mathcal{X} \to \mathcal{Z}^m$; black-box access to the
       baseline algorithm $\mathcal{B} : \mathcal{X} \to \mathcal{Z}^m$; model family
       $\Theta \subset \mathbb{R}^r$; regulation parameter (or false positive rate)
       $\alpha \in [0, 1/n]$.
Output: PASS if $\mathcal{F}$ is $(X, \alpha, \Theta)$-decision robust; FAIL if not.
1  $\tau \leftarrow \frac{2}{m}\chi_r^2(1-\alpha)$;
2  for $\mathbf{x} \in X$ do
3      $Z \leftarrow \mathcal{F}(\mathbf{x})$;
4      $Z_B \leftarrow \mathcal{B}(\mathbf{x})$;
5      $(Z', Z'') \leftarrow (Z, Z_B)$ with probability $\frac{1}{2}$ and $(Z_B, Z)$ o.w.;
6      $\hat{\theta}' \leftarrow$ MLE of $\theta' \in \Theta$ given $Z'$;
7      $\hat{\theta}'' \leftarrow$ MLE of $\theta'' \in \Theta$ given $Z''$;
8      if $(\hat{\theta}' - \hat{\theta}'')^\top I(\hat{\theta}')(\hat{\theta}' - \hat{\theta}'') \geq \tau$ or
         $(\hat{\theta}' - \hat{\theta}'')^\top I(\hat{\theta}'')(\hat{\theta}' - \hat{\theta}'') \geq \tau$ then
9          Return FAIL;
10     end
11 end
12 Return PASS;
```

any input $\mathbf{x} \in X$, then the platform fails the audit. Otherwise, the platform passes the audit.

Notably, the audit requires only black-box access to the filtering algorithm $\mathcal{F}$ and baseline algorithm $\mathcal{B}$. In addition, it does not need access to users—it can be run on synthetically generated inputs X. The auditor must also pick a model family Θ. Choosing Θ to be large increases the audit's comprehensiveness while choosing Θ to be small makes it easier to compute the MLE. We recommend choosing Θ to be a family of multivariate Gaussians or an exponential family. For the most part, the auditor's main degree of freedom $\alpha \in [0, 1/n]$, which can be interpreted as the maximum false positive rate of the hypothesis test.[5] The closer α is to $1/n$, the stricter the regulation; and the closer it is to 0, the looser the regulation. One of the benefits of the audits is that α is interpretable and one can think of $n\alpha \in [0, 1]$ as the cumulative false positive rate.

4.3 Audit's guarantee

The following result shows that the audit enforces asymptotic, approximate decision robustness. As such, Algorithm 1 provides a procedure that ensures the filtered and baseline feeds are close for all $\mathbf{x} \in X$ and, consequently, that the platform upholds its end of the social contract.

Theorem 1 (Adapted from Theorem 1 in [11]). *Let Assumptions 1-3 hold. Consider an input $\mathbf{x} \in X$, and let Z' and Z'' be as defined in Assumption 3, i.e., $z'_i \overset{i.i.d.}{\sim} p_{\mathbf{z}}(\cdot;\theta')$ and $z''_i \overset{i.i.d.}{\sim} p_{\mathbf{z}}(\cdot;\theta'')$ for all $i \in [m]$ and unknown $\theta', \theta'' \in \Theta \subset \mathbb{R}^r$. Let there be two hypotheses:*

$$H_0 : \theta' = \theta'', \qquad H_1 : \theta' \neq \theta''.$$

Let $\theta^* = (\theta' + \theta'')/2$. Let $\mathcal{P} = \{p_{\mathbf{z}}(\cdot;\theta) : \theta \in \Theta\}$ denote a regular exponential family that meets the regularity conditions stated in the Appendix. If $\hat{H}$ is defined such that $\hat{H} = H_1$ if and only if

$$(\mathcal{E}^+(Z') - \mathcal{E}^+(Z''))^\top I(\theta^*)(\mathcal{E}^+(Z') - \mathcal{E}^+(Z'')) \geq \frac{2}{m}\chi_r^2(1-\alpha),$$

then $P(\hat{H} = H_1 | H = H_0) \leq \alpha$ as $m \to \infty$. If $r = 1$, then $\hat{H}$ is the UMPU test as $m \to \infty$.

Recall that $\mathcal{E}^+$ denotes the MLE (Definition 3), and observe that the hypothesis test $\hat{H}$ defined in Theorem 1 closely resembles the test in Line 8 of Algorithm 1.

The main difference is that, in reality, θ^* in Theorem 1 is unknown. As such, Line 8 provides a data-driven version of the test $\hat{H}$ in Theorem 1. Therefore, if $\mathcal{F}$ passes the audit, it is asymptotically, approximately decision robust. In fact, when $r = 1$, the audit ensures that $\mathcal{F}$ is *exactly* decision robust as $m \to \infty$. When $r > 1$, the audit executes the UMPU test, as required by decision robustness (Definition 1). As such, the audit ensures that $\mathcal{F}$ is approximately decision robust as $m \to \infty$, where it is approximate because there is no guarantee that the audit is the UMPU test (determining the UMPU test for $r > 1$ is a hard problem).

5 SIMULATIONS

In this section, we provide simulations that illustrate the effects of auditing the social contract on algorithmic filtering. In particular, we examine what types of feeds pass the audit, what the platform is incentivized to filter, and therefore how enforcing the social contract changes users' feeds. We discuss how the platform is always incentivized to ensure that the filtered feeds contain diversity (e.g., in terms of viewpoints and across topics) and how this plays nicely with existing work on the importance of content diversity.

Setup. For simplicity, suppose that $\mathcal{Z} = \mathbb{R}$, i.e., each piece of content is described by a 1-D feature. As a toy example, suppose that a platform only shows posts estimating today's local temperature.

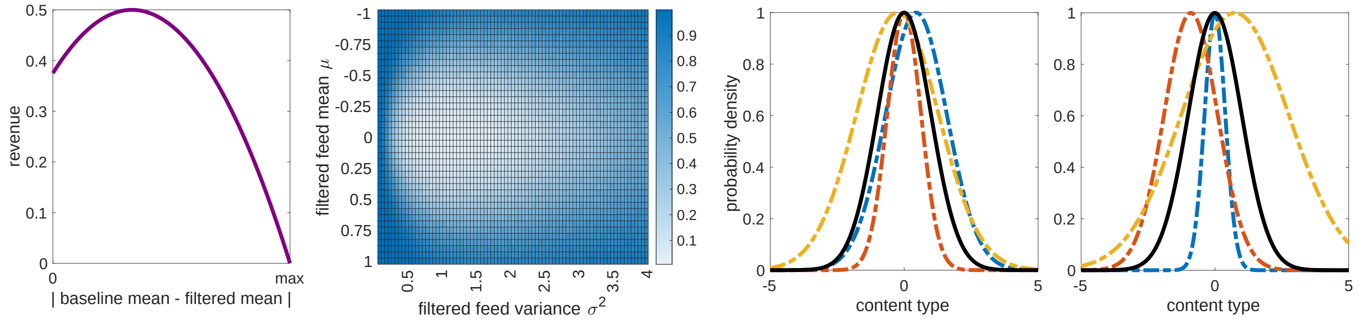
Now, consider a specific platform and user (i.e., consider a single input $X = \{\mathbf{x}\}$). We are interested in studying how decision robustness constrains the platform's filtering algorithm $\mathcal{F}$ and, consequently, affects the filtered feed that is ultimately shown to the user. To this end, suppose that the user's baseline content follows a normal distribution $\mathcal{N}(0, 1)$. Recalling that the social contract on algorithmic filtering requires that the filtered content is decision robust with respect to (i.e., "close" to) the baseline content, we examine what distributions $\mathcal{N}(\mu, \sigma^2)$ yield feeds that passes the audit. In other words, suppose the platform composes the user's filtered feed by, first, deciding on parameters (μ, σ^2), then drawing content i.i.d. from $\mathcal{N}(\mu, \sigma^2)$. When do such feeds pass the audit?

We answer this question in our discussion below. For the following analysis, let $\Theta = \mathbb{R} \times \mathbb{R}_{\geq 0}$, $\alpha = 0.01$, and $m = 30$.

Feeds that pass the audit. In Fig. 1(b), we give a heatmap, in which each cell's color indicates the proportion of filtered feeds drawn from $\mathcal{N}(\mu, \sigma^2)$ that fail the audit over 1000 simulations. As expected, the rate of failure increases as the filtered content (μ, σ^2) moves away from $(0, 1)$.

Fig. 1(c) plots the baseline distribution (in solid black) and three filtered distributions that generated feeds that passed the audit

[5] Although α can be viewed as a "rate", it is possible for the platform to verify that $\mathcal{F}$ passes the audit for a set X with complete certainty. In this sense, α simply quantifies the allowed distance between $\hat{\theta}$ and $\hat{\theta}_B$ or, equivalently, $\mathcal{F}(\mathbf{x})$ and $\mathcal{B}(\mathbf{x})$ for all $\mathbf{x} \in X$.



(a)

(b)

(c)

(d)

Figure 1: Illustration of the audit and its effect on users' feeds. In (a), we plot an example revenue function that we use in our discussion. This example captures instances where revenue is concave in the distance between the center of the baseline and filtered feeds. Suppose the baseline content follows a standard normal distribution (mean 0 and variance 1). In (b), we plot a heatmap, where cell (σ^2, μ) indicates the proportion of filtered feeds drawn from $\mathcal{N}(\mu, \sigma^2)$ that fail the audit (i.e., are decision robust with respect to the baseline feed $\mathcal{N}(0, 1)$) Lighter cells indicate passing the audit with higher frequency, and vice versa for darker cells. In (c) and (d), we illustrate this intuition. Both (c) and (d) plot the distribution of the baseline content in solid black. (c) plots three filtered feeds that *pass* the audit in dotted lines (specifically, the distributions used to generate filtered feeds that pass the audit). Similarly, (d) plots three feeds that fail the audit.

more than 80 percent of the time. Similarly, Fig. 1(d) plots the baseline distribution (in solid black) and three filtered distributions that generated feeds that failed the audit more than 80 percent of the time. Although filtered feeds that differ significantly from the baseline feed fail, Fig. 1(c)-(d) demonstrate that the platform still has flexibility in how it filters under the social contract.

Interestingly, however, the heatmap is *not* symmetric. Building on the intuition developed by [11], ensuring that the downstream effects of two feeds are (effectively) indistinguishable—in other words, ensuring decision robustness—is easier when the content is less concentrated (i.e., less peaky). This observation is reflected in Line 8 of Algorithm 1. Specifically, when the Fisher information is low (i.e., the content is diverse), the statistic on the left-hand side is small. Holding everything else equal, low Fisher information therefore reduces the likelihood of failing the audit.

The role of diversity. Consider the revenue function plotted in Fig. 1(a). In this simple setup, revenue is a strictly concave function of the distance between the means of the filtered and baseline feeds. This relationship is built on the following intuition. First, the platform earns some baseline amount of revenue by showing the user content that is close to the user's baseline content. Second, as the platform shows the user content that is close to but different from their baseline content, the revenue increases—after all, if platforms earned their maximum revenue by filtering baseline content, then there would be no need to develop sophisticated filtering algorithms. Lastly, at some point, content that is too far from the user's baseline content is no longer relevant to the user, so the revenue that the platform obtains decreases.

As illustrated by Fig. 1(b), ensuring the filtered content is sufficiently diverse can improve the platform's revenue. Specifically, consider the column in Fig. 1(b) corresponding to $\sigma^2 = 1.1$. As one moves upwards or downwards from $\mu = 0$, the proportion of audit failures increases. However, for a given μ (say, 0.5), the platform can increase the pass rate by slightly increasing the filtered feed variance. To understand why, recall that decreasing the Fisher information improves the platform's chances of passing the audit. In the context of one-dimensional Gaussians, lowering the Fisher information corresponds to increase variance.

Intuitively, decreasing the Fisher information—i.e., increasing the filtered feed's variance σ^2—increases the allowable distance between means (as long as σ^2 is not too far from 1). As such, the platform is *incentivized* to include a bit of content diversity in order to maximize its revenue while passing the audit. This result shows that the audit plays nicely with efforts to increase content diversity as a remedy for bursting filter bubbles, countering misinformation, and reducing political polarization.

6 CONCLUSION

By moderating the content that users see, social media platforms wield a great deal of influence. Although lawmakers in the US are eager to regulate social media, many efforts to do so are impeded by existing laws and protections. In particular, proposals to directly regulate harmful content or regulate platforms' content policies run into problems with Section 230 [9] and free speech rights [20].

In this work, we propose an alternate approach that can complement these efforts. In particular, we propose to regulate the *implied social contract* between users and platforms. One benefit of this approach is that regulating contracts does not require a global sense of what is"right" or "wrong" [23] (e.g., defining "misinformation" or "harmful content" in social media). Rather it increases the *accountability* of platforms to users.

In this piece, we explain how the implied social contract arises in social media. We then focus our discussion on *algorithmic filtering*—the process platforms use to curate content—and translate the social contract on algorithmic filtering into an auditable requirement on a platform's filtering algorithm. We lay out an auditing procedure (adapted from Cen and Shah [11]) and conclude with simulations.

ACKNOWLEDGMENTS

We would also like to thank the many people who provided feedback along the way, including but certainly not limited to Andrew Ilyas, Aspen Hopkins, and Zachary Schiffer. This work was supported in parts by the MIT-IBM project on "Representation Learning as a Tool for Causal Discovery" and the NSF TRIPODS Phase II grant towards Foundations of Data Science Institute.

REFERENCES

[1] 1959. Smith v. California. , 147 pages.

[2] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics* 6, 2 (2019), 2053168019848554.

[3] Julia Angwin, Ariana Tobin, and Madeleine Varner. 2017. Facebook (Still) Letting Housing Advertisers Exclude Users by Race. https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin. (November 2017).

[4] R. R. Bahadur. 1964. On Fisher's Bound for Asymptotic Variances. *The Annals of Mathematical Statistics* 35, 4 (1964), 1545–1552.

[5] Mohammed Bedjaoui, Nadia Elouali, Sidi Mohamed Benslimane, and Erhan Şengel. 2022. Suggestion pattern on online social networks: between intensity, effectiveness and user's satisfaction. *The Visual Computer* 38, 4 (2022), 1331–1343.

[6] Brenda L Berkelaar. 2014. Cybervetting, online information, and personnel selection: New transparency expectations and the emergence of a digital social contract. *Management Communication Quarterly* 28, 4 (2014), 479–506.

[7] Michael Bossetta. 2018. The digital architectures of social media: Comparing political campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 US election. *Journalism & mass communication quarterly* 95, 2 (2018), 471–496.

[8] D. Boucher and P. Kelly. 2003. *The Social Contract from Hobbes to Rawls.* Taylor & Francis. https://books.google.com/books?id=Z4OJAgAAQBAJ

[9] Valerie Brannon and Eric Holmes. 2021. Section 230 of the Communications Decency Act: An Overview. Congressional Research Service.

[10] George Casella and Roger L Berger. 2021. *Statistical inference.* Cengage Learning.

[11] Sarah H. Cen and Devavrat Shah. 2021. Regulating algorithmic filtering on social media. *Advances in Neural Information Processing Systems* 34 (2021), 6997–7011.

[12] New Jersey Courts. 2018. Express Or Implied Contract (Charge 4.10E). https://www.njcourts.gov/attorneys/civilcharges.html

[13] Richard L Cruess and Sylvia R Cruess. 2008. Expectations and obligations: professionalism and medicine's social contract with society. *Perspectives in biology and medicine* 51, 4 (2008), 579–598.

[14] Joan Donovan. 2020. Redesigning consent: Big data, bigger risks. *The Harvard Kennedy School Misinformation Review* (2020).

[15] Steven W. Feldman. 2016. Statutes and Rules of Law as Implied Contract Terms: The Divergent Approaches and a Proposed Solution. *University of Pennsylvania Journal of Business Law* 19 (2016), 809–869.

[16] Roberto Garvin. 2019. How social networks influence 74 percent of shoppers for their purchasing decisions today. https://awario.com/blog/how-social-networks-influence-74-of-shoppers-for-their-purchasing-decisions-today. (May 2019).

[17] Michael Gibbons. 1999. Science's new social contract with society. *Nature* 402, 6761 (1999), C81–C84.

[18] John C. Harsanyi. 1953. Cardinal Utility in Welfare Economics and in the Theory of Risk-taking. *Journal of Political Economy* 61 (1953), 434 – 435.

[19] Thomas Hobbes. 2011. *Leviathan.* CreateSpace Independent Publishing Platform. https://books.google.com/books?id=RfAxXwAACAAJ

[20] Daphne Keller. 2021. Amplification and its discontents: Why regulating the reach of online content is hard. *J. FREE SPEECH L.* 1 (2021), 227–268.

[21] E. L. Lehmann and George Casella. 1998. *Theory of Point Estimation* (2nd ed.). Springer-Verlag, New York, NY, USA.

[22] Alexander Ly, Maarten Marsman, Josine Verhagen, Raoul Grasman, and Eric-Jan Wagenmakers. 2017. A Tutorial on Fisher Information. *Journal of Mathematical Psychology* 80 (2017), 40–55.

[23] NL. 1856. On the Theory of Implied Contracts. *American Law Register* (1856), 321–334.

[24] Victor Pickard. 2021. A new social contract for platforms. *Regulating Big Tech: Policy responses to digital dominance* (2021), 323–337.

[25] John Rawls. 1971. *A Theory of Justice.* Cambridge (Mass.).

[26] Samuel C Rhodes. 2022. Filter bubbles, echo chambers, and fake news: how social media conditions individuals to be less critical of political misinformation. *Political Communication* 39, 1 (2022), 1–22.

[27] Jessica L. Rich. 2014. Director of the Federal Trade Commission Bureau of Consumer Protection, to Erin Egan, Chief Privacy Officer, Facebook, and to Anne Hoge, General Counsel, WhatsApp Inc.

[28] Michel Rosenfeld. 1984. Contract and justice: The relation between classical contract law and social contract theory. *Iowa L. Rev.* 70 (1984), 769.

[29] Warren L. Shattuck. 1959. Contracts in Washington, 1937-1957. *Washington Law Review & State Bar Journal* 34 (1959), 24–77.

[30] Jaime E Sidani, Ariel Shensa, Beth Hoffman, Janel Hanmer, and Brian A Primack. 2016. The Association Between Social Media Use and Eating Concerns Among US Young Adults. *Journal of the Academy of Nutrition and Dietetics* 116, 9 (2016), 1465–1472.

[31] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social Media, Political Polarization, and Political Disinformation: A Review of The Scientific Literature. (March 2018).

[32] William Vickrey. 1945. Measuring marginal utility by reactions to risk. *Econometrica: Journal of the Econometric Society* (1945), 319–333.

[33] Georgia Wells, Jeff Horwitz, and Deepa Seetharaman. 2021. Facebook knows Instagram is Toxic for Teen Girls, Company Documents Show. *The Wall Street Journal* (2021).

A DEFINITIONS

Definition 4 (Fisher information matrix). *For a family of distribution $\{p_{\mathbf{z}}(\cdot\,;\theta) : \theta \in \Theta\}$, the* Fisher information matrix $I(\theta) \in \mathbb{R}^{r\times r}$ *at $\theta \in \Theta$ is a positive semi-definite matrix, where*

$$[I(\theta)]_{ij} = \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\cdot\,;\theta)}\left[\frac{\partial}{\partial\theta_i}\log p_{\mathbf{z}}(\mathbf{z};\theta)\frac{\partial}{\partial\theta_j}\log p_{\mathbf{z}}(\mathbf{z};\theta)\right].$$

Recall from Assumption 1 that Z comprises m samples $\mathbf{z}_i \in \mathcal{Z}$ drawn i.i.d. from $p_{\mathbf{z}}(\cdot\,;\theta)$, for $\theta \in \Theta$.

Definition 5 (Asymptotic normality and efficiency). *An estimator $\mathcal{E} : \mathcal{Z}^m \to \Theta$ is* asymptotically normal and efficient *if:*

$$\sqrt{m}\,(\mathcal{E}(Z) - \theta) \xrightarrow{d} \mathcal{N}(\mathbf{0}_r, I^{-1}(\theta)),$$

as $m \to \infty$ for all $\theta \in \Theta$ where $I^{-1}(\theta)$ denotes the inverse of the Fisher information matrix at θ.

Definition 6 (Uniformly most powerful test). *Suppose that $\hat{G} \in \{G_0, G_1\}$ denotes a binary hypothesis test, where*

$$G_0 : \phi = \phi', \qquad G_1 : \phi \neq \phi',$$

for $\phi, \phi' \in \Phi$. Let G denotes the true hypothesis. If $\hat{G}$ solves

$$\max_{\hat{G}'} P(\hat{G}' = G_1 | G = G_1) \quad s.t. \quad P(\hat{G} = G_1 | G = G_0) \leq \alpha,$$

for all $\phi, \phi' \in \Phi$ (i.e., maximizes the true positive rate while having a false positive rate—or significance level—no greater than α for all $\phi, \phi' \in \Phi$), then it is the uniformly most powerful (UMP) *test.*

Definition 7 (Unbiased test). *Consider the setup in Definition 6. A hypothesis test $\hat{G}$ is* unbiased *if $P(\hat{G} = G_1 | G = G_1) \geq \delta \geq P(\hat{G} = G_1 | G = G_0)$ for some $\delta \in [0, 1]$.*

Definition 8 (Uniformly most powerful unbiased test). *The* uniformly most powerful unbiased (UMPU) *test is the UMP test among all unbiased tests.*

B REGULARITY CONDITIONS

Let $\mathcal{P} = \{p_{\mathbf{z}}(\cdot;\theta) : \theta \in \Theta\}$. The regularity conditions used in Theorem 1 are given next.

(1) Θ is a compact and open set of $\mathbb{R}^r$.
(2) $\mathbf{z} \overset{\text{i.i.d.}}{\sim} p_{\mathbf{z}}(\cdot;\theta)$ for $\theta \in \Theta$ and $\theta_1 \neq \theta_2$ implies $p_{\mathbf{z}}(\cdot;\theta_1)$ and $p_{\mathbf{z}}(\cdot;\theta_2)$ are distinct.
(3) The support of $p_{\mathbf{z}}(\cdot;\theta)$ is independent of $\theta \in \Theta$.
(4) All second-order partial derivatives of $\log p_{\mathbf{z}}(\mathbf{z};\theta)$ with respect to θ exist and are continuous in θ.

(5) For any $\theta_0 \in \Theta$, there exists a neighborhood of θ_0 and a function $\Pi(\mathbf{z})$, where $\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\,\cdot\,;\,\theta_0)}[\Pi(\mathbf{z})] < \infty$ and

$$\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_{\mathbf{z}}(\mathbf{z};\theta)\right| \leq \Pi(\mathbf{z}),$$

for all $\mathbf{z} \in \mathcal{Z}$, all θ in the neighborhood of θ_0, and $i, j \in [r]$.

(6) If θ^* is the data generating parameter,

(a) $\frac{\partial}{\partial\theta_i}\log p_{\mathbf{z}}(\mathbf{z};\theta^*)$ is square integrable for all $i \in [r]$.

(b) $\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\cdot;\theta^*)}\left[\frac{\partial}{\partial\theta_i}\log p_{\mathbf{z}}(\mathbf{z};\theta^*)\right] = 0$.

(c) Fisher information:

$$\begin{aligned}[I(\theta^*)]_{ij} &= \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\cdot;\theta^*)}\left[\frac{\partial}{\partial\theta_i}\log p_{\mathbf{z}}(\mathbf{z};\theta^*)\frac{\partial}{\partial\theta_j}\log p_{\mathbf{z}}(\mathbf{z};\theta^*)\right]\\ &= -\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\cdot;\theta^*)}\left[\frac{\partial^2}{\partial\theta_i\theta_j}\log p_{\mathbf{z}}(\mathbf{z};\theta^*)\right]\end{aligned}$$

(d) $I(\theta^*)$ is positive-definite and invertible.

(7) Θ is a convex set.

As stated in [11], there are variations on these regularity conditions, cf. [4, 21, 22].

C PROOF OF THEOREM 1

As stated in the theorem statement, Theorem 1 is adapted from Theorem 1 in [11]. As such, the proof below adapts the proof in [11] as well.

Proof. The definitions and regularity conditions required for Theorem 1 are given in Appendices A-A. Condition 1 (specifically, compactness) and Condition 4 ensure that the MLE exists. From Conditions 1-6, the MLE is asymptotically normal and efficient [4, 21, 22]. Under Condition 7, $I((\theta' + \theta'')/2)$ exists for $\theta', \theta'' \in \Theta$. Under Condition 7, $I((\theta + \theta')/2)$ exists for $\theta', \theta'' \in \Theta$.

By the asymptotic normality and efficiency of the MLE,

$$\begin{aligned}\sqrt{m}(\mathcal{E}^+(Z') - \theta') &\xrightarrow{d} \mathcal{N}(\mathbf{0}_r, I^{-1}(\theta'))\\ \sqrt{m}(\mathcal{E}^+(Z'') - \theta'') &\xrightarrow{d} \mathcal{N}(\mathbf{0}_r, I^{-1}(\theta''))\end{aligned}$$

as $m \to \infty$, where $\mathbf{z}_i' \overset{\text{i.i.d.}}{\sim} p(\cdot;\theta')$ and $Z' = (\mathbf{z}_1', \dots, \mathbf{z}_m')$, and similarly for Z''. Therefore, as $m \to \infty$,

$$\sqrt{m}(\mathcal{E}^+(Z') - \theta' - \mathcal{E}^+(Z'') + \theta'') \xrightarrow{d} \mathcal{N}(\mathbf{0}_r, I^{-1}(\theta') + I^{-1}(\theta''))$$

Under the hypothesis test statement in Theorem 1, $H_0 : \theta' = \theta'' = \theta^*$. Therefore, if $H = H_0$,

$$\sqrt{m}(\mathcal{E}^+(Z') - \mathcal{E}^+(Z'')) \xrightarrow{d} \mathcal{N}(\mathbf{0}_r, 2I^{-1}(\theta^*)) \qquad (1)$$

as $m \to \infty$. As a result, the two-sample, two-sided hypothesis test becomes a two-sample, one-sided test of on the mean of a multivariate Gaussian random variable as $m \to \infty$. Under (1),

$$(\mathcal{E}^+(Z') - \mathcal{E}^+(Z''))^\top I(\theta^*)(\mathcal{E}^+(Z') - \mathcal{E}^+(Z'')) \sim \frac{2}{m}\chi_r^2$$

Therefore, if $\hat{H}$ satisfies:

$$\begin{gathered}\hat{H} = H_1\\ \Longleftrightarrow\\ (\mathcal{E}^+(Z') - \mathcal{E}^+(Z''))^\top I(\theta^*)(\mathcal{E}^+(Z') - \mathcal{E}^+(Z'')) \geq \tfrac{2}{m}\chi_r^2(1-\alpha)\end{gathered}$$

then $\hat{H}$ has a FPR $\leq \alpha$, as desired.

$\hat{H}$ as defined above is the UMPU test with significance level α when $r = 1$ (cf. Section 8.3 of [10]). As such, $\hat{H}$ would *exactly* test whether $\mathcal{F}$ is decision robust and the platform upholds the social contract on algorithmic filtering as $m \to \infty$. When $r > 1$, there is no guarantee that $\hat{H}$ is asymptotically the UMPU test (in general, determining the UMPU test is known to be difficult for $r > 1$). □