Social Media and Misleading Information in a Democracy: A Mechanism Design Approach

Aditya Dave, Student Member, IEEE, Ioannis Vasileios Chremos, Student Member, IEEE, and Andreas A. Malikopoulos, Senior Member, IEEE

Abstract—In this article, we present a resource allocation mechanism to incentivize misinformation filtering among strategic social media platforms and, thus, to indirectly prevent the spread of fake news. We consider the presence of a strategic government and private knowledge of how misinformation affects the users of the social media platforms. Our proposed mechanism strongly implements all generalized Nash equilibria for efficient filtering of misleading information in the induced game, with a balanced budget. We also show that for quasi-concave utilities, our mechanism implements a Pareto efficient solution.

Index Terms—Fake news, mechanism design, misinformation, Nash implementation, social media.

I. INTRODUCTION

For the past few years, political commentators have been indicating that we live in a post-truth era [1], wherein the deluge of information available on the Internet has made it extremely difficult to identify facts. As a result, individuals have developed a tendency to form their opinions based on the believability of presented information rather than its truthfulness [2]. This phenomenon is exacerbated by the business practices of social media platforms who often seek to maximize the engagement of their users at all costs. The algorithms developed by platforms for this purpose often promote conspiracy theories among their users [3]. Thus, social media platforms form an ideal terrain to conduct political misinformation campaigns. Such campaigns are effective tools to disrupt democratic institutions because the functioning of stable democracies relies on common knowledge about the political actors and the processes they can use to gain public support [4]. The trust held by the citizens of a democracy on common knowledge includes that: 1) all political actors act in good faith when contesting for power; 2) elections lead to a fair transfer of power; and 3) democratic institutions ensure that elected officials wield their power responsibly. In contrast, citizens of democracies often have a contested knowledge regarding who should hold power and how they should use it [4]. The introduction of alternative facts can diminish the trust on common knowledge about democracy, especially if they become accepted beliefs. Such disruptions in the trust on common knowledge can be found in the 2016 U.S. elections [5] and 2016 Brexit Campaign [6], where the spread of misinformation on social media platforms resulted in mistrust toward the voting results.

In this article, we seek to tackle the phenomenon of misinformation by considering a group of social media platforms, whose users represent the citizens in a democracy and a democratic government. Every post on the platforms is associated with a parameter that captures its informativeness, which takes values between two extremes: 1) completely factual and 2) complete misinformation. In our article, posts that exhibit misinformation can lead to a decrease in trust on common knowledge among the users [4], [7]. We consider that social media platforms pose the technologies to filter, or label, posts, which can eventually diminish trust on common knowledge. Thus, the government seeks to incentivize the social media platforms to use these technologies to filter misinformation.

Motivated by capitalistic values, we introduce a misinformation filtering game to describe the interactions between the social media platforms and the government. In this game, each platform acts as a strategic player seeking to maximize the advertisement revenue from user engagement [8]. User engagement is a metric that can be used to quantify the interactions of users with a platform. Recent results from the literature on misinformation in social media platforms have indicated that increasing filtering of misinformation may decrease user engagement [9]. There are many possible reasons for this phenomenon. First, filtering reduces the total number of posts on the platform. Second, users whose opinions are filtered may perceive this action as dictatorial censorship [10], and as a result, they may choose to express their opinions on other platforms. Finally, misinformation tends to elicit stronger reactions, e.g., surprise, joy, as compared to factual posts [11], which may increase user engagement. Thus, each platform is reluctant to filter misinformation. In our framework, the government is also a strategic player, whose utility increases as the trust on common knowledge of any user increases. Consequently, the government’s utility increases with an increase in misinformation filtering by the social media platforms. The government seeks to make an investment to incentivize the platforms to filter misinformation. We use mechanism design [12]–[15] to optimally distribute this investment among the platforms and implement an optimal level of filtering.

Our primary contribution is to design an indirect mechanism to incentivize social media platforms to filter misinformation. We show that our mechanism is feasible, budget balanced, individually rational, and strongly implementable at the equilibria of the induced game. We prove the existence of at least one generalized Nash equilibrium (GNE) and show that our mechanism is Pareto efficient.

The rest of this article is organized as follows. In Section II, we provide the problem formulation. In Section III, we present our mechanism, and in Section IV, we establish its main properties and present a descriptive example. In Section V, we conclude and present some directions for future research.

II. PROBLEM FORMULATION

We consider a democratic society with a nonempty set of social media platforms \( \mathcal{I} = \{1, \ldots, I\} \), \( I \in \mathbb{N} \), and a government. We refer to the social media platforms and the government collectively as the...
players and denote the set of all players by \( \mathcal{J} = \mathcal{I} \cup \{0\} \), where the index 0 corresponds to the government. The players strategically take actions in a misinformation filtering game, as described next.

Let the informativeness of a post on platform \( i \in \mathcal{I} \) be \( x_i \in [0, 1] \), where \( x_i = 0 \) indicates that the post contains complete misinformation, and \( x_i = 1 \) indicates that the post is completely factual. Our hypothesis states that the emergence of posts with many falsehoods, i.e., \( x_i \rightarrow 0 \), decreases the trust of the users on common knowledge about democracy [4], [7]. Recall that the common knowledge refers to knowledge of the political actors and the processes to gain public support. Each platform \( i \in \mathcal{I} \) has the technological means to detect and filter misinformation. In the misinformation filtering game, the action \( a_i \in \mathcal{A} = [0, 1] \) of platform \( i \) represents the level of filtering imposed by \( i \). Action \( a_i \) minimizes the spread of a post with informativeness \( x_i < a_i \), while a post with \( x_i \geq a_i \) is unaffected. In practice, misinformation filters can be implemented by either placing warnings on each post with \( x_i < a_i \) or limiting the reach of such posts. Thus, the action \( a_i \) is a lower threshold on informativeness that is accepted by platform \( i \). To this end, we call \( a_i \) the filter of platform \( i \).

Each platform \( i \in \mathcal{I} \) generates revenue by monetizing the engagement of their users with advertisements [8]. With an increase in filtering, there is a decrease in user engagement [9]. Users may perceive filters as censorship [10], and as a result, they may choose to express their opinions on other platforms. Consider, for example, platform \( i \in \mathcal{I} \) with a filter \( a_i > a_i \). Some users of \( i \), whose posts have been marked up by the filter, may migrate to platform \( i \) and increase the engagement of \( i \). This motivates us to define a set of competing platforms.

**Definition 1:** For each platform \( i \in \mathcal{I} \), the set \( \mathcal{C}_i \subseteq \mathcal{I} \), with \( i \in \mathcal{C}_i \), is the set of competing platforms whose choice of filters has an impact on the engagement of platform \( i \in \mathcal{I} \).

To simplify our exposition, we consider that for \( i, k \in \mathcal{I} \), if \( i \in \mathcal{C}_k \), then \( k \in \mathcal{C}_i \). However, our mechanism can easily be extended to allow asymmetric competition among platforms.

**Definition 2:** The valuation function of a social media platform \( i \in \mathcal{I} \) is \( v_i(a_k : k \in \mathcal{C}_i) : \mathcal{A}^{|\mathcal{C}_i|} \rightarrow \mathbb{R}_{\geq 0} \). It is a decreasing function with respect to \( a_i \) and strictly increasing with respect to \( a_k \) for all \( k \in \mathcal{C}_i \), where \( \mathcal{C}_i = \mathcal{C} \setminus \{i\} \).

The valuation function \( v_i(a_k : k \in \mathcal{C}_i) \) gives the revenue of platform \( i \) from user engagement after filtering by all platforms. A higher value of \( a_i \) will decrease the revenue of platform \( i \). A higher value of \( a_k \) for another competing platform \( k \in \mathcal{C}_i \) will increase the revenue of platform \( i \). Recall from the discussion in the previous section that filtering of misinformation in a platform increases the trust of their users on common knowledge about democracy. Thus, for each \( i \in \mathcal{I} \), we define the average trust function on common knowledge.

**Definition 3:** The average trust function on common knowledge of the users of platform \( i \in \mathcal{I} \) is \( h_i(a_i) : \mathcal{A} \rightarrow [0, 1] \) and it is a strictly increasing function with respect to \( a_i \).

The average trust function \( h_i(a_i) \) captures the impact of filter \( a_i \) on the trust on common knowledge across the users of platform \( i \). A low value of \( h_i(a_i) \) implies that \( a_i \) leads to low trust on common knowledge for the users of platform \( i \), and vice versa. In practice, platform \( i \) can measure the opinions of their users through surveys [16] and, thus, eventually estimate the impact of filter \( a_i \) using the average trust function \( h_i(a_i) \).

Recall that, in our framework, the government is the strategic player \( 0 \in \mathcal{J} \) who seeks to maximize the trust on common knowledge of the users of all social media platforms. Therefore, the government selects an action \( a_0 \in \mathcal{A} = [0, 1] \) that designates a lower bound, which must be satisfied by the aggregate average trust of all platforms in \( \mathcal{I} \). To this end, we refer to the action \( a_0 \) as the government’s lower bound on trust on common knowledge.

Let \( N_i \in \mathbb{N} \) be the total number of users of the social media platform \( i \in \mathcal{I} \). The fraction of the number of users of \( i \) with respect to the total number of users of all platforms is \( n_i = \sum_{k \in \mathcal{I}} N_i = 1 \). The fraction \( n_i \) represents the contribution of users in platform \( i \) on the average trust on common knowledge. Since \( \sum_{k \in \mathcal{I}} n_i = 1 \), the aggregate average trust is \( \sum_{k \in \mathcal{I}} n_i \cdot h_i(a_k) \). In our framework, the government’s role is to select the lower bound \( a_0 \) for the aggregate average trust. After the government decides on \( a_0 \), each platform \( i \in \mathcal{I} \) that participates in the game must select a filter \( a_i \) that satisfies

\[
a_0 - \sum_{k \in \mathcal{I}} n_i \cdot h_i(a_k) \leq 0.
\]

Next, we define the government’s valuation function.

**Definition 4:** The valuation function of the government is \( v_0(a_0) : [0, 1] \rightarrow \mathbb{R}_{\geq 0} \), and it is an increasing function with respect to the lower bound \( a_0 \).

The government’s valuation function \( v_0(a_0) \) assigns a monetary value to the lower bound \( a_0 \). Recall that the government seeks to increase the trust on common knowledge among the users of all platforms. Thus, the government’s valuation increases as the lower bound on aggregate average trust increases. The government also has a fixed finite budget \( b_0 \in \mathbb{R}_{\geq 0} \), which denotes the maximum possible investment.

The private and public information structure corresponding to each player is as follows.

1. **Public information:** The set of competing platforms \( \mathcal{C}_i \), set of feasible actions \( \mathcal{A} \), and fraction of users \( n_i \) of each platform \( i \in \mathcal{I} \) are known to all players in set \( \mathcal{J} \).
2. **Valuation functions:** The valuation function \( v_i(\cdot) \) of each platform \( i \in \mathcal{I} \) is known only to platform \( i \). Similarly, the valuation function \( v_0(\cdot) \) and the budget \( b_0 \) of the government are known only to the government.
3. **Average trust functions:** The average trust function \( h_i(\cdot) \) of platform \( i \in \mathcal{I} \) is known only to platform \( i \).

We impose the following assumptions in our game.

**Assumption 1:** For each platform \( i \in \mathcal{I} \), \( |\mathcal{C}_i| \geq 3 \).

This assumption simplifies the exposition of our mechanism. Assumption 1 implies that each user frequents multiple social media platforms. For the case with \( |\mathcal{C}_i| \geq 2 \) and extended results, see our online preprint [17].

**Assumption 2:** The valuation function \( v_i(a_k : k \in \mathcal{C}_i) : \mathcal{A}^{|\mathcal{C}_i|} \rightarrow \mathbb{R}_{\geq 0} \) of each social media platform \( i \in \mathcal{I} \) is a concave and differentiable function with respect to \( a_k \).

The concavity of \( v_i(a_k : k \in \mathcal{C}_i) \) captures the diminishing marginal change in engagement due to additional filtering. The higher the value of \( a_i \), the more users of platform \( i \) will perceive the filter as censorship. Thus, for platform \( i \), increasing a low-value filter may lead to a smaller loss in engagement as compared to increasing a high-value filter.

**Assumption 3:** The average trust function \( h_i(a_k) : \mathcal{A} \rightarrow [0, 1] \) of each social media platform \( i \in \mathcal{I} \) is a concave and differentiable function with respect to \( a_k \).

The concavity of \( h_i(a_k) \) implies that, for large values of \( a_k \), a small incremental change in \( a_k \) would not have a significant impact on the average trust of users. Practically, this implies that low values of \( a_k \) will have a major impact on the average trust.

**Assumption 4:** The valuation function of the government \( v_0(a_0) : [0, 1] \rightarrow \mathbb{R}_{\geq 0} \) is a concave and differentiable function with respect to the lower bound \( a_0 \).

Practically, for high values of \( a_0 \), the government might not be interested in investing additional resources to increase \( a_0 \) even more, as the impact on average trust would not be significant. Nevertheless, we also analyze our system by relaxing Assumptions 2–4 in Section IV-A.
**Assumption 5:** The output of the function $h_i(a_i)$ can be monitored by any competing platform $l \in C_i$, and a violation of the condition (1) can be detected by the government.

Assumption 5 helps us enforce the mechanism, presented in Section III, in a static environment. In the mechanism, each platform $i \in \mathcal{I}$ commits to a minimum value of their average trust function, to be achieved by choosing an appropriate $a_i$. Consider that a platform $i$ selects a value $a_i$ that fails to satisfy this commitment. The government can detect a violation of (1) by gauging public opinion on the Internet. However, the government does not know the function $h_i(\cdot)$ and, thus, would penalize each platform in $\mathcal{I}$ equally for the violation of (1).

To avoid a penalty for the failure of platform $i$, a competing platform $l \in C_i$ can report the violation. Thus, it is reasonable to consider that each platform $i \in \mathcal{I}$ monitors the output $h_i(a_i)$ of each competing platform $l \in C_i$ to maximize their own utility. In future research, we could potentially relax Assumption 5 using a dynamic mechanism [14].

**Assumption 6:** The government ensures that any social media platform $i \in \mathcal{I}$ that does not participate in the mechanism receives no benefits from the filters of participating platforms.

In static mechanisms, the ability to exclude a player from receiving benefits of some common resource is a necessary condition for voluntary participation of players without any monetary investment [18]. This condition is often assumed implicitly in the literature [12]. In our mechanism, the government can make an investment up to the budget $b_0$. Thus, we assume partial excludability in Assumption 6, where a nonparticipating platform $i$ still receives the maximum valuation for selecting filter $a_i = 0$, but cannot receive benefits from the filters of any participating platforms. In practice, the government can publicize that platform $i$ is noncooperative in a collective endeavor to filter misinformation. The resulting loss in credibility among the users of participating platforms will minimize their migration to platform $i$. In future research, we can relax this assumption using a dynamic mechanism [19].

### A. Problem Statement

To resolve the conflict of interest between the government and the platforms, the government hires a social planner to design a mechanism and impose the misinformation filtering game. The mechanism must 1) incentivize all platforms to voluntarily participate in the game and 2) induce a selection of filters that maximizes the social welfare. The social welfare is the sum of utilities of all players, defined next.

To meet these objectives, the social planner asks each player $i \in \mathcal{I}$ to send a message $m_i$ from a set $\mathcal{M}_i$. Using the message profile $m = (m_0, \ldots, m_{|\mathcal{I}|})$, the social planner assigns a tax $\tau_i(m) \in \mathbb{R}$ for each platform $i \in \mathcal{I}$, and an investment $\tau_0(m) \in \mathbb{R}_{\geq 0}$ for the government. The message and tax of each player are defined in Section III-B. By convention, a tax $\tau_i(m) > 0$ is a payment made by player $i \in \mathcal{J}$, and a tax $\tau_i(m) < 0$ is a subsidy given to player $i$. While the taxes of the platforms can be either payments or subsidies, the government may never collect a subsidy from any platform. Note that the social planner must not receive any profit, nor incur any losses, for designing and implementing the mechanism, i.e., the mechanism should be budget balanced with $\sum_{i \in \mathcal{I}} \tau_i(m) = 0$.

**Definition 5:** The utility of platform $i \in \mathcal{I}$ is $u_i(m, a_i : k \in \mathcal{C}_i) := v_i(a_i : k \in \mathcal{C}_i) - \tau_i(m)$, while government’s utility is $u_0(m, a_0) := v_0(a_0) - \tau_0(m)$.

Then, the social welfare to be maximized by the social planner is $u_0(m, a_0) + \sum_{i \in \mathcal{I}} u_i(m, a_i : k \in \mathcal{C}_i)$.

**Problem 1:** The social planner’s optimization problem is

$$\max_{a_i} \left( v_0(a_0) - \tau_0(m) + \sum_{i \in \mathcal{I}} (v_i(a_i : k \in \mathcal{C}_i) - \tau_i(m)) \right)$$

subject to $0 \leq a_i \leq 1, \quad \forall i \in \mathcal{I}$

$$a_0 - \sum_{i \in \mathcal{I}} n_i \cdot h_i(a_i) \leq 0$$

$$0 \leq \tau_0(m) \leq b_0$$

$$\sum_{i \in \mathcal{I}} \tau_i(m) = 0$$

where $a = (a_0, \ldots, a_{|\mathcal{I}|})$ and $\tau(m) = (\tau_0(m), \ldots, \tau_{|\mathcal{I}|}(m))$.

In Problem 1, (3) ensures that the aggregate average trust of all users satisfies the government’s lower bound $a_0$. (4) restricts the government’s investment $\tau_0(m)$ to be within the available budget, and (5) ensures budget balance. The objective function of Problem 1 is differentiable and concave, and the set of feasible solutions is nonempty, convex, and compact. Thus, Problem 1 has a unique solution. However, this solution cannot be computed directly by the social planner because she has no knowledge of the functional form of either the valuation function $v_i(\cdot)$ of any player $i \in \mathcal{J}$ or the average trust function $h_i(\cdot)$ of any platform $i \in \mathcal{I}$. If the social planner simply asks the players to report their private information, the players may not be truthful. Thus, the social planner seeks to design the taxes $\tau_i(m)$ for each player $i \in \mathcal{J}$ that will incentivize the players to be truthful while also maximizing the social welfare.

**Remark 1:** The government has no compelling reason to misreport their budget $b_0$ to the social planner. Thus, we consider that the social planner has knowledge of $b_0$.

**Remark 2:** By maximizing the social welfare, the utility of each player is maximized in Problem 1. Hence, players have an incentive to participate in the mechanism. Note that the government cannot design the mechanism because they would optimize only their own utility $u_0(m, a_0)$. Thus, the social planner is essential to design and implement our mechanism.

### III. MECHANISM DESIGN APPROACH

In this section, we present a two-step mechanism to incentivize the filtering of misinformation among social media platforms. The aim of step 1 is to ensure the voluntary participation of all platforms. The aims of step 2 are to: 1) gain truthful information from the platforms; 2) derive the optimal investment; and 3) maximize the social welfare.

**A. Step 1: The Participation Step**

In step 1 of the mechanism, each social media platform $i \in \mathcal{I}$ must decide whether to participate in the mechanism. This decision is taken with complete knowledge of the rules of step 2, described in the next subsection. Let platform $i \in \mathcal{I}$ choose to not participate. Platform $i$ neither pays taxes nor receives subsidies, i.e., $\tau_i(m) = 0$, and they are free to select the filter $a_i = 0$ that maximizes $v_i(a_i : k \in \mathcal{C}_i)$. Meanwhile, another competing platform $l \in \mathcal{C}_i \setminus \{i\}$ may decide to participate in the mechanism and, subsequently, implement a nonzero filter $a_l$. From Assumption 6, the government ensures that platform $i$ receives no utility from the competing filter $a_l$. Thus, the utility of platform $i$ is $v_i(a_k = 0 : k \in \mathcal{C}_i)$. We will use this utility in Theorem 4 to establish the voluntary participation of all platforms in our mechanism.

**B. Step 2: The Bargaining Step**

In step 2, the social planner asks each player $i \in \mathcal{J}$ to broadcast a message $m_i \in \mathcal{M}_i$. For all $i \in \mathcal{I}$, let $\mathcal{D}_i = \mathcal{C}_i \cup \{0\}$, and $\mathcal{D}_{-i} = \mathcal{D}_i \setminus \{i\}$. The message of platform $i$ is

$$m_i := (\tilde{h}_i, \tilde{p}_i, \tilde{a}_i)$$

where $\tilde{h}_i$, $\tilde{p}_i$, and $\tilde{a}_i$ are the components of $m_i$ that are broadcast by platform $i$. The social planner seeks to design the messages $m_i$ for each player $i \in \mathcal{J}$ that will incentivize the players to be truthful while also maximizing the social welfare.
where $\tilde{h}_i \in \mathbb{R}_{>0}$ is the minimum average trust that platform $i$ proposes to achieve through filtering; $\tilde{p}_i := \{\tilde{p}_i^l : l \in D_i\}, \tilde{p}_i \in \mathbb{R}^{\mathbb{R}_{>0}}$, is the collection of prices that platform $i$ is willing to pay or receive per unit change in the filters of other competing platforms (except $i$) and the government’s lower bound; and $\tilde{a}_i = (\tilde{a}_i^l : k \in D_i), \tilde{a}_i \in \mathbb{R}^{\mathbb{R}_{>0}}$, is the profile of filters proposed by platform $i$ for all competing platforms (including $i$) and government’s lower bound.

Remark 3: Note that each platform proposes a filter for themselves, denoted by $\tilde{a}_i^l$, in their message $m_i$. However, platform $i$ does not propose a price for $\tilde{a}_i^l$. Thus, every platform can influence their filter, but not the associated price.

The message of the government is $m_0 := (\tilde{p}_0, \tilde{a}_0^0)$, where $\tilde{p}_0 \in \mathbb{R}_{>0}$ is the price that the government is willing to pay or receive per unit change in the average trust, and $\tilde{a}_0^0 \in \mathbb{R}$ is the lower bound proposed by the government. Note that our social planner respects the privacy of each platform $i \in I$ since the valuation function $v_i(a_k : k \in C_i)$ or average trust function $h_i(a_i)$ are not requested. Similarly, the government is not forced to reveal the functional form of $v_0(a_0)$. Each player $i \in J$ is free to send any feasible value of the message $m_i$. Given the messages $m := (m_0, \ldots, m_J), \text{the social planner allocates the following parameters to the players:}$

1) The social planner allocates a filter to each platform $i \in I$ and a lower bound to the government such that the constraints of Problem 1 are satisfied. The filter allocated by the social planner to platform $i$ is $h_i(m) := \sum_{k \in C_i} \tilde{h}_k \cdot \tilde{a}_k^l$, i.e., the average of the filters proposed by all competing platforms including $i$. The lower bound allocated by the social planner to the government is $a_0(m) = \sum_{k \in D_i} \tilde{a}_k^l$, i.e., the average of the lower bounds proposed by all platforms and the government.

2) The social planner allocates a minimum average trust $\eta_i(m) := \min\left(\frac{1}{|C_i|} \sum_{k \in C_i} n_k \cdot h_k, 1\right), \eta_i(m) \in [0, 1]$ to each platform $i \in I$, where the social planner will not accept a message $m_i$ that might lead to $\sum_{k \in C_i} n_k \cdot h_k = 0$. The allocated minimum average trust, $\eta_i(m)$, is a lower bound on average trust that must be achieved by platform $i$. Let the filter implemented by platform $i$ be $\alpha_i$. Then, platform $i$ must ensure that $\eta_i - h_i(\alpha_i) \geq \eta_i(m)$. Recall from the information structure that a violation of this condition cannot be detected by the social planner since she does not have explicit knowledge of the function $h_i(\cdot)$. However, by Assumption 5, the output of $h_i(a_i)$ can be monitored by any other competing platform $l \in C_i$. Platform $l$ will then report any violation of $\eta_i - h_i(\alpha_i) \geq \eta_i(m)$ to ensure that platform $i$ implements the largest filter $a_i$ and maximizes the utility $u_l(m, a_k : k \in C_i)$. This prevents platforms from violating the constraint imposed by $\eta_i(m)$.

3) The social planner allocates a price $\pi^l_i := \sum_{k \in C_l \setminus k \neq l} \pi^l_k \cdot \pi^l_i \in \mathbb{R}_{>0}$, to each platform $i \in I$, corresponding to the allocated filter $\alpha_i(m)$ of every other competing platform $l \in C_i \setminus i$. This is the average of prices proposed for the allocated filter $\alpha_i(m)$ by all competing platforms in $C_i \setminus i$. This way, the allocated price $\pi^l_i$ is independent of the prices proposed by both platforms $i$ and $l$. Similarly, the social planner allocates the price $\pi^0_i = \sum_{k \in D_i} \pi^0_k \cdot \pi^0_i$ to the government. We write the prices allocated to the players without the argument $m_i$ to simplify the notation.

4) The social planner allocates the following tax to each social media platform $i \in I$:

$$\tau_i(m) := -\tilde{p}_i \cdot \eta_i(m) - \sum_{l \in C_i} \pi^l_i \cdot \alpha_i(m) + \sum_{l \in C_i} \pi^l_i \cdot \alpha_l(m)$$

$$+ \sum_{l \in C_i} \tilde{p}_i \cdot (\tilde{a}_i^l - \tilde{a}_i^{l+1})^2$$

(7)

where $\tilde{a}_i^{l+1} = \sum_{k \in C_i \setminus k \neq l} \tilde{a}_k^{l+1}$, for each $l \in C_i$, is the average of the proposed filters for $l$ by all competing platforms except $i \in I$, and $\tilde{a}_i^{l+1} = \sum_{k \in D_i} \tilde{a}_k^{l+1}$ is the average of lower bounds proposed by all players except $i$. The tax $\tau_i(m)$ of platform $i$ in (7) can be interpreted as follows: 1) the first term is a subsidy given by the government to platform $i$ for the increase in average trust among the users of $i$; 2) the second term is a collection of subsidies given by each competing platform $l \in C_i \setminus i$ to platform $i$ for the increase in valuation $v_i(a_k : k \in C_i)$ due to the allocated filter $\alpha_l$; 3) the third term is a payment by platform $i$ for the increase in valuation $v_i(a_k : k \in C_i)$ due to the allocated filter $\alpha_l$ of each competing platform $l \in C_i \setminus i$; and 4) the fourth term is a collection of penalties to platform $i$ if either the filter proposed in message $m_i$ for any competing platform $l \in C_i \setminus i$ is inconsistent with the filters proposed by other platforms or the proposed lower bound is inconsistent with that proposed by other players. The fourth term also penalizes platform $i$ for higher values of $\tilde{p}_i$ and, thus, incentivizes proposing lower prices for other players.

The social planner also allocates an investment to the government as $\tau_0(m) := \pi_0(m) + (\tilde{p}_0 - \pi_0)^2$, where the first term is the total investment for the allocated lower bound $\alpha_0(m)$, and the second term is a penalty for any deviation between the proposed price $\tilde{p}_0$ and the allocated price $\pi_0$.

Remark 4: For some filter $a_i > 0$ of any platform $i$ in (7), the social planner takes a payment from each competing platform $l \in C_i \setminus i$ and allocates an equal subsidy to $i$. This subsidy incentivizes platform $i$ to implement the filter $a_i$ and helps to fairly distribute the government’s investment.

Remark 5: In the bargaining step, we have used all platforms in $I$ when defining the allocations, for, e.g., $\tau_0$. However, this does not cause any issues due to nonparticipating platforms because, as we prove in Theorem 4, all platforms eventually participate in the mechanism in the participation step.

Step 2 of the mechanism is characterized by the tuple $(\mathcal{M}, g(\cdot))$, where $\mathcal{M} = \mathcal{M}_0 \times \cdots \times \mathcal{M}_J$ is the message space, and $g(\cdot) : \mathcal{M} \rightarrow \mathcal{O}$ maps it to a set of outcomes $\mathcal{O} := \{\alpha_0(m), \ldots, \alpha_J(m)\}, \tau_0(m), \tau_j(m) = \alpha_i(m) \in \mathcal{A}, \tau_i(m) \in \mathcal{R}, i \in J$. The mechanism $(\mathcal{M}, g(\cdot))$ together with the utility functions $(u_i : i \in J)$ induces a game, in which the social planner allocates the filters $(\alpha_1(m), \ldots, \alpha_J(m))$ to the platforms and the lower bound $\alpha_0(m)$ to the government. Each platform $i \in I$ that participates in the mechanism must implement the filter $\alpha_i - \alpha_i(m)$, and the government must select the lower bound $\alpha_0 = \alpha_0(m)$. Platform $i$ can influence their allocated filter $\alpha_i(m)$ with their message $m_i$. Thus, the strategy of platform $i$ is given by the message $m_i \in \mathcal{M}_i$, with the constraint $\alpha_i(m) \in \mathcal{S}_i(m)$, where $\mathcal{S}_i(m) = \{a_i = a_i : h_i(a_i) \geq \eta_i(m)\}$. The set of feasible allocations $\mathcal{S}_i(m)$ for $i \in I$ is a function of the messages of all players. The government’s strategy is the message $m_0 \in \mathcal{M}_0$. For such a game, we select the solution concept of the GNE [20]. Let $m = (m_0, m_1, m_2, \ldots, m_J)$. A message profile $m^*$ is a GNE of the induced game, if

$$u_i((m^*_i, m^*_{-i}), \alpha_i(m^*_i, m^*_{-i})) : k \in C_i) \geq u_i((m_i, m^*_{-i}), \alpha_i(m_i, m^*_{-i})) : k \in C_i)$$

(8)

for all $m_i \in \mathcal{M}_i$ and $\alpha_i \in \mathcal{S}_i(m)$, for all $i \in I$, and the message $m^*_i$ of the government satisfies $u_0((m^*_0, m^*_{-0}), \alpha_0(m^*_0, m^*_{-0})) \geq u_0((m_0, m^*_0), \alpha_0(m_0, m^*_0)), \text{for all } m_0 \in \mathcal{M}_0$. In the rest of this article, we denote the utility of player $i \in J$ by $u_i(m_i, m_{-i})$.

Remark 6: In general, the GNE solution concept is defined for a game with complete information. However, we adopt this solution in our
induced game despite the fact that the valuation function $v_i(\alpha_k : k \in C_i)$ and the average trust function $h_i(\alpha)$ are the private information of platform $i$. We resolve this discrepancy by considering that the induced game is repeatedly over multiple iterations, and thus, the players can iteratively converge to a GNE [12]–[15].

Remark 7: We have summarized our notation in Table I.

### IV. PROPERTIES OF THE MECHANISM

In this section, we establish the properties of our mechanism. Recall that each social media platform $i \in I$ is a strategic player who seeks to maximize $u_i(m_i, m_{-i})$ through the choice of $m_i \in M_i$. Thus, we can define the following optimization problem for platform $i \in I$ in the induced game.

**Problem 2:** Platform $i$’s optimization problem is

$$\max_{m_i \in M_i} v_i(\alpha_k (m) : k \in C_i) - \tau_i(m),$$

subject to $0 \leq \alpha_k (m) \leq 1$

where (9) is the utility $u_i(m_i, m_{-i})$ of platform $i$, (10) ensures the feasibility of the allocated filter $\alpha_k (m)$, and (11) ensures that the allocated minimum average trust is achieved.

Note that the social planner can ensure that (10) and (11) are hard constraints by imposing a tax $\tau_i(m) \rightarrow \infty$ when they are violated. Also recall that the government strategically selects message $m_0 \in M_0$ to maximize their utility $u_0(m_0, m_{-0})$.

**Problem 3:** The government’s optimization problem is

$$\max_{m_0 \in M_0} v_0(\alpha_k (m) - \tau_0(m))$$

subject to $0 \leq \alpha_k (m) \leq 1$

$$\tau_0(\alpha_k (m)) - \delta_0 \leq 0$$

where the objective in (12) is the utility $u_0(m_0, m_{-0})$, (13) ensures that the government’s lower bound $\delta_0$ is feasible, and (14) is the budgetary constraint on total investment.

Remark 8: Consider optimal solutions $m^*_i \in M_i$ of Problem 2 for each platform $i \in I$, and $m^*_0 \in M_0$ of Problem 3 for the government. Then, the profile $m^* = (m^*_0, \ldots, m^*_n)$ satisfies (8) and, thus, is a GNE of the induced game.

Next, we establish some basic properties of the mechanism in Lemmas 1 and 2 at any GNE. We prove later in Theorem 3 that a GNE for the induced game always exists. In Lemma 1, we show that the government’s proposed price at any GNE is equal to the average price proposed by all platforms.

**Lemma 1:** Let the message profile $m^* \in M$ be a GNE of the induced game. Then, $p_i^*_0 = \pi^*_i$ for the government.

**Proof:** We note that (12) is concave with respect to $\bar{p}_0$. At GNE, we have $\frac{\partial \bar{p}_0}{\partial p_0} |_{p_0} = 2 \cdot (\bar{p}_0 - \pi^*_0) = 0$; thus, $\bar{p}_0 = \pi^*_0$.

Similarly, in Lemma 2, we show that, at any GNE, the filters proposed by all social media platforms in $C_i$ for platform $i$ are equal, and the lower bound proposed by all platforms is the same, unless the corresponding price proposal is 0.

**Lemma 2:** Let the message profile $m^* \in M$ be a GNE of the induced game. Then, for $\tilde{p}_k \neq 0$, we have $\tilde{\alpha}_k^* = \tilde{\alpha}_k^*$ for each social media platform $i \in I$, for each $k \in D_{-i}$.

**Proof:** The proof is similar to Lemma 1.

Next, we show that our proposed mechanism is budget balanced at any GNE, i.e., the social planner redistributes all the payments it collects from the players as subsidies to them.

**Theorem 1 (Budget Balance):** Consider any GNE $m^* \in M$ of the induced game. The proposed mechanism is budget balanced at GNE, i.e., $\sum_{i \in I} \tau_i(m^*) = 0$.

**Proof:** From Lemmas 1 and 2, the tax $\tau_i^*(m^*)$ for social media platform $i$ at GNE is $\tau_i^* = -\tilde{p}_i^* \cdot h_i(\alpha_i (m^*)) - \sum_{k \in D_{-i}} \pi_i^* \cdot \alpha_i (m) + \sum_{k \in C_i} \pi_i^* \cdot \alpha_i (m^*)$. The tax $\tau_i^*$ for the government at GNE is $\tau_i^* = \tilde{p}_i^* - \alpha_i (m^*)$, where $\tilde{p}_i^*$ is the price per unit change on average trust at GNE. Since $\sum_{k \in C_i} h_i(\alpha_i (m)) = \alpha_i (m)$, then, at GNE, we have $\sum_{i \in I} \tau_i^* = \sum_{i \in C_i} [\sum_{k \in C_i} \pi_i^* \cdot \alpha_i (m) + \sum_{k \in C_i} \pi_i^* \cdot \alpha_i (m^*)] = 0$.

In Lemma 3, we establish that every GNE of the induced game leads to an allocation of a filter profile and a lower bound such that all constraints of Problem 1 are satisfied.

**Lemma 3 (Feasibility):** Every GNE message profile $m^* \in M$ leads to a filter profile $(\alpha_1 (m^*), \ldots, \alpha_n (m^*))$ and lower bound $\alpha_0 (m^*)$, which is a feasible solution of Problem 1.

**Proof:** Every GNE message profile $m^*$ satisfies (10), (11), (13) and (14). From Theorem 1, $\sum_{i \in I} \tau_i^*(m^*) = 0$. For each $i \in I$, $h_i(\alpha_i (m)) \leq \alpha_0 (m)$, and $\sum_{i \in I} h_i(\alpha_i (m)) = \alpha_0 (m)$. Hence, $\sum_{i \in I} \tau_i^*(m^*) = 0$.

Next, we establish that each platform $i \in I$ can unilaterally deviate from the message $m_i \in M_i$ to achieve any desired allocation of filter profile. This property ensures that platform $i$ can always attain any filter $\tilde{\alpha}_i \in A$ for themselves.

**Lemma 4:** Given the message profile $m_{-i} \in M_{-i}$, the social media platform $i \in I$ can unilaterally deviate in their message $m_i \in M_i$ to attain any filter $\tilde{\alpha}_i \in A$ as the allocated filter $\alpha_k (m_i) \in S_k (m_i)$, for all $k \in C_i$.

**Proof:** Let $m_{-i}$ be the message profile of all players in $I_{-i}$. Then, platform $i$ can propose a filter $\tilde{\alpha}_k^* = \tilde{\alpha}_k^* - \sum_{i \in A_k \cup \cdot} \tilde{\alpha}_k^*$, to ensure that $\alpha_k (m_i) = \tilde{\alpha}_k^*$ for each $k \in C_i$. Moreover, platform $i$ can propose a lower bound $\tilde{\alpha}_k^* = -\sum_{i \in A_k \cup \cdot} \tilde{\alpha}_k^*$ for the government, to ensure that $\alpha_0 (m) = 0$ and, subsequently, $\alpha_k (m) = \tilde{\alpha}_k^*$ for all $k \in C_i$.

Next, we show that, at any GNE, the allocated filters for all platforms and the allocated lower bound for the government result in the optimal solution of Problem 1.

**Theorem 2 (Strong Implementation):** Consider any GNE $m^* \in M$ of the induced game. The allocated filter profile

### TABLE I

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_i$</td>
<td>The message broadcast by player $i \in J$</td>
</tr>
<tr>
<td>$\alpha_i$</td>
<td>The filter of platform $i \in I$</td>
</tr>
<tr>
<td>$\alpha_k^*$</td>
<td>The filter proposed by platform $i \in I$ for platform $k \in C_i$</td>
</tr>
<tr>
<td>$\alpha_i (m)$</td>
<td>The filter allocated to platform $i \in I$</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>The government’s lower bound on trust</td>
</tr>
<tr>
<td>$\alpha_0 (m)$</td>
<td>The lower bound proposed by the government</td>
</tr>
<tr>
<td>$\alpha_0 (m)$</td>
<td>The lower bound allocated to the government</td>
</tr>
<tr>
<td>$v_i(\cdot)$</td>
<td>The valuation function of player $i \in J$</td>
</tr>
<tr>
<td>$h_i(\cdot)$</td>
<td>The average trust function of platform $i \in I$</td>
</tr>
<tr>
<td>$\eta_i(m)$</td>
<td>The minimum average trust of platform $i \in I$</td>
</tr>
<tr>
<td>$\pi_i^<em>(m^</em>)$</td>
<td>The allocated minimum average trust for platform $i \in I$</td>
</tr>
<tr>
<td>$p_i^*$</td>
<td>The price proposed by platform $i \in I$ for player $l \in D_{-i}$</td>
</tr>
<tr>
<td>$\tau_i^<em>(m^</em>)$</td>
<td>The price allocated to platform $i \in I$ for player $l \in D_{-i}$</td>
</tr>
<tr>
<td>$\bar{p}_0$</td>
<td>The price proposed by the government</td>
</tr>
<tr>
<td>$\bar{\pi}_0$</td>
<td>The price allocated to the government</td>
</tr>
<tr>
<td>$\tau_i^<em>(m^</em>)$</td>
<td>The tax allocated to player $i \in J$</td>
</tr>
</tbody>
</table>
\[ (\alpha_1(m^*), \ldots, \alpha_\ell(m^*)) \] and lower bound \( \alpha_0(m^*) \) at equilibrium are equal to the optimal solution \( \alpha^* \) of Problem 1.

**Proof:** Let \( \alpha(m^*) = (\alpha_1(m^*), \ldots, \alpha_\ell(m^*)) \). Then, the GNE message profile \( m^* \) satisfies, for platform \( i \in I \), the following Kush–Kahn–Tucker (KKT) conditions for optimality:

1. \( \frac{\partial v_i}{\partial \alpha_i} \big|_{\alpha(m^*)} + \sum_{l \in C_i} \pi^*_l - \lambda^*_l + \mu^*_l + \nu^*_i = 0 \);
2. \( \frac{\partial h_i}{\partial \alpha_i} \big|_{\alpha(m^*)} - \pi^*_i = 0 \), for all \( l \in C_i \);
3. \( \lambda^*_l = 0; \)
4. \( \mu^*_i; \)
5. \( \nu^*_i; \)
6. \( \pi^*_i \).

Consider the optimal solution \( \lambda^*_M \in D \) of Problem 1, for all \( i \in I \). Let the valuation function \( v_i \) be \( \ell \)-concave and differentiable with respect to \( \alpha_i(m^*) \). Then, every social media platform \( i \) acts in an individually rational manner, \( u_i(m^*) = v_i(m^*), \) where user engagement is a primary driver of advertisement revenue.

Next, we show that our mechanism guarantees the existence of at least one GNE for the induced game. This ensures that the results in this section are always valid for the mechanism.

**Theorem 3 (GNE existence):** Let \( a^* = (a^*_0, a^*_1, \ldots, a^*_\ell) \) be the optimal solution of Problem 1. There is a GNE message profile \( m^* \in M \) of the induced game that guarantees the filter profile \( (\alpha_1(m^*), \ldots, \alpha_\ell(m^*)) \) and lower bound \( \alpha_0(m^*) \) at GNE satisfy \( \alpha_i(m^*) = a^*_i \), for all \( i \in I \).

**Proof:** Consider the optimal solution \( a^* \), which satisfies the KKT conditions for Problem 1 with the corresponding Lagrange multipliers \( (\lambda^*_1, \mu^*_1, \nu^*_1, \omega^*_1) \). The optimal solution \( \alpha^* = (a^*_0, a^*_1, \ldots, a^*_\ell) \) satisfies the KKT conditions of Problem 1 with the Lagrange multipliers \( (\lambda^*_1, \mu^*_1, \nu^*_1, \omega^*_1, \ldots, \lambda^*_i, \mu^*_i, \nu^*_i, \omega^*_i) = (\alpha_i(m^*)) \). Thus, the efficient allocation of filters for all platforms and lower bound for the government is implemented by all GNEs.

Next, we consider the step 2 of the mechanism to show that in step 1, every social media platform \( i \) is quasi-concave and differentiable for all \( \mu_i \in \mathbb{R}^+ \), and (iii), \( \sum_{l \in C_i} \frac{\partial v_i}{\partial \alpha_i} \big|_{\alpha(m^*)} - \lambda^*_l + \mu^*_l + \nu^*_i = 0 \), for all \( l \in C_i \). Similarly, we can write the KKT conditions for Problem 3 with the Lagrange multipliers \( (\lambda^*_l, \mu^*_l, \nu^*_l, \omega^*_l) \). The optimal solution \( \alpha^* = (a^*_0, a^*_1, \ldots, a^*_\ell) \) satisfies the KKT conditions of Problem 1 with the Lagrange multipliers \( (\lambda^*_1, \mu^*_1, \nu^*_1, \omega^*_1, \ldots, \lambda^*_i, \mu^*_i, \nu^*_i, \omega^*_i) = (\alpha_i(m^*)) \). Thus, the efficient allocation of filters for all platforms and lower bound for the government is implemented by all GNEs.

Next, we consider the participation step from Section III-A. The optimal solution \( \alpha^* = (a^*_0, a^*_1, \ldots, a^*_\ell) \) of each platform \( i \in I \) and the government, respectively, are quasi-concave and differentiable; and 2) the average trust function \( h_i(a_i) \) for all \( i \in I \) is differentiable. Thus, we cannot use the KKT conditions to prove the existence of an induced GNE and strong implementation. However, if a GNE exists, the proposed mechanism is still budget balanced, individually rational, and Lemmas 1–4 hold in Theorem 5, we show that there still exists a GNE, and it induces a Pareto efficient equilibrium in the game. Pareto efficiency is the condition where we cannot improve the utility of any player without decreasing the utility of another player [12]. This is weaker than Theorem 2.

**Theorem 5:** Let the valuation function \( v_i(a_i : k \in C_i) \) be quasi-concave and differentiable for all \( i \in J \) in the game \( (M, g(\cdot), (u_i)_{i \in I}) \). Then, there exists an efficient allocation of filters for all platforms and lower bound for the government is implemented by all GNEs.

**Proof 1) Existence:** By Lemma 2, \( M := \{m_i \in M_i : \tilde{p}_i \cdot (\tilde{a}_i - a^*_i) = 0, \forall i \in D_i \} \) at GNE. For all \( m_i \in M_i \), \( u_i(a_i) = v_i(a_i) : k \in C_i + \tilde{p}_i \cdot h_i(a_i) + \sum_{l \in C_i} \lambda^*_l \cdot a_i^l \). Thus, platform \( i \)'s action is to select \( \alpha_i(m^*) = a^*_i \), which is convex, compact, and independent of the profile message \( m^* \), where \( \alpha_i \in \mathbb{R}^+ \), and (iii), \( \sum_{l \in C_i} \lambda^*_l \cdot a_i^l \). Thus, the efficient allocation of filters for all platforms and lower bound for the government is implemented by all GNEs.

Next, we consider the participation step from Section III-A. The optimal solution \( \alpha^* = (a^*_0, a^*_1, \ldots, a^*_\ell) \) of each platform \( i \in I \) and the government, respectively, are quasi-concave and differentiable for all \( \mu_i \in \mathbb{R}^+ \), and (iii), \( \sum_{l \in C_i} \frac{\partial v_i}{\partial \alpha_i} \big|_{\alpha(m^*)} - \lambda^*_l + \mu^*_l + \nu^*_i = 0 \), for all \( l \in C_i \). Similarly, we can write the KKT conditions for Problem 3 with the Lagrange multipliers \( (\lambda^*_l, \mu^*_l, \nu^*_l, \omega^*_l) \). The optimal solution \( \alpha^* = (a^*_0, a^*_1, \ldots, a^*_\ell) \) satisfies the KKT conditions of Problem 1 with the Lagrange multipliers \( (\lambda^*_1, \mu^*_1, \nu^*_1, \omega^*_1, \ldots, \lambda^*_i, \mu^*_i, \nu^*_i, \omega^*_i) = (\alpha_i(m^*)) \). Thus, the efficient allocation of filters for all platforms and lower bound for the government is implemented by all GNEs.

2) **Pareto Efficiency:** It is sufficient to show that the NE can be characterized by a Walrasian equilibrium, thus Pareto efficient. Consider any NE action profile \( \beta^* = (\beta^*_1, \ldots, \beta^*_\ell) \in A \times B_1 \times \cdots \times B_{\ell+1} \). By the NE definition, for every platform \( i \in I \), it holds that \( u_i(\beta^*) \geq u_i(\beta), \) for all \( \beta \in B_i \). Then, we have \( \beta^*_i = \arg \max_{\beta \in B_i} (v_i(a_i : k \in C_i) + \tilde{p}_i \cdot h_i(a_i) - \sum_{l \in C_i} \lambda^*_l \cdot a_i^l) \), for all \( i \in I \). Similarly, for the government’s utility \( u_0(\alpha_0) \) is quasi-concave and differentiable with respect to \( \alpha_0 \). It follows from Glicksberg’s theorem that there exists a GNE for the induced game.

**Remark 9:** With quasi-concave valuations, the induced GNE may not lead to allocations that are optimal for Problem 1. However, Theorem 5 shows that there still exists a GNE and that it is a Pareto efficient allocation, and thus, our mechanism incentivizes filtering with suboptimal social welfare.

**Example**

In this subsection, we present an example of how our proposed mechanism may be executed. Consider three major social media platforms: Facebook, Twitter, and Reddit. These platforms allow users from different political backgrounds to obtain the latest news. Typically, users engage with these platforms by scrolling, liking, or sharing posts featuring news and personal opinions. The time spent by all users on a platform defines the total engagement in the platform [5].

Since user engagement is a primary driver of advertisement revenue, Facebook, Twitter, and Reddit regularly optimize their post recommendation algorithms to maximize user engagement without accounting for...
the impact on user opinions [3]. Thus, many users form echo chambers, where they repeatedly interact only with biased posts. The biases of many users expose them to misinformation. This might cause uncertainty regarding the integrity of democratic institutions [7] or the results of the elections [4]. In practice, each platform can filter misinformation by flagging inaccurate posts. However, filtering is expensive because of 1) the large cost to identify inaccurate posts [21], and 2) the potential decrease in engagement of censored users [9]. Thus, the government allocates a budget for the problem and appoints an independent agency to design monetary incentives for the platforms. This agency seeks a mechanism that (i) induces voluntary participation among all platforms and (ii) maximizes the social welfare. Such a mechanism incentivizes platforms to implement filters. A sufficiently high lower bound ensures that some platforms implement nonzero filters. Using the mechanism in Section III, the agency must achieve (i) and (ii) without knowledge of how the engagement and average trust on common knowledge evolve.

In step 1 (the participation step), the agency asks each platform whether they wish to participate in the mechanism since the government is not dictatorial. However, the agency assures the three platforms that they need not reveal private information, and that they can avoid filtering misinformation even after participating in the mechanism (see Lemma 4). The government announces that platforms that choose not to participate will be labeled as uncooperative. Then, the weakly dominant action of every platform in step 1 is to participate in the mechanism (see Theorem 4), ensuring property (i).

In step 2 (the bargaining step) of the mechanism, the agency asks each platform for a message proposing filtering levels for competing agents and corresponding prices, a lower bound for the government, and a minimum level for the average trust of their own users. Similarly, the government also proposes a lower bound and a price associated with the lower bound. The agency then publicly reveals all proposals and uses the rules of the mechanism to assign a potential subsidy/payment and potential filtering level to each platform. Similarly, the agency assigns a potential amount of investment and minimum average to the government. Note that the subsidy given to any platform is proportional to their assigned minimum average trust and filtering level.

This process is repeated until all stakeholders reach a consensus, known as a GNE. The mechanism ensures that a consensus exists (see Theorem 3) and that it maximizes the social welfare among all stakeholders (see Theorem 2), thus establishing property (i). Furthermore, the mechanism also ensures that each stakeholder is consistent in their messages with respect to the messages of other stakeholders (see Lemmas 1 and 2) and that the agency makes neither profit nor loss (see Theorem 1). Therefore, as long as the government is committed to addressing the problem of misinformation, the mechanism ensures that the platforms will eventually agree to implement filters. The allocations become binding on all stakeholders, and the agency collects the government’s investment. This investment is distributed to Facebook, Twitter, and Reddit as a subsidy, only after they implement the assigned filters.

V. CONCLUSION AND FUTURE WORK

In this article, we designed a mechanism to induce a GNE solution in the misinformation filtering game, where 1) each platform agrees to participate voluntarily, and 2) the collective utility of the government and the platforms is maximized. Our proposed mechanism also satisfies budget balance. We also analyzed our mechanism under relaxed assumptions. Our ongoing work focuses on improving estimates of the valuation and average trust functions of the platforms using data and explicitly considering modeling uncertainty. These refinements of the modeling framework will allow us to make our mechanism more practically useful. Future research should include extending the mechanism to a dynamic setting, where the platforms react in real time to the proposed taxes/subsidies.

REFERENCES