**INTRODUCTION**

For the last 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, which often seek to maximize the *engagement* of their users at all costs. In fact, the algorithms developed by platforms for this purpose often promote conspiracy theories among their users [3].

The sensitivity of users of social media platforms to conspiratorial ideas makes them an ideal terrain to conduct political misinformation campaigns [4], [5]. Such campaigns are especially effective tools to disrupt democratic institutions, because the functioning of stable democracies relies on *com-* *mon knowledge* about the political actors and the processes they can use to gain public support [6]. The trust held by the citizens of a democracy on common knowledge includes: (i) trust that all political actors act in good faith when contesting for power, (ii) trust that elections lead to a free and fair transfer of power between the political actors, and (iii) trust that democratic institutions ensure that elected officials wield their power in the best interest of the citizens. In contrast, citizens of democracies often have a *contested knowledge* regarding who should hold power and how they should use it [6]. The introduction of *alternative facts* can reduce the trust on common knowledge about democracy, especially if they become accepted beliefs among the citizens. Such disruptions on the trust on common knowledge can be found in the 2016 U.S. elections [7] and Brexit Campaign in 2016 [8], where the spread of misinformation through social media platforms resulted in a large number of citizens mistrusting the results of voting. To tackle this growing phenomenon of misinformation, in this paper, we consider a finite group of social media platforms, whose users represent the citizens in a democracy, and a democratic government. Every post in the platforms is associated with a parameter that captures its informativeness , which can take values between two extremes: (i) completely factual and (ii) complete misinformation. In our framework, posts that exhibit misinformation can lead to a decrease in trust on common knowledge among the users [9]–[12]. In addition, social media platforms are considered to have the technologies to *filter*, or label, posts that intend to sacrifice trust on common knowledge. Thus, the government seeks to incentivize the social media platforms to use these technologies and filter any misinformation included in the posts.

Motivated by capitalistic values, we induce a *misinformation* *filtering game* to describe the interactions between the social media platforms and the government. In this game, each platform acts as strategic player seeking to maximize their advertisement revenue from the engagement of their users [7], [13]. User engagement is a metric that can be used to quantify the interaction of users with a platform, and subsequently, how much time they spend on the platform. Recent efforts reported in the literature on misinformation in social media platforms have indicated that increasing filtering of misinformation leads to decreasing of user engagement [14]. There are many possible reasons for this phenomenon. First, filtering reduces the total number of posts propagating across the social network. Second, the users whose opinions are filtered may perceive this action as dictatorial censorship [15], and as a result, they may chose to express their opinions in other platforms. Finally, misinformation tends to elicit stronger reactions, e.g., surprise, joy, sadness, as compared to factual posts [16], which may increase user engagement. Thus, each platform is reluctant to filter misinformation.

In our framework, we consider that the government is also a strategic player, whose utility increases as the trust of the users of social media platforms on common knowledge increases. Consequently, increasing filtering of misinformation by the social media platforms increases the utility of the government. Thus the government is willing to make an investment to incentivize the social media platforms to filter misinformation. In our approach, we use mechanism design to distribute this investment among the platforms optimally, and in return, implement an optimal level of filtering.

Mechanism design was developed for the implementation of system-wide optimal solutions to problems involving multiple rational players with conflicting interests, each with private information about preferences [17]. Note that this approach is different from traditional approaches to decentralized control with private information [18]–[21] because the players are not a part of the same time, but in fact, have private and competitive utilities. The fact that Mechanism design optimizes the behaviour of competing players has led to broad applications spanning different fields including economics, politics, wireless networks, social networks, internet advertising, spectrum and bandwidth trading, logistics, supply chain, management, grid computing, and resource allocation problems in decentralized systems [22]–[28].

The contribution of this paper is as follows. We present an indirect mechanism to incentivize social media platforms to filter misleading information. We show that our proposed mechanism is (i) feasible, (ii) budget balanced, (iii) individual rational, and (iv) strongly implementable at the equilibria of the induced game. We prove the existence of at least one generalized Nash equilibrium and show that our mechanism induces a Pareto efficient equilibrium.

The rest of the paper is organized as follows. In Section II, we provide the modeling framework and problem formulation. In Section III, we present our mechanism, and in Section IV, we prove the associated properties of the mechanism. In Section V, we interpret the mechanism and present a descriptive example. Finally, in Section VI we conclude and present some directions for future research.