OPTIMAL ALLOCATION OF FACT-CHECKING RESOURCES ON LONG-TERM PREVALENCE OF FALSE NARRATIVES

Mustafa Alam*

Last Updated: July 2025

Abstract

I use a stylized compartmental model to analyze the long-term dynamics of misinformation propagation in social networks, focusing on the allocation of fact-checking resources. I conceptualize a false narrative as spreading through multiple types of claims, which can differ in their virality and resistance to fact-checking interventions. The analysis reveals that harder-to-debunk claims can persist when fact-checkers concentrate on easy-to-debunk claims—an approach commonly arising from crowd-sourced, consensus-based systems such as Community Notes—and ultimately become the primary vector sustaining the false narrative over time. I characterize the optimal allocation of fact-checking effort and show that, given sufficient resources, effective long-term mitigation of misinformation requires devoting resources to both easy and hard-to-debunk claims, no matter their initial virality or perceived cost. These findings challenge the prevailing focus on short-term fact-checking "successes" and underscore the need to supplement crowd-sourced interventions with targeted professional fact-checking of complex or resilient misinformation. The theoretical framework provides actionable guidance for platforms and policymakers seeking to minimize the long-run societal impact of persistent false narratives.

*Department of Economics, 312J Wilbur O. and Ann Powers College of Business, Clemson University, Clemson, South Carolina 29634-1309, USA; email: mustafa@g.clemson.edu.

1 Introduction

Misinformation poses a growing threat to society and its important institutions. The widespread dissemination of false or misleading content erodes trust in credible media sources (Ognyanova et al., 2020), distorts political beliefs and democratic processes (Ecker et al., 2024; Allcott and Gentzkow, 2017), and undermines the effectiveness of public health interventions (Loomba et al., 2021; Pierri et al., 2022). As social media platforms become an increasingly important source of news and information for a large fraction of the global population, the ease with which content can be produced and consumed can magnify both the prevalence and the impact of misinformation. Therefore, understanding how to mitigate the societal risks these false narratives pose effectively is of first-order importance.

A large and diverse body of research has explored who is susceptible to misinformation[1], how certain types of false or misleading claims spread[2], and the effectiveness of various mitigation interventions.[3] A related strand of literature highlights limitations of fact-checking interventions, focusing on the timing of fact-checking messages (Brashier et al., 2021), the relative reach of corrections compared to misinformation (Chido-Amajuoyi et al., 2019), and the "continued influence effect," whereby people persist in using corrected misinformation in their reasoning.[4] There have also been concerns about a potential "backfire" effect, in which corrections strengthen belief in false information (Nyhan et al., 2014; Nyhan and Reifler, 2015)—though recent studies generally fail to confirm this phenomenon (Swire-Thompson et al., 2020; Wood and Porter, 2019). Nonetheless, most existing research focuses on individual claim types, relatively short-run behavioral measures, and individual reasoning about misinformation before or after exposure to corrections.

However, little work addresses how competing misinformation claims supporting a broader narrative can persist in the *long-run* and how fact-checkers or platforms should allocate limited resources to undermine it. For instance, individuals in a population may be susceptible to evolving or different claims supporting the same narrative even though they witnessed corrections on similar claims in the past if the false narrative continues to persist in their social network. Furthermore, prior literature rarely considers how targeting one misinformation claim may affect the spread or endurance of other related claims or how fact-checking strategies and incentive structures shape the overall prevalence of a broader narrative. As major platforms move away from professional fact-checking organizations and increasingly rely on crowd-sourced, consensus-based approaches, understanding the long-term implications of these resource allocation choices is both urgent and

[1] See, e.g., Dechêne et al. (2010); Wang et al. (2016); Guess et al. (2019); Saunders and Jess (2010); Imhoff et al. (2022); Pennycook and Rand (2019); Bronstein et al. (2019).

[2] See, e.g., Cinelli et al. (2020); Vosoughi et al. (2018); Juul and Ugander (2021).

[3] See, e.g., Chan et al. (2017); Yousuf et al. (2021); Smith et al. (2011); Paynter et al. (2019); Barbera et al. (2024).

[4] See Lewandowsky et al. (2012); Walter and Tukachinsky (2020); Desai et al. (2020).

unresolved.

This paper presents a simple theoretical exposition to illustrate how differential fact-checking of distinct claim types within a broader false narrative affects its long-term discussion in a social network. By modeling the joint spread of multiple, heterogeneous claims, the analysis reveals that the choice of which claims to target can significantly alter the long-run persistence of misinformation. Fact-checking systems that lead to the prioritization of easily debunked claims or maximization of short-run reductions in the prevalence of individual claims may be sub-optimal and, under plausible conditions, even counterproductive in reducing the long-term discussion of false narratives. Assuming that persistent discussion of false narratives is linked to undesirable actions and broader societal harms, this work underscores the importance of fact-checking systems that explicitly account for long-term transmission dynamics and the interactions among different types of claims.

I model information spread through a social network using an epidemiology-style compartmental framework. A well-mixed population of individuals may be susceptible to sharing misinformation, actively propagating distinct claims, or temporarily disengaging from the narrative. Each claim type differs in transmissibility and resistance to fact-checking interventions, and fact-checking resources are allocated under a fixed constraint. This approach complements existing empirical work[5] that draws analogies between misinformation and viral contagion with emphasis on the role of the reproduction number—which measures the number of people who start sharing false content following contact with someone already sharing that content, in a fully susceptible population. The analysis in this paper probes deeper into this analogy by considering a multi-strain transmission model.[6] The central focus is to determine how interventions targeting a specific claim type in the presence of multiple viral claims influence both its own long-term prevalence and the persistence of related claims within the same narrative, as well as to elucidate the implications of fact-checking incentive structures for these long-term outcomes.

The results demonstrate that fact-checking systems that prioritize interventions on the easiest-to-debunk claims, or those that yield the largest immediate reductions in visibility, may be insufficient to minimize the long-term discussion of the overarching narrative. For example, if the harder-to-debunk claims are also more viral, devoting resources to fact-checking easy claims may reduce transmission in the short run but will not affect the long-run persistence of the false narrative since the harder-to-debunk claims would be the vector with which the false narrative would propagate in the long run even without the intervention. Conversely, if easy-to-debunk claims are more viral, concentrating significant resources on suppressing them can facilitate the eventual dominance of the more persistent, harder-to-debunk claims. The central insight is that fact-checking interven-

[5] See, e.g., Kucharski (2016); Cinelli et al. (2020); Scales et al. (2021).

[6] See Andreasen et al. (1997) for a classic treatment.

tions do not operate in isolation: targeting one set of claims affects the competitive landscape for others. From a transmission dynamics perspective, removing easy-to-fact-check claims can create space for more resistant claims to spread with less competition, magnifying their long-run reach. From a resource allocation perspective, devoting more than the optimal amount of resources to easier claims leads to a waste of resources, leaving fewer resources available to suppress the claims most resistant to correction and potentially more consequential for societal harm.

These findings are particularly salient given recent shifts in platform moderation practices. The move toward consensus-based, crowd-sourced fact-checking systems, such as Community Notes, introduces an inherent bias toward identifying and intervening in claims broadly recognizable as false or uncontroversial. As a result, more sophisticated, ambiguous, or divisive misinformation often escapes annotation—even when these claims may disproportionately impact long-term beliefs and behaviors. The analysis suggests that metrics focused solely on the number of claims checked or short-term behavioral change risk reinforcing this bias. To reduce the long-term societal impact of false narratives more effectively, platforms and fact-checkers should design mitigation strategies that complement the incentive structures of consensus-based systems, for example, by dedicating targeted professional fact-checking efforts to claims unlikely to be addressed by crowd-sourcing alone.

2 Model

The structure of the model is as follows. Section 2.1 sets up the environment. Section 2.2 illustrates the dynamics of narrative propagation. Section 2.3 characterizes the long-run equilibria. Section 2.4 presents an analysis of optimal fact-checking to minimize long-term costs to society.

2.1 Set-up

Consider a well-connected, homogeneous network of n individuals where a false or misleading narrative propagates from a small subset of the population. The narrative is propagated via two types of claims: type-1 and type-2. Individuals can be in one of five states at any time, t: 'Susceptible,' 'Infected with a type-1 claim', 'Infected with a type-2 claim', 'Recovered from a type-1 claim' and 'Recovered from a type-2 claim'. Susceptible individuals are open to adopting and sharing claims from individuals 'infected' with a claim of either type. Those in the infected states actively share and discuss these claims on social media. After a period of being infected, individuals transition to a 'Recovered from type-j claim' state ($j \in \{1,2\}$), where individuals take a break from discussing this narrative altogether. After spending some non-zero time in the 'Recovered from type-j claim' state, the individuals become 'Susceptible" again, open to sharing/discussing

new or evolving claims of either type. Let the number of individuals in the 'Susceptible' state at time t be denoted $X(t)$. The number of individuals in the 'Infected with a type-j claim' state is denoted $Y_j(t)$, and the number of individuals in the 'Recovered from type-j claim' state is denoted $Z_j(t)$.

An individual in the 'Susceptible' state has a mean of k contacts per unit time. Two distinct fractions of these contacts are infected by either type of claim. I denote this fraction as $I_j(t) = Y_j(t)/n$. Over a small time interval of δt, the expected number of contacts with people sharing the j-type claim is given by $\frac{kY_j(t)\delta t}{n}$. Let p_j be the probability of sharing a claim after interacting with an individual 'infected' with a j-type claim. By independence of contacts, the probability that an individual in the 'Susceptible' state escapes infection of or does not share a j-type claim following $\frac{kY_j(t)\delta t}{n}$ contacts is $(1-p_j)^{\frac{kY_j(t)\delta t}{n}}$. Let δq_j be the probability of sharing the type-j claim following any of the $\frac{kY_j(t)\delta t}{n}$ contacts. Then, the following identity holds

$$1-\delta q_j = (1-p_j)^{\frac{kY_j(t)\delta t}{n}}. \tag{1}$$

Let $\beta_j = -k\log(1-p_j)$ be the effective transmission rate of the j-type claim. Equation (1) can then be rewritten as

$$\delta q_j = 1 - e^{-\frac{\beta_j Y_j(t)\delta t}{n}}$$

$$\implies \delta q_j = \frac{\beta_j Y_j(t)\delta t}{n} - \frac{(\frac{\beta_j Y_j(t)\delta t}{n})^2}{2!} + \frac{(\frac{\beta_j Y_j(t)\delta t}{n})^3}{3!} - \ldots$$

$$\implies \frac{\delta q_j}{\delta t} = \frac{\beta_j Y_j(t)}{n} - \frac{(\frac{\beta_j Y_j(t)}{n})^2 \delta t}{2!} + \frac{(\frac{\beta_j Y_j(t)}{n})^3 (\delta t)^2}{3!} - \ldots \tag{2}$$

To determine the transmission rate per susceptible individual, one can take the limit,

$$\lim_{t\to 0} \frac{\delta q_j}{\delta t} = \frac{dq_j}{dt} = \frac{\beta_j Y_j(t)}{n}. \tag{3}$$

The rate at which susceptibles become infected (i.e., the outflow from the susceptible state due to infection) is

$$-X(t)\left[\frac{dq_1}{dt} + \frac{dq_2}{dt}\right] = -X(t)\left[\frac{\beta_1 Y_1(t)}{n} + \frac{\beta_2 Y_2(t)}{n}\right]. \tag{4}$$

Upon infection with a type-j claim, an individual remains in the infected state for an exponentially distributed duration with rate γ_j. Thus, the expected period of sharing or discussing the claim (the 'infectious period') is $1/\gamma_j$ without any fact-checking intervention. This can be interpreted as

natural disengagement from the narrative, possibly due to loss of interest, competing content, or claim evolution.

Fact-checking interventions are assumed to shorten this infectious period. The effectiveness of fact-checking on claim j depends on two factors: $\alpha_j > 0$, which reflects the resistance of the claim type to fact-checking, and $T_j > 0$, which captures the intensity, quality or quantity of fact-checking. Under intervention, T_j, the expected duration of sharing for type-j claims becomes $1/(\gamma_j + \alpha_j T_j)$.

When individuals stop sharing a type-j claim, they enter the 'Recovered from type-j claim' state. Individuals in this state are temporarily disengaged from the broader narrative and remain in this state for an exponentially distributed duration with rate ν_j. The expected duration of disengagement (the 'immunity period') is $1/\nu_j$. This may represent users moving on to discussing or sharing other narratives or using their time elsewhere. Note that interventions could also be modeled to *extend* this immunity period. If such a characterization is chosen, the key take-aways of the paper remain unchanged.

2.2 Narrative Propagation Dynamics

Let $S(t) = \frac{X(t)}{N}$, $I_j(t) = \frac{Y_j(t)}{N}$ and $R_j(t) = \frac{Z_j(t)}{N}$ be the fraction of individuals in each of the five states. By construction, $S(t) + I_1(t) + I_2(t) + R_1(t) + R_2(t) = 1, \forall t$. Taken together, the dynamics of narrative spread through both types of claims are as follows

$$\frac{dS}{dt} = -S[\beta_1 I_1 + \beta_2 I_2] + \nu_1 R_1 + \nu_2 R_2 \tag{5}$$

$$\frac{dI_1}{dt} = \beta_1 S I_1 - [\gamma_1 + \alpha_1 T_1] I_1 \tag{6}$$

$$\frac{dI_2}{dt} = \beta_2 S I_2 - [\gamma_2 + \alpha_2 T_2] I_2 \tag{7}$$

$$\frac{dR_1}{dt} = [\gamma_1 + \alpha_1 T_1] I_1 - \nu_1 R_1 \tag{8}$$

$$\frac{dR_2}{dt} = [\gamma_2 + \alpha_2 T_2] I_2 - \nu_2 R_2 \tag{9}$$

2.3 Long-run Equilibrium

The *basic reproduction number* of a type-j claim is given by

$$R_0^j = \frac{\beta_j}{\gamma_j + \alpha_j T_j}.$$

This quantity represents the expected number of individuals who adopt and begin sharing the claim due to a single individual's sharing activity in a fully susceptible population.
I impose the following restriction in the absence of fact-checking interventions

$$R_0^j\big|_{T_j=0} = \frac{\beta_j}{\gamma_j} > 1, \quad \text{for } j \in \{1,2\}. \tag{10}$$

This restriction focuses the analysis on scenarios where both claim types have sufficient characteristics to propagate on their own at the onset of spread. More precisely, it ensures that an individual actively sharing a type-j claim is expected to generate more than one additional adopter in a fully susceptible population. If instead $R_0^j < 1$, the claim would fail to generate sustained spread and eventually die out, even without intervention.

If fact-checking is highly effective—such that it drives $R_0^j < 1$ for one or both claims—then the corresponding claim(s) will never persist in the long run. In this case, a single individual fails to generate at least one additional adopter in a fully susceptible population, and the claim dies out as the susceptible population declines. While such an outcome may be desirable in practice, it is analytically uninteresting.

I instead focus on claims that are relatively difficult to fact-check by tightening the restriction in Equation (10) to ensure that reproduction numbers remain above one even in the presence of intervention,

$$R_0^j = \frac{\beta_j}{\gamma_j + \alpha_j T_j} > 1, \quad \text{for } j \in \{1,2\}. \tag{11}$$

This assumption ensures that even the maximum fact-checking intervention will not lead to the vanishing of both types of claim in the long run, although the fraction of the population discussing this claim can be arbitrarily near zero. This allows us to study how differential interventions influence long-term narrative dynamics.

Given the setting, the following is a well-known and relevant implication about long-term propagation in compartmental models.[7]

Remark. *If $R_0^j > R_0^{-j}$, the j-type claim is the only type of claim supporting the narrative that survives in the population. That is* $\lim_{t\to\infty} I_j > 0$ *and* $\lim_{t\to\infty} I_{-j} = 0$.

Proof sketch and intuition. *To develop an intuition for this result, consider the dynamics of each claim type. At any time t, if the effective reproduction number $R_0^j S(t) > 1$, then the fraction of individuals sharing the j-type claim will increase—each individual, on average, causes more than one additional person to adopt. Conversely, if $R_0^j S(t) < 1$, the infected fraction will decrease. Suppose the $(-j)$-type claim is in equilibrium. Then, by definition, the infected fraction is constant, which implies $R_0^{-j} S(t) = 1$. But since $R_0^j > R_0^{-j}$, it follows that $R_0^j S(t) > 1$, and thus the j-type*

[7] See van den Driessche and Watmough (2002)

claim must be growing. This means both claims cannot simultaneously be in equilibrium unless $R_0^j = R_0^{-j}$. Therefore, a positive steady state for both claim types is not possible. The fraction of the population discussing the $(-j)$-type claim can only be stable when it is zero. Hence, in the long run, the j-type claim with the higher reproduction number will dominate, and the other will vanish from the population.

For any pair (T_1, T_2) such that $R_0^1 > R_0^2$, type-1 claim is the only one that persists in the long run. In equilibrium, the fraction of the population in each of the five states is as follows

$$\left(S^* = \frac{\gamma_1 + \alpha_1 T_1}{\beta_1}, I_1^* = \frac{1 - \frac{\gamma_1 + \alpha_1 T_1}{\beta_1}}{1 + \frac{\gamma_1 + \alpha_1 T_1}{\nu_1}}, R_1^* = \frac{1 - \frac{\gamma_1 + \alpha_1 T_1}{\beta_1}}{1 + \frac{\nu_1}{\gamma_1 + \alpha_1 T_1}}, I_2^* = 0, R_2^* = 0\right). \tag{12}$$

Alternatively, if (T_1, T_2) is such that $R_0^1 < R_0^2$, the type-2 claim is the only one that persists in the long run. In equilibrium, the fraction of the population in each of the five states is then

$$\left(S^* = \frac{\gamma_2 + \alpha_2 T_2}{\beta_2}, I_1^* = 0, R_1^* = 0, I_2^* = \frac{1 - \frac{\gamma_2 + \alpha_2 T_2}{\beta_2}}{1 + \frac{\gamma_2 + \alpha_2 T_2}{\nu_2}}, R_2^* = \frac{1 - \frac{\gamma_2 + \alpha_2 T_2}{\beta_2}}{1 + \frac{\nu_2}{\gamma_2 + \alpha_2 T_2}}\right). \tag{13}$$

2.4 Effectiveness of Fact-Checking Interventions on the Long-term Discussion of the Narrative

In this section, I analyze the effect of allocating fact-checking resources on the two types of claims to reduce a notion of "cost to society" due to the propagation of the false narrative. To formalize this cost, consider the following cost function

$$\mathscr{C}(I_1^*, I_2^*) = C_1\left(I_1^*\left(T_1(r_1), T_2(r_2)\right)\right) + C_2\left(I_2^*\left(T_1(r_1), T_2(r_2)\right)\right), \tag{14}$$

where r_1 and r_2 are resources devoted to fact-check type-1 and type-2 claims respectively. I assume that the cost is increasing in the long-run fraction of the population discussing either type of claim, $C_j'(I_j^*) > 0$. One can interpret this cost function as the externalities generated through the actions taken by individuals in the population due to the long-term prevalence of the false narrative. It could also include the cost imposed on oneself due to actions based on misleading/false information.

Assume that a fact-checker or a fact-checking system has fixed resources, R, to reduce the prevalence of the false narrative in the long run. They can divide their resources to fact-check either type of claim by choosing (r_1, r_2). Devoting more resources to fact-checking type-j claim improves treatment ($T_j'(r_j) > 0$) and devoting zero resources to fact-checking leads to no intervention $T_j(0) = 0$. With no loss of generality, also assume

$$R_0^1 > R_0^2. \tag{15}$$

For a given T_2, the condition in (14) holds if

$$T_1 < \frac{\beta_1(\gamma_2 + \alpha_2 T_2) - \beta_2\gamma_1}{\alpha_1\beta_2}. \tag{16}$$

This threshold defines the region where increasing T_1 continues to reduce the long-run narrative prevalence I_1^*, since $\frac{\partial I_j^*}{\partial T_j} < 0$. However, once T_1 crosses this threshold, the basic reproduction number of the type-2 claim exceeds that of the type-1 claim ($R_0^2 > R_0^1$). Beyond that point, any resources devoted to combatting type-1 claims have no impact on the long-term discussion of the false narrative.

Consider the recent evolution in misinformation mitigation strategies on major social media platforms. Notably, X (formerly Twitter) has shifted from reliance on professional fact-checking organizations to crowd-sourced interventions through its Community Notes system (The Associated Press, 2022). Similarly, Meta has ended its third-party fact-checking partnerships in the U.S., adopting a user-driven Community Notes-type model to provide contextual information on posts (Kaplan, 2025). The central algorithmic feature of Community Notes is its "bridge-based" design, wherein user-submitted notes are surfaced when found helpful by raters with diverse rating histories (Wojcik et al., 2022). The intended benefit is clear: By requiring agreement from users with differing perspectives for fact-check annotation, the system offers a scalable approach that guards against unilateral or partisan manipulation, thus enhancing legitimacy and public trust.

However, this design introduces a structural bias toward identifying and fact-checking only those claims for which broad consensus is possible—typically, misinformation that is egregious, non-controversial, or already widely recognized as false. In practice, this means that more sophisticated, ideologically resonant, or ambiguous claims—precisely those most likely to shape long-term beliefs and behaviors—often escape annotation, as cross-group consensus on their veracity is harder to attain. The result is a systematic targeting of "easy-to-refute" narratives, while persistent, high-resistance claims remain largely unaddressed. This is of particular consequence given the diminishing role of institutional fact-checkers, as community-driven models are increasingly positioned as the primary—sometimes sole—fact-checking mechanism on large platforms.

Importantly, this structural issue is not unique to crowd-sourced or bridge-based algorithms. Any fact-checking organization whose effectiveness is measured by its ability to reduce the spread of the claims it selects for intervention will tend to prioritize "easy-to-refute" claims. For example, suppose the primary metric of success is the average reduction in the duration an individual continues to share a claim after exposure to a fact-check. In that case, fact-checkers are naturally incentivized to target claims for which interventions yield the largest and most reliable behavioral

change. Claims already less resilient to correction are more likely to be targeted and successfully suppressed. In contrast, more resistant claims—due to ambiguity, motivated reasoning, or partisan resonance—are deprioritized or avoided entirely.

To see the effect of these incentive structures on the long-term propagation of false narratives, consider the following example.

Example: ***Community Notes.*** Consider a fact-checking system that incentivizes fact-checking of only the easier-to-refute claims. If type-2 claims are the easier to fact check claims ($\alpha_2 < \alpha_1$), all the resources (R) will be devoted to it. However, due to condition (15), type-2 claims will not survive in the long run regardless of the resources devoted. It will reduce the short-term spread due to increasing T_2, but there is no impact on the long-term fraction of people propagating the false narrative. Alternatively, if $\alpha_1 < \alpha_2$, all the resources will be devoted to fact-checking the type-1 claim. This will decrease the long-run propagation of the false narrative if R is sufficiently low such that condition 16 is not violated. Let $r_1 = r_1^m$ be the level of resources devoted to combat type-1 claim such that condition 16 is violated. For a fact-checking system with sufficiently high resources $R > r_1^m$, type-2 claims will be the medium through which the false narrative will survive in the long run and any extra resources spent on T_1 beyond r_1^m will be wasted.

Furthermore, devoting more than r_1^m resources to intervene on type-1 claims can lead to a higher fraction of the population discussing the false narrative compared to a situation where the fact-checker devotes less than r_1^m resources. Figure 1 demonstrates this situation, assuming that type-1 claims are more sensitive to fact-checking $(\alpha_1 > \alpha_2)$ and holding constant the duration of sharing with no interventions ($\gamma_1 = \gamma_2$) and duration in the recovery state ($\nu_1 = \nu_2$). The parameter assumptions are $\beta_1 = 5, \beta_2 = 1.5, \gamma_1 = \gamma_2 = 0.5, \alpha_1 = 2.5, \alpha_2 < \alpha_1$ and $\nu_1 = \nu_2 = 2$.

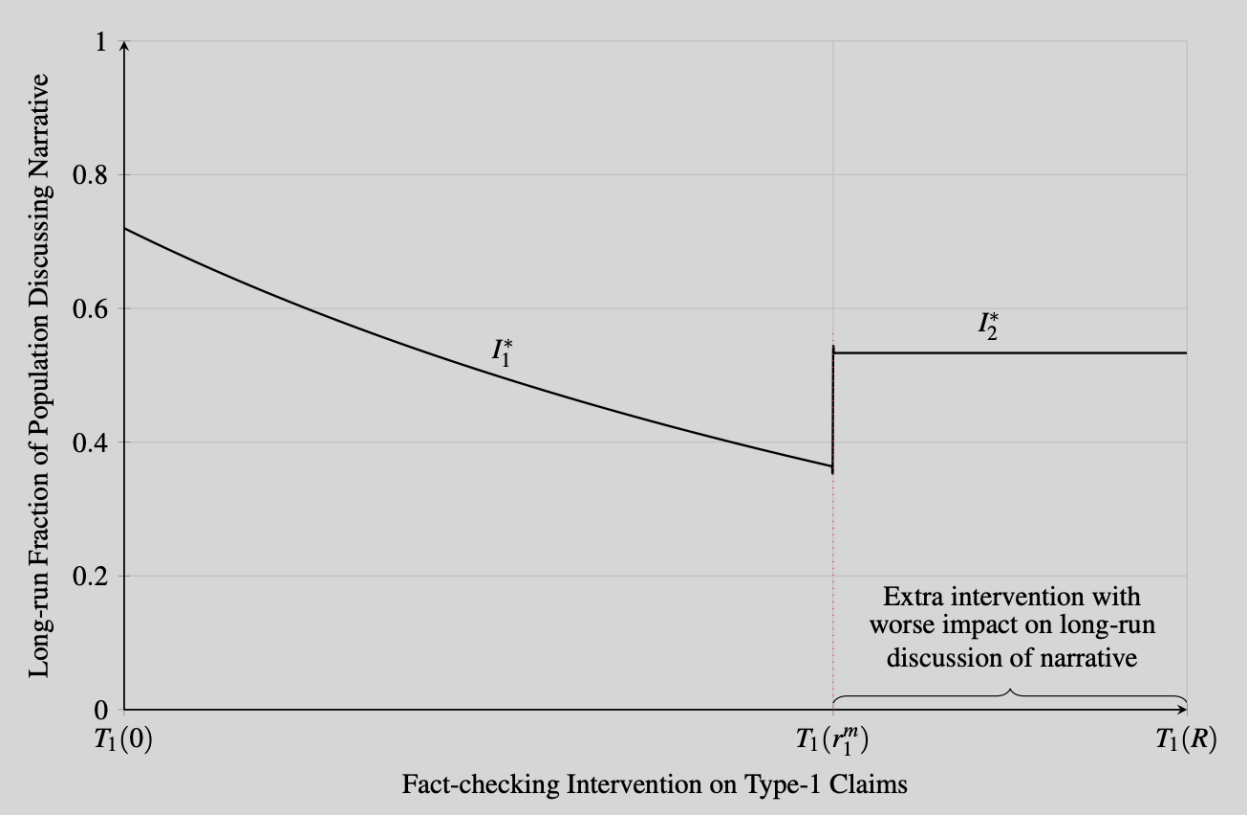

Figure 1: *An Example of Long-run Discussion of Narrative vs.* T_1

The example illustrates that when a claim type has both a higher basic reproduction number and is easier to fact-check, a success criterion based on reducing the expected duration of sharing that type of claim can be effective if $R < r_1^m$ however if $R > r_1^m$, additional resources devoted to intervening on type-1 claims beyond r_1^m have no impact on the long-run discussion of the narrative and may even increase the fraction of people discussing it if $I_2^* > I_1^*(T_1(r_1^m))$.

In contrast, if the easier-to-fact-check claim also has a lower basic reproduction number, allocating all resources to that claim does not affect long-run prevalence: the harder-to-suppress claim remains dominant regardless of the intervention. In this case, the fact-checking effort satisfies the institutional success metric but fails to reduce the persistence of the false narrative.

If the goal of the fact-checker is to minimize the long-term costs imposed by the narrative—potentially conditional on claim type—the general optimization problem can be stated as follows

$$\begin{aligned} \min_{r_1,r_2\geq 0} \quad & C_1\left(I_1^*\left(T_1(r_1),T_2(r_2)\right)\right)+C_2\left(I_2^*\left(T_1(r_1),T_2(r_2)\right)\right) \\ \text{subject to} \quad & r_1+r_2\leq R,\ C_j'(I_j^*)>0,\ C_j(0)=0. \end{aligned} \tag{17}$$

One can define the level of resources devoted to intervening on type-1 claims that leads to a switch to the endemic equilibrium in which type-2 claims dominate. Let r_1^m be such that

$$T_1(r_1^m) = \frac{\beta_1\gamma_2 - \beta_2\gamma_1}{\beta_2\alpha_1} + \frac{\beta_1\alpha_2}{\beta_2\alpha_1} T_2(r_2). \tag{18}$$

It is easy to see that r_1^m increases in r_2 and $r_1^m(r_2 = 0) > 0$ due to (14). To solve the optimization problem, one must consider two cases.

Case 1: If $r_1^m(r_2 = 0) > R$, all feasible allocation of resources would lead to a long-run discussion of the narrative via type-1 claims. Therefore, since $\frac{\partial I_1^*}{\partial T_1} < 0$ and T_1 is increasing in r_1, the optimal allocation of resources would be $(r_1, r_2) = (R, 0)$.

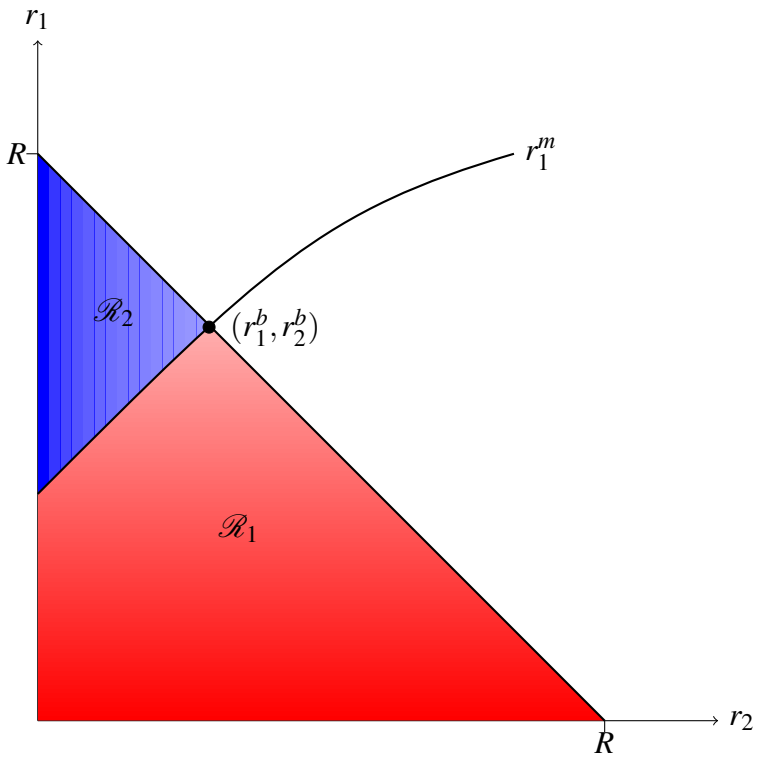

Figure 2: *Resource allocation space for Case 2*

Case 2: If $r_1^m(r_2 = 0) < R$, there are two regions in the feasible allocation space of resources with different claim types dominating in the long run. Let the unique intersection of the equilibrium switch line and the budget constraint be defined as follows

$$(r_1^b, r_2^b) = \{(r_1, r_2) | r_1 = r_1^m(r_2)\} \cap \{(r_1, r_2) | r_1 = R - r_2\}. \tag{19}$$

The region where (r_1, r_2) leads to a long-run discussion of the narrative via type-1 claims is

$$\mathscr{R}_1 = \{(r_1, r_2) | r_1 < r_1^m(r_2), r_2 \in [0, r_2^b]\} \cup \{(r_1, r_2) | r_1 < R - r_2, r_2 \in [r_2^b, R]\}. \tag{20}$$

And the region where (r_1, r_2) leads to a long-run discussion of the narrative via type-2 claims is

$$\mathscr{R}_2 = \{(r_1, r_2) | r_1^m(r_2) < r_1 \leq R - r_2, r_2 \in [0, r_2^b]\} \tag{21}$$

See Figure 2 for a visualization of the resource allocation space. Lighter colors are used to denote lower costs for both endemic equilibria. Since r_1^b is the highest possible allocation to $r_1 \in \mathscr{R}_1$ and r_2^b is the highest possible allocation to $r_2 \in \mathscr{R}_2$, the minimum cost resource allocation is

$$(r_1^*, r_2^*) = \arg\min \left\{ \begin{array}{c} \lim\limits_{\substack{(r_1,r_2)\to(r_1^b,r_2^b) \\ (r_1,r_2)\in\mathscr{R}_1}} C_1\left(I_1^*\left(T_1(r_1)\right)\right), \\ \lim\limits_{\substack{(r_1,r_2)\to(r_1^b,r_2^b) \\ (r_1,r_2)\in\mathscr{R}_2}} C_2\left(I_2^*\left(T_2(r_2)\right)\right) \end{array} \right\}.$$

This implies that with a sufficiently high resource budget, any fact-checking system that aims to minimize the cost of long-term discussion of the narrative will invest in corrections of both type-1 and type-2 claims even though the type 1 claim has a higher reproduction number with no fact-checking interventions. The central insight of this optimization is summarized in the following proposition.

Proposition: *If the goal of a fact-checking system is to reduce the cost of long-run discussion of the false narrative, given sufficient resources ($R > r_1^m(r_2 = 0)$), a fact-checker must devote non-zero resources to fact-check both types of claims.*

Note that the result that the optimal fact-checking system must allocate resources to both claim types, given sufficient resources, does not depend on the specific costs associated with the endemic equilibria of each claim. Even if type 1 claims are highly costly and type 2 claims are relatively benign, allocating excessive resources to type 1 claims to force a region switch can be inefficient. The region change will occur at a higher resource threshold by strategically devoting some resources to suppressing type 2 claims before the switch. However, the eventual prevalence—and thus the cost—of type 2 in the $\mathscr{R}_2$ can be even lower. Conversely, if type 1 claims are low cost and type 2 claims are highly costly in their endemic equilibria, targeting type 2 claims first frees up more resources to suppress type 1, achieving a lower total societal cost in $\mathscr{R}_1$. In both scenarios, the lowest-cost outcome is realized not by exclusively focusing on one claim or forcing a region switch but by jointly managing both claims in anticipation of region transitions and long-run impacts.

3 Conclusion

This paper presents a multi-claim compartmental model of misinformation to show that *which* claims a fact-checking system chooses to fact-check can shape the long-run survival and costs of a false narrative. Because claims differ in virality and resistance to correction, concentrating limited fact-checking effort on the easiest-to-debunk or most viral content can produce the illusion of short-term success, potentially ceding the field to harder-to-debunk claims that sustain the narrative in the long run. Furthermore, such an approach can lead to inefficient allocation of resources, especially if the easiest-to-debunk claims are relatively less viral or are over-targeted. The optimal allocation, therefore, always involves a positive share of resources for *all* salient claim types once resources are large enough, even when some claims appear low-impact or low-virality ex-ante.

The model yields three primary takeaways for platforms and policymakers.

1. **Beware of short-run metrics.** Evaluations that reward immediate visibility reductions or "debunk counts" can perversely incentivize efforts that can worsen long-run prevalence.

2. **Hybrid mitigation is important.** Crowd-sourcing excels at flagging blatant falsehoods in a scalable fashion. However, professional fact-checkers (or algorithmic tools) must systematically target the complex, high-resistance claims that community consensus struggles to address.

3. **Design incentives around narrative-level outcomes.** Fact-checking allocations and incentives should be guided by their effect on the overall narrative's prevalence rather than focusing on individual claims separately.

Limitations and future research. The stylized model presented in this paper abstracts away from network structure, heterogeneous user susceptibilities and influences, significant claim mutation, and cross-group or cross-platform spillovers. Theoretical extensions in these directions are needed to better elucidate how the allocation of fact-checking resources affects the long-term persistence of false narratives. Empirical work with platform data may investigate realistic spread mechanisms and the role of competition between claims. Experimental work could uncover biases in professional fact-checking and crowd-sourced methods, sharpening policy guidance for platforms and policymakers.

References

Hunt Allcott and Matthew Gentzkow. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236, May 2017.

V. Andreasen, J. Lin, and S. A. Levin. The dynamics of cocirculating influenza strains conferring partial cross-immunity. *Journal of Mathematical Biology*, 35(7):825–842, August 1997.

David La Barbera, Eddy Maddalena, Michael Soprano, Kevin Roitero, Gianluca Demartini, Davide Ceolin, Damiano Spina, and Stefano Mizzaro. Crowdsourced Fact-checking: Does It Actually Work? *Information Processing & Management*, 61(5):103792, September 2024.

Nadia M. Brashier, Gordon Pennycook, Adam J. Berinsky, and David G. Rand. Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, 118(5): e2020043118, February 2021.

Michael V. Bronstein, Gordon Pennycook, Adam Bear, David G. Rand, and Tyrone D. Cannon. Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, 8(1):108–117, 2019.

Man-Pui Sally Chan, Christopher R. Jones, Kathleen Hall Jamieson, and Dolores Albarracín. Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*, 28(11):1531–1546, November 2017.

Onyema Greg Chido-Amajuoyi, Robert K. Yu, Israel Agaku, and Sanjay Shete. Exposure to Court-Ordered Tobacco Industry Antismoking Advertisements Among US Adults. *JAMA Network Open*, 2(7):e196935, July 2019.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The COVID-19 social media infodemic. *Scientific Reports*, 10(1):16598, October 2020.

Alice Dechêne, Christoph Stahl, Jochim Hansen, and Michaela Wänke. The Truth About the Truth: A Meta-Analytic Review of the Truth Effect. *Personality and Social Psychology Review*, 14(2): 238–257, May 2010.

Saoirse A. Connor Desai, Toby D. Pilditch, and Jens K. Madsen. The rational continued influence of misinformation. *Cognition*, 205:104453, December 2020.

Ullrich K. H. Ecker, Li Qian Tay, Jon Roozenbeek, Sander van der Linden, John Cook, Naomi Oreskes, and Stephan Lewandowsky. Why misinformation must not be ignored. *American Psychologist*, 2024.

Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5(1):eaau4586, January 2019.

Roland Imhoff, Felix Zimmer, Olivier Klein, João H. C. António, Maria Babinska, Adrian Bangerter, Michal Bilewicz, Nebojša Blanuša, Kosta Bovan, Rumena Bužarovska, Aleksandra Cichocka, Sylvain Delouvée, Karen M. Douglas, Asbjørn Dyrendal, Tom Etienne, Biljana Gjoneska, Sylvie Graf, Estrella Gualda, Gilad Hirschberger, Anna Kende, Yordan Kutiyski, Peter Krekó, Andre Krouwel, Silvia Mari, Jasna Milošević Đorđević, Maria Serena Panasiti, Myrto Pantazi, Ljupcho Petkovski, Giuseppina Porciello, André Rabelo, Raluca Nicoleta Radu, Florin A. Sava, Michael Schepisi, Robbie M. Sutton, Viren Swami, Hulda Thórisdóttir, Vladimir Turjačanin, Pascal Wagner-Egger, Iris Žeželj, and Jan-Willem van Prooijen. Conspiracy mentality and political orientation across 26 countries. *Nature Human Behaviour*, 6(3):392–403, March 2022.

Jonas L. Juul and Johan Ugander. Comparing information diffusion mechanisms by matching on cascade size. *Proceedings of the National Academy of Sciences*, 118(46):e2100786118, November 2021.

Joel Kaplan. More speech and fewer mistakes. `https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/`, January 2025.

Adam Kucharski. Study epidemiology of fake news. *Nature*, 540(7634):525–525, December 2016.

Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, 13(3): 106–131, December 2012.

Sahil Loomba, Alexandre de Figueiredo, Simon J. Piatek, Kristen de Graaf, and Heidi J. Larson. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, 5(3):337–348, March 2021.

Brendan Nyhan and Jason Reifler. Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. *Vaccine*, 33(3):459–464, January 2015.

Brendan Nyhan, Jason Reifler, Sean Richey, and Gary L. Freed. Effective messages in vaccine promotion: a randomized trial. *Pediatrics*, 133(4):e835–842, April 2014.

Katherine Ognyanova, David Lazer, Ronald E. Robertson, and Christo Wilson. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*, June 2020.

Jessica Paynter, Sarah Luskin-Saxby, Deb Keen, Kathryn Fordyce, Grace Frost, Christine Imms, Scott Miller, David Trembath, Madonna Tucker, and Ullrich Ecker. Evaluation of a template for countering misinformation—Real-world Autism treatment myth debunking. *PLOS ONE*, 14(1): e0210746, January 2019.

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

Francesco Pierri, Brea L. Perry, Matthew R. DeVerna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer, and John Bryden. Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Scientific Reports*, 12(1):5966, April 2022.

Jo Saunders and Alice Jess. The effects of age on remembering and knowing misinformation. *Memory*, 18(1):1–11, January 2010.

David Scales, Jack Gorman, and Kathleen H. Jamieson. The Covid-19 Infodemic — Applying the Epidemiologic Model to Counter Misinformation. *New England Journal of Medicine*, 385(8): 678–681, August 2021.

Philip Smith, Maansi Bansal-Travers, Richard O'Connor, Anthony Brown, Chris Banthin, Sara Guardino-Colket, and K. Michael Cummings. Correcting over 50 years of tobacco industry misinformation. *American Journal of Preventive Medicine*, 40(6):690–698, June 2011.

Briony Swire-Thompson, Joseph DeGutis, and David Lazer. Searching for the Backfire Effect: Measurement and Design Considerations. *Journal of Applied Research in Memory and Cognition*, 9(3):286–299, September 2020.

The Associated Press. Musk's twitter has dissolved its trust and safety council. `https://www.npr.org/2022/12/12/1142083744/twitter-trust-and-safety-council-elon-musk`, December 2022.

P. van den Driessche and James Watmough. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences*, 180: 29–48, 2002.

Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, March 2018.

Nathan Walter and Riva Tukachinsky. A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication Research*, 47(2):155–177, 2020.

Wei-Chun Wang, Nadia M. Brashier, Erik A. Wing, Elizabeth J. Marsh, and Roberto Cabeza. On Known Unknowns: Fluency and the Neural Mechanisms of Illusory Truth. *Journal of Cognitive Neuroscience*, 28(5):739–746, May 2016.

Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, M. B. Fallin Hunzaker, Keith Coleman, and Jay Baxter. Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation, October 2022. arXiv:2210.15723.

Thomas Wood and Ethan Porter. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 41(1):135–163, 2019.

Hamza Yousuf, Sander van der Linden, Luke Bredius, G. A. (Ted) van Essen, Govert Sweep, Zohar Preminger, Eric van Gorp, Erik Scherder, Jagat Narula, and Leonard Hofstra. A media intervention applying debunking versus non-debunking content to combat vaccine misinformation in elderly in the Netherlands: A digital randomised trial. *eClinicalMedicine*, 35, May 2021.