Synergies between deplatforming and inoculation against online extremism: an agent-based simulation and cost-benefit analysis

Matteo Vergani, Andrea Giovannetti, Stephanie Zi Xin Ng, Chee Peng Lim, James Zhang, Robin Scott

Classification: Social Sciences

Keywords: Inoculation, Online extremism, Agent-based modelling, Terrorism, Violent extremism, Countering Violent Extremism, Online radicalisation

Abstract

Inoculation is known to effectively reduce the spread of online extremism; however, it is unclear to what extent inoculation is needed when combined with other counter-extremism strategies such as monitoring and removing extremist users. We develop an agent-based model of diffusion and control of online extremism calibrated using data from systematic reviews and empirical studies. A population of Bayesian agents (exemplified by their belonging to an anti-immigration Facebook group) forms a binary belief about far-right content unilaterally spread by extremists embedded within the social network. We introduce two interventions to mitigate the spread of extremism: deplatforming – consisting in monitoring and removing extremist users – and inoculation – consisting in exposing users to weakened forms of extremist arguments to build belief resistance. We find strong evidence that inoculation is synergetic to deplatforming: for given levels of effectiveness, inoculation can *partially substitute* deplatforming. We pin down this synergetic relationship to a protective effect exerted by inoculation on individual belief formation, which, at group level, translates into a stabilisation of opinion dynamics. Therefore, the introduction of inoculation on top of deplatforming makes the outcome of a deplatforming intervention less arbitrary on average and more effective in controlling the spread of extremism. We exploit the partial substitution effect to identify the minimum amount of inoculation required to control the spread of extremism for each additional dollar spent on repression. Cost-benefit analysis highlights the efficiency advantages of having a comprehensive policy approach encompassing both types of measures. Our findings offer important evidence to regulators and businesses in supporting investments in inoculation interventions alongside repressive efforts.

Introduction

Extremism and its active control put a heavy toll on the functioning of democracies and their budget, with online platforms being a key channel for the spread of extremist ideas globally. Although extremist content generally does not appeal to viewers (Hassan et al., 2020), it can increase support for extremist views among those who share a salient identity and ideological leanings with extremists. This has been suggested in previous qualitative studies with former terrorists (Koehler, 2014) and survey research with general population samples (Schumann et al., 2024). Individual differences also matter in determining susceptibility to extremist ideas: a systematic review by Wolfowicz et al. (2020) found that personality factors (e.g., low self-control, thrill-seeking) and demographic characteristics (e.g., being younger, male, and less educated) predict the risk of developing extremist ideas. Previous research has drawn parallels between contagion dynamics and the spread of extremism on social media channels (Ferrara, 2017).

Technology companies and governments all over the world heavily invest in so-called 'hard' counter-extremism approaches such as monitoring and ad-hoc removal of extremists from online platforms via deplatforming and/or arrest. This a class of strategies is costly and prone to legal and technical inefficiencies (Corbeil & Rohozinski, 2021; Vu, Hutchings & Anderson, 2023)Such measures pose challenges to democratic societies, including potential backlash and perceived infringements on freedom of speech (Dugan & Fisher, 2023). In the long term, deplatforming may inadvertently erode trust in democratic institutions (Spalek, 2010). Paradoxically, deplatforming can backfire and actually strengthen the target of the policy intervention (see Vu, Hutchings & Anderson, 2023, for an emblematic example of a large-scale deplatforming failure) Last, and equally important, deplatforming potentially carries a string of social and health-related costs: monitoring is largely a human-intensive, psychologically tolling task (see Steiger et al., 2021 for a contemporary review), with an estimated 100,000 people working in the commercial content moderation industry, a sector subject to global competitive pressure and weak work rights(Steiger et al., 2021).

Conversely, so-called 'soft' primary counter-extremism approaches that aim at preventing the onset of radicalisation in the general population (Hardy, 2022) encompass public education, training, outreach activities, and strategic communications, which align more closely with democratic principles by promoting engagement, empowerment, and resilience within communities (Stephens et al., 2021). Among 'soft' approaches, inoculation is one of the most promising evidence-based interventions to curb the spread of online extremism (Braddock, 2019). The main idea at the basis of inoculation is that by exposing individuals to weakened forms of an argument, they develop resistance to it, much like a biological vaccine works by introducing a harmless version of the virus to stimulate the immune system (Lewandowsky & Van Der Linden, 2021). Inoculation theory operates through two primary mechanisms: the presentation of a pre-emptive warning that activates a perception of threat, motivating individuals to protect their existing beliefs, and the use of refutational pre-emption, or pre-bunking, which involves presenting arguments that challenge these beliefs while simultaneously offering counterarguments (Roozenbeek et al., 2020). Over decades, empirical research across various fields, including health communication and political campaigning, has demonstrated the effectiveness of inoculation strategies in conferring resistance to persuasion (Banas and Rains 2010). Two randomised controlled trials tested the effects of inoculation to

counter cognitive radicalisation outcomes (Braddock, 2019; Lewandowsky and Yesilada, 2021) confirming the value of inoculation in curbing the spread of online extremism.

Although the value of inoculation in reducing the spread of online extremism is demonstrated, previous studies were conducted with small samples and in isolation from other counter-extremism policies. Therefore, it remains unclear to what extent inoculation is needed when integrated with other counter-extremism approaches, such as monitoring and removing extremist users. In this study, we develop an agent-based model of diffusion and control of online extremism to test the synergies between inoculation and deplatforming. While our approach within this realm is general and could be applied to various forms of extremism, for simplicity we frame the discussion on the context of far-right extremism, which is characterised by the advocacy for an ideological core made of White superiority and violence against diverse groups and minorities within a liquid and fungible set of narratives (Mudde, 2019). These ideas are often advocated by organisations that effectively use mainstream online platforms such as Facebook to recruit new members through spreading extremist propaganda and forging personal relationships with online users (McSwiney, 2021).

In this article, we perform two exercises. In the first exercise, we investigate the mechanism of action and effectiveness of these two structurally diverging policies and their interactions. Artificial data allows us to identify a "possibility frontier" (Sickles and Zelenyuk, 2019), that is the minimum amount of inoculation required to control the spread of extremism for each additional dollar spent on deplatforming. We operationalise the frontier to identify whether *structural* synergies between the two policies exist. Structural synergies imply that given a level of effectiveness of deplatforming, inoculation can, at least partially, *substitute* it. In the second exercise, we attempt an assessment of the economic value of inoculation policies. To do so, we calibrate the model to real-world cost figures and run a cost-benefit analysis. If inoculation is cheaper than deplatforming, structural synergies translate into *economic* synergies by effectively reducing the net load on resources required to attain that level of effectiveness. Therefore, if economic synergies exist, the possibility frontier provides clear *normative* guidance to enforcers and policy designers on efficient cost planning of inoculation policies: the *maximum* cost that a rational enforcer *should* be willing to pay for enforcing a specific inoculation policy should be *at most* equal the dollar-savings on deplatforming policy caused by the introduction of the inoculation.

Model

We model an online community made of individuals with a shared interest and political identity broadly aligned with far-right extremist views, exemplified by their belonging to an anti-immigration Facebook group. Within this Facebook group, we assume heterogenous leanings toward extremism, depending on social structures and demographic characteristics of each individual (Wolfowicz et al. 2020). As in a realistic scenario, although all group members oppose immigration, not all of them will have the same predispositions towards accepting far-right extremist ideas inciting violence against immigrants. As we know from personal accounts of former extremists and studies on online recruitment (Speckhard & Ellenberg, 2020; Weimann, 2021; McSwiney, 2021), extremists often integrate themselves into online groups comprising individuals potentially receptive to their ideologies. They start conversations and disseminate selective propaganda through personal interactions, aiming to cultivate interest and

facilitate radicalisation. Should these efforts prove successful, the frequency of contact increases, with interactions frequently extending to offline engagements. Our study attempts to replicate these dynamics, in particular the first stages of online interactions and cognitive radicalisation. Our agent-based model consists of four agent types: users, extremists, inoculators, and enforcers that interact along discrete time iterations, $t = 1, \dots, T$. Details on the behavioural assumptions and calibration are contained in Appendix I.

Users. Users are Bayesian rational agents that dynamically form a belief about the extremist opinion, the *radicalisation level* B_i , a dynamic variable ranging from 0 (no radicalisation) to 1 (full radicalisation). We track radicalisation by means of a population-level matrix $\boldsymbol{B}$ where each column contains an individual whose radicalisation is stored through the rows. Users are heterogenous in two broad classes of features. First, they potentially diverge in their individual characteristics. These are measurable items such as ideological leanings, degree of education, age and other demographic features that contribute in determining their *level of susceptibility* to the extremist opinion $\gamma_i \in [0,1]$. We report the full list of individual characteristics (based on Wolfowicz et al., 2020) as well as the determination of γ in Appendix I. Second, albeit embedded within a single Facebook group, they may diverge in the structure of social relationships. We generically refer to linked individuals as "friends" and model the web of connections through a simple inter-personal network $\boldsymbol{M}$. To generate our network, we adopt a small-world protocol (Jackson & Patil, 2022) using the calibration of Aiello et al. (2016) of real-world Facebook group networks. In our network, two types of information circulate: extremist information and inoculation against extremism. These are the two types of information relevant for updating the radicalisation level of users. This information flows through three channels: other users, extremists, and/or inoculators. In every period, users might receive extremist content through private messages on Facebook. Each user will then assess the *internal credibility* $c_i \in [0,1]$ of the content as described in the Appendix materials.

Enforcers. These are online moderators, broadly defined (e.g. Facebook moderators, contractors, police, antiterrorism), that is individual actors with full access to users' online activity that monitor and moderate the network and identify and deplatform extremists. We impose two realistic procedural rules to the behaviour of enforcers. First, identification of extremists goes through two parallel channels: user reporting and direct monitoring. While we assume that enough user reporting of any given extremist will lead to their deterministic removal, we allow direct monitoring to be potentially *imperfect*. While granular data on the efficiency of online counter-extremism are scarce and ad-hoc to specific online contexts (Baruch, Ling, and Hofman, 2018), there is an exponential increase in storage, machine, and human-based costs (see ACMA, 2023). This increase is coupled with a progressive expansion of public budgets and stricter regulations in the realm of online policing (Winter et al., 2020), indicating that current practices are imperfect due to various factors, including overloading. Therefore, we introduce *inefficiency* by assuming that the set is randomly picked across the user base of the entire Facebook group. Furthermore, the pooling is made with a monthly time frequency $\delta^R \geq 0$ to reflect the fact that investigations are not instantaneous processes. As a result, the mechanism (r^R, δ^R) captures two important and dimensions of efficiency for a police operation: the precision and turnarounds of investigations.

Extremists. Extremists are individuals who spread extremist content among users. They are assumed to be randomly embedded within the Facebook group, through which they gain exposure to users and, via Facebook friendships, obtain the ability to message them privately.

To keep our model's predictions tractable, we impose that extremists are heterogeneous only in their positioning within the social network. In every period t, until removed by enforcers, each extremist can only decide whether to perform one action, that is to spread disinformation characterised by a fixed salience $\epsilon > 0$ across all of her contacts. This action is chosen at a fixed frequency $\delta^E \geq 0$.

Inoculators. Inoculators are broadly defined as entities (for example, community organisations that deliver education-based interventions in offline settings such as schools, sports clubs, and churches) that implement inoculation policies targeting individuals who may belong to the anti-immigration Facebook group. Once deployed, their positions remain fixed throughout the entire simulation. We model inoculators as external to the network of the Facebook group because they represent real-world initiatives impacting dimensions of social interactions that are only partially reflected in the social network considered in this work. This modelling choice aligns with real-world inoculation policies, such as the recent mix of online and community-based countering violent extremism measures implemented by governments globally in various settings, including schools, youth centres, and religious venues (Barton et al., 2022). Similar to enforcers, inoculators are also imprecise and have limited direct reach of the user community. Simulation setup and timeline are described in the Appendix materials.

Analytical strategy and hypotheses

Assumptions and baseline calibrations

Let $(r^R, \delta^R, r^I, \delta^I)$ be a *policy mix*. We simplify the space of policies through calibration of the inoculation radius r^I and the investigation length δ^R. We conservatively maintain a very small radius, $r^I = 1$, thus obtaining a workable lower bound on the efficacy of the inoculation policy. Second, following the data-driven calibration of moderators' productivity described in Appendix 1, we fix $\delta^R = 1$, equivalent to one day of work per enforcer (measured at full-time employment, i.e. 1.0 FTE). This way, the policy mix maps into a tractable two-dimensional space (r^R, δ^I).

We generate a baseline model made of $K = 11$ simulations characterised by no inoculation policy, and varying levels of monitoring that range from $r^R = 0$ (no monitoring) to $r^R = 10$, such that 1% of the population is monitored every $\delta^R = 10$ days. We interact each of these scenarios with an interval of inoculation policies, ranging from one training event every 10 days per month, $\delta^I = 10$, to one where the inoculation takes place on daily basis, $\delta^I = 1$. For each policy mix, the model is simulated for a total of $T = 365$ periods and replicated across $M = 100$ runs. As a result, our analysis is grounded on a space made of $11 \times 11 = 121$ distinct policy mixes, each simulated for $m = 1, \dots, M = 100$ times. This procedure generates a pool of $11 \times 11 \times 100 \times 365 = 4{,}416{,}500$ artificial data points.

Last, to run the cost-benefit analysis we include information on productivity and wage of moderators (USD valued, March 2024). To construct a convincing lower bound on costs, we adopt a costing range spanning from \$2.20 per hour to \$18 per hour, corresponding to contemporary outsourced Facebook content moderator hourly market wage in a low-income country (i.e. Kenya), and in the U.S, respectively (Adeyemi, 2022).

Study 1: Measuring performance and structural synergies

To evaluate the effectiveness of alternative policing strategies and test the existence of structural synergies, we start with a parametric approach and run two sets of regressions, respectively named *Model 1* and *Model 2*, on two performance measures, (1) and (2), constructed as follows. Let $n_m\,(t) \geq 0$ be the fraction of vulnerable users, that is users in high risk or extreme state (corresponding to an individual belief $B \geq 0.65$) in period t and simulation run $m = 1, \dots, 100$. The performance measures are: (1) the *mean fraction* of vulnerable users $\bar{n}_m$ computed, for each simulation m, across periods $t = 1, \dots, 365$; and (2) the *end fraction* of vulnerable users $n_m\,(T)$ observed the end period $T = 365$. Given the fractional nature of performance measures, we opt for a fractional regression approach (Papke and Woolridge, 1996) where the output variable is the selected performance measure. The fractional regression model is the standard estimation framework for fractional output variables (Woolridge, 2010).

Next, as interaction effects can involve complex non-linearities which is hard to capture through a parametric model (Woolridge, 2010), to understand synergies between policies we take a non-parametric, descriptive approach and construct a simple empirical *possibility frontier* (Sickles and Zelenyuk, 2019). The possibility frontier metricizes the effect of policy mixes across the full support (r^R, δ^I). To do so, we proceed in three steps. First, we map each of the 11×11 policy mixes (r^R, δ^I) to simulated performance measures (1) and (2), averaged across the $M = 100$ simulations. We call these averages $\bar{n}$ and $\bar{n}\,(T)$. To account for cross-simulation heterogeneity, we further augment each the 121×2 averages by one standard deviation as obtained from the $M = 100$ simulations, $\bar{\sigma}$ and $\bar{\sigma}\,(T)$. We refer to these as the *augmented* performance measures (1*) and (2*). Second, we use (1*) and (2*) to split policy configurations in three groups, *High-Risk*, *Medium-Risk* and *Low-Risk*, depending on whether the average of augmented performance (1*) and (2*) is above 5%, between 5% and 3% and below 3%, respectively. Third, to obtain the graphical description of effects, we map these three sets in the 11×11 policy mix space (r^E, δ_I) on a heatmap against the augmented average performance measures (1*) and (2*). We explore the mechanism of functioning of inoculation by disaggregating measures (1*) and (2*) in an analysis of means and variances. Last, to exemplify the role in variance-reduction of inoculation, we pick three mixes belonging to the *Medium-Risk* set and describe the temporal evolution of the fraction of vulnerable users $n_m\,(t)$ across the $t = 1, \dots, 365$ periods for each of the $m = 1, \dots, 100$ simulation.

Study 2: Cost-benefit analysis and economic synergies

We conduct a cost-benefit analysis of policy mixes to assess whether structural synergies can translate in economic synergies. For simplicity, we only focus on performance measure (1). Using estimates of Model 1, we consider the set of policy mixes attaining a given predicted performance. Within such set, we define the economic value of inoculation as the difference between the cost of monitoring attracted by the policy with no inoculation and the minimum level of monitoring versus the cost of monitoring for a policy that uses the minimum positive amount of inoculation and minimum monitoring. The difference in cost identifies the economic value of inoculation, hence its *maximum rational cost*. The empirical building block of the

analysis is given by the productivity and cost of online moderation as described in Section 4.1. We assess that the cost of a single Facebook contractor running a one-day investigation (equivalent to 7 working hours) targeting the 0.1% of the user population ranges between $15.4 and $126, depending on whether the firm is based in a low-income country like Kenya or the U.S.. A linear costing structure implies that a one-off investigation of the 2% of the population will cost between $30.8 and $252, and so forth. On yearly basis, the per-user cost of monitoring 1% of the population ranges between $56.21 and $459.90. Practically, we run the analysis by mapping the productivity and cost information above into an analysis of the marginal effects of policy mixes as extracted from the regression models of Section 4.2.1. The procedure is described in Appendix 2.2.

Model sensitivity and validation

To guarantee robustness of results, models are fed and evaluated along a continuous interval of policy mixes (r^E, δ^I). Sensitivity tests of outputs is at two levels: first, confidence intervals are constructed for the mean value of each relevant variable through a monte-carlo procedure (with $M = 100$ repetitions). Second, the parametric space of calibrated parameters δ^R and r^I is explored in the neighbourhood of the chosen values to confirm the qualitative structure of results.

Results

In Table 2 we collect the estimate of fractional regression models described in Section 4.2.1, where for sign readability, inoculation frequency δ^I has been transformed to its inverse, $-\delta^I$.

From the table, the coefficient of each policy, r^R and $(-\delta^I)$ is associated to a pronounced and significant reduction of the mean and the end number of vulnerable users ($p < 0.01$). Furthermore, the interaction term $r^R \times (-\delta^I)$ is positive and significative, thus implying that a significant substitution effect exists between the two policies: *ceteris paribus,* increasing the inoculation frequency causes a *decrease* in the absolute efficacy of deplatforming. In other words, structural synergies exist between the two policies as the use of one policy, *given that the total fraction of vulnerable users is constrained to 1* (regardless to the selected performance measure), decreases the marginal impact of the other policy.

In the regression specification of Table 1, interactions (hence, substitution effects) are assumed to be linear across the support of the policies. To better understand the interacting effects of the two policies, we construct a possibility frontier of the two policies by mapping the full set of policy mixes (r^R, δ^I) into the augmented performance measures (1*) and (2*) and construct the policy sets *High-Risk, Medium-Risk* and *Low-Risk* as described in Section 4.2.1. We report the average effect of each set in Table 2.

We note that performance between these three sets differs significantly. *High-Risk* obtains an augmented average and final fraction of vulnerable individuals of 34% and 16.6% ($p < 0.01$), respectively. For policies in *Medium-Risk,* the fractions reduce to 3.9% and 2.7% ($p < 0.01$),

respectively. Lastly, for policies in *Low-Risk,* the amounts corresponds to 1.6% and 1.3% ($p < 0.01$), respectively.

	Model 1 Effect on (1)	Model 2 Effect on (2)
r^R	-0.319***	-0.319***
	(0.01)	(0.00)
$-\delta^I$	-0.422***	-0.422***
	(0.01)	(0.00)
$r^R \times (-\delta^I)$	0.057***	0.057***
	(0.00)	(0.00)
Constant	-1.732***	-1.732***
	(0.03)	(0.02)
AIC	5988.524	5988.524
BIC	6018.095	6018.095
r^2	0.357	0.358

Table 1. *AIC* and *BIC* stand for Akaike Information criterion and Bayesian Information Criterion, respectively. Robust standard errors are in parenthesis. Symbols *, ** and *** represent statistical significance at 0.1, 0.05 and 0.01 level, respectively.

	Model 3 Effect on (1*)	Model 4 Effect on (2*)
High-Risk	0.342***	0.166***
	-0.04	(0.02)
Medium-Risk	-0.303***	-0.139***
	-0.04	(0.02)
Low-Risk	-0.326***	-0.153***
	-0.04	(0.02)
AIC	-45.513	-248.41
BIC	-37.125	-240.023
r^2	0.434	0.434

Table 2. Average difference in the augmented performance measures (1*) and (2*) between the three sets of policy mixes defined in Section 4.2.1. Coefficients are reported in percentages. *AIC* and *BIC* stand for Akaike Information criterion and Bayesian Information Criterion, respectively. Robust standard errors are in parenthesis. Symbols *, ** and *** represent statistical significance at 0.1, 0.05 and 0.01 level, respectively.

Third, in Figure 1 we map the three policy sets in the 11×11 policy mix space (r^E, δ_I) against a heatmap of (1*) and (2*) respectively plotted in the left and right panel of the figure. For each panel, areas are coloured according to the policy set mix belongs to: black areas corresponds to mixes that fall within *High-Risk*, whereas grey and white areas corresponds to mixes in *Medium-Risk* and *Low-Risk*, respectively. From visual comparison of the areas (i.e. performance measures), we obtain two implications. First, a substitution effect is noticeable as movement along the axes (i.e. the expansion of one policy, *ceteris paribus*) causes an increase of performance. Second, it appears that such substitution is possible only in the space defined

by $r^R \geq 2$. Below such level of moderation, *no frequency of inoculation exists* so that the fraction of individuals in critical state can be bounded below 5%. In other words, the substitution effect is *partial.*

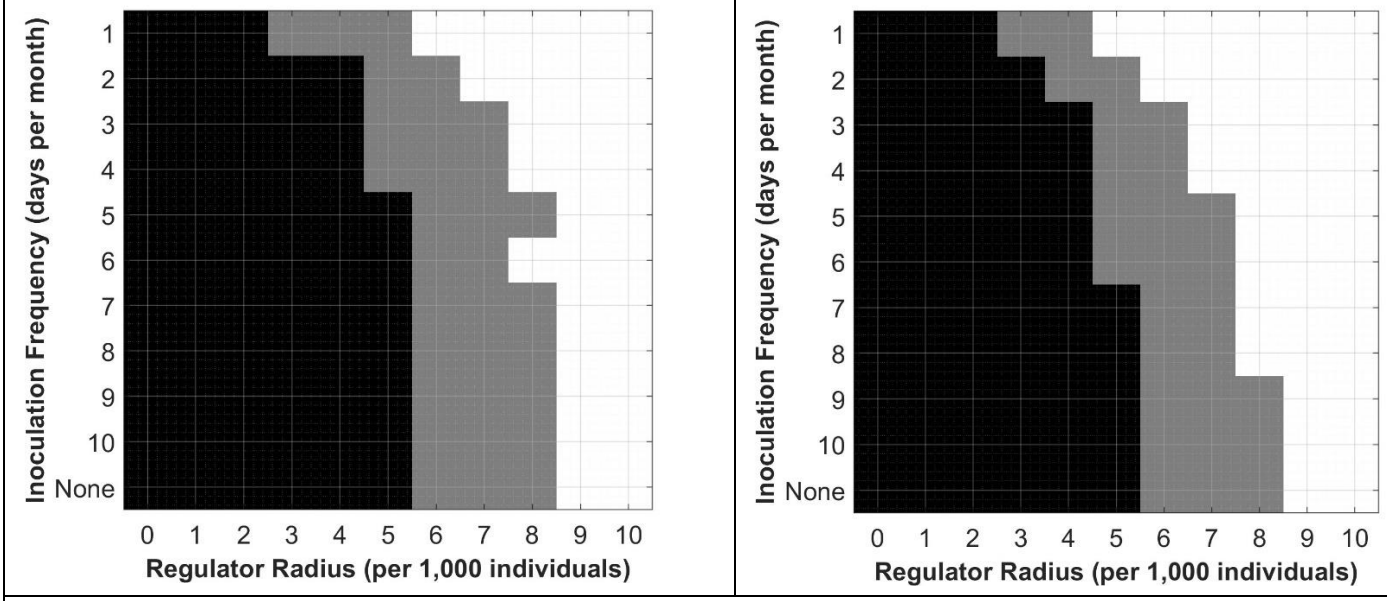

Figure 1. Possibility Frontier of policy mixes: the effect of various policy mixes on augmented performance measures as categorized in *High-Risk, Medium-Risk* and *Low-Risk* policies, respectively given by black, grey and white areas. (*Left Panel.*) Performance measure (1*). (*Right Panel.*) Performance measure (2*).

Next, in Table 3 we turn to exploring the mechanism of action of the two policies by investigating the elements used to build performance measures (1*) and (2*). To do so, we individually measure the mean fractions of vulnerable users $\bar{n}_M$ and $\bar{n}\,(T)$ along with the associated standard deviations $\bar{\sigma}_M$ and $\bar{\sigma}\,(T)$. For brevity, we report on policies corresponding to the *Medium-Risk* set. Critically, we remark that for comparable levels of average values, standard deviations are generally smaller for policy mixes featuring a *higher level of inoculation* versus deplatforming.

To validate and visualise the mechanism at the basis of the results of the variance analysis of Table 4, on the top panel of Figure 4 we plot the time series $n_m(t)$ for three configurations belonging to the *Medium-Risk* set, $(r^E, \delta^I) = (7, \text{none}), (6, 7), (4, 1)$. From Table 4, these policies are associated to $\bar{n} = 2.35\%, 2.48\%$ and 2.58%, and $\bar{\sigma} = 1.14\%, 1.09\%$ and 0.84%. We emphasize the difference between the mechanism of action of the two policies by plotting (bottom panel) the average mean fraction of vulnerable users $\bar{n}\,(t)$ and the confidence interval obtained from standard deviations $\bar{\sigma}(t)$ computed at every time period $t = 1, .., 365$ across the $M = 100$ simulations.

Cost-benefit analysis

For simplicity, we frame the cost-benefit analysis by grouping mixes attaining three distinct levels of predicted performance: $3-5\%$, $1-3\%$ and below 1%, respectively classified as *High-Tolerance*, *Medium-Tolerance and Low-Tolerance* mixes. Within each category, we compare the cost of the policy mix with no inoculation and minimum cost of moderation (i.e. the *benchmark policy*) versus the policy mix that jointly maximizes the amount of inoculation and minimizes the cost of moderation. Results for both U.S. and Kenya are reported in Figure 5 (monetary amounts measured at per-user level, in USD), as well as in Table 4, where a formal cost-benefit analysis of the value of inoculation (measured at single inoculation event) is produced for the three policy sets.

Vulnerable Users, Medium-Risk Policy Set : Performance Measure (1)

Inoculation Frequency \ Mean Enforcer Radius	2	3	4	5	6	7
1	-	3.72%	2.58%	-	-	-
2	-	-	3.85%	2.73%	-	-
3	-	-	-	3.43%	2.24%	-
4	-	-	-	3.49%	2.48%	-
5	-	-	-	3.65%	2.69%	-
6	-	-	-	3.85%	2.71%	2.01%
7	-	-	-	4.09%	2.68%	2.05%
8	-	-	-	4.08%	2.72%	-
9	-	-	-	4.29%	2.80%	2.02%
10	-	-	-	3.98%	2.54%	2.05%
none	-	-	-	-	2.98%	2.35%

Inoculation Frequency \ S.D Enforcer Radius	2	3	4	5	6	7
1	-	1.21%	0.84%	-	-	-
2	-	-	1.45%	1.03%	-	-
3	-	-	-	1.46%	1.00%	-
4	-	-	-	1.50%	1.09%	-
5	-	-	-	1.79%	1.06%	-
6	-	-	-	1.68%	1.18%	0.82%
7	-	-	-	1.96%	1.18%	1.03%
8	-	-	-	1.87%	1.37%	-
9	-	-	-	1.84%	1.47%	1.01%
10	-	-	-	1.82%	1.17%	1.04%
none	-	-	-	-	1.47%	1.14%

Vulnerable Users, Medium-Risk Policy Set: Performance Measure (2)

Inoculation Frequency \ Mean Enforcer Radius	2	3	4	5	6	7
1	4.49%	2.97%	-	-	-	-
2	-	4.28%	2.34%	-	-	-
3	-	-	3.75%	2.32%	-	-
4	-	-	4.86%	2.72%	-	-
5	-	-	-	3.15%	2.22%	-
6	-	-	-	3.42%	2.34%	-
7	-	-	-	3.83%	2.42%	-
8	-	-	-	3.80%	2.51%	-
9	-	-	-	4.08%	2.61%	-
10	-	-	-	3.82%	2.36%	-
none	-	-	-	-	3.11%	2.43%

Inoculation Frequency \ S.D Enforcer Radius	2	3	4	5	6	7
1	0.92%	1.04%	-	-	-	-
2	-	1.44%	0.96%	-	-	-
3	-	-	1.71%	0.99%	-	-
4	-	-	1.87%	1.19%	-	-
5	-	-	-	1.58%	0.87%	-
6	-	-	-	1.51%	1.02%	-
7	-	-	-	1.93%	1.04%	-
8	-	-	-	1.82%	1.28%	-
9	-	-	-	1.77%	1.36%	-
10	-	-	-	1.78%	1.07%	-
none	-	-	-	-	1.57%	1.21%

Table 3. Performance of policy mixes within the Medium-Risk policy set as defined in the main text.

	U.S.					
Policy Mix	**Performance (no inoc.)**	**Monitoring Cost (no inoc.)**	**Performance (with inoc.)**	**Max Reduction of Monitoring**	**Value of Inoculation (per inoc.)**	**Total Cost**
High Tolerance	4.27%	0.793	4.95%	38.09%	0.302	0.794
Medium Tolerance	1.96%	0.90720	2.83%	20.83%	0.189	0.907
Low Tolerance	0.52%	1.096	0.89%	6.90%	0.076	1.096
	Kenya					
Policy Mix	**Performance (no inoc.)**	**Monitoring Cost (no inoc.)**	**Performance (with inoc.)**	**Max Reduction of Monitoring**	**Value of Inoculation (per inoc.)**	**Total Cost**
High Tolerance	4.27%	0.097	4.95%	38.09%	0.037	0.097
Medium Tolerance	1.96%	0.111	2.83%	20.83%	0.023	0.111
Low Tolerance	0.52%	0.134	0.89%	6.90%	0.009	0.134

Table 4. Cost-Benefit analysis for three policy mix sets (see Section 5.2) for U.S and Kenya, respectively. Costs are per-user. Performance is measured with measure (1) as defined in Section 4.2.1. Value of inoculation is computed at the single inoculation event. Cost estimates are built from Model 1 (Section 4.2.1).

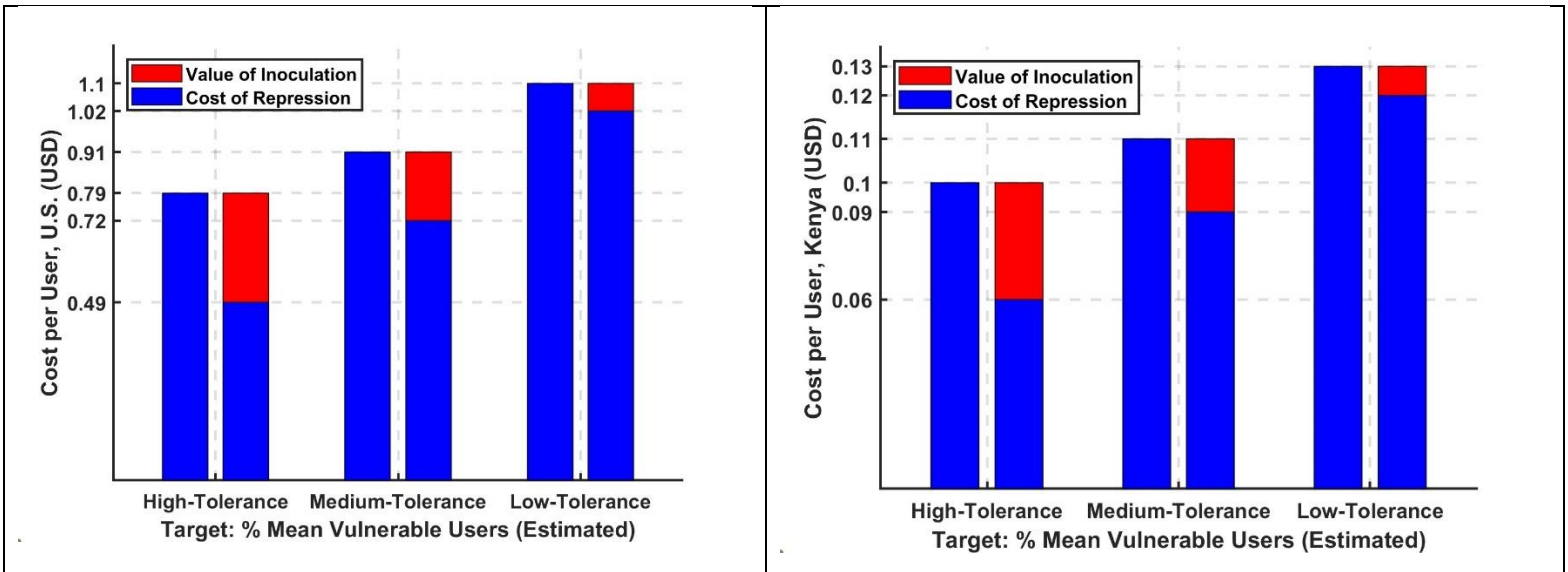

Figure 2. Cost-equivalent policy mixes under three target scenarios as defined in Section 5.2. (Left.) Moderator firm based in U.S. (Right.) Moderator firm based in Kenya. Estimates are built from Model 1 (Section 4.2.1).

Discussion

Our study is the first to demonstrate the economic and structural synergies that result from combining inoculation strategies with the monitoring and removal of online extremists. We find that these strategies are not only effective independently but also significantly more effective when integrated, thereby enhancing overall policy efficacy. The economic viability of this combined approach is confirmed through an analytical cost-benefit analysis, which highlights the scalability and sustainability of inoculation interventions across various socio-political environments. Beginning with realistic benchmarks for the costs of deplatforming, we project these into achievable and realistic estimates for inoculation costs, confirming that the cost per user is reasonable.

We highlight three key findings from this study. Firstly, we have demonstrated that inoculation and deplatforming are synergistic because there is a substitution between the two. Importantly, the substitution is only partial. There are performance levels (i.e., expected number of vulnerable users) that cannot be achieved by inoculation alone; deplatforming is necessary to reach these levels. Secondly, we discovered a unique effect of inoculation at the base of its functioning and its complementarity with deplatforming: it reduces the variance of simulation outcomes. This is very important and noteworthy because it means that inoculation makes a society more resilient to contagion and also makes the effect of moderation more effective and predictable. By maintaining a higher inoculation rate, societal responses become more predictable, thus enhancing the system's resilience as proposed by McGuire's original conceptualisation of inoculation theory (Lewandowsky & Van Der Linden, 2021). This predictability is essential for effectively mitigating the spread of extremist ideologies. Thirdly, we have developed a cost-benefit analysis to demonstrate that it is possible to price inoculation. From this analysis, it emerged that there is an increasing relationship between regulation costs and the reduction in the target of vulnerable users. Additionally, as the target increases, the value of inoculation decreases, indicating less substitutability between policies. We have proposed a straightforward method to assess the cost-benefit of policies that have potentially complex, far-reaching effects, which are difficult to quantify in experimental settings. Our evaluation method can be easily integrated into standard fiscal analysis to assess for example whether and how online inoculation policies should be subsidised, or whether inoculation can be made self-sustainable through fines.

This research has valuable implications for policy. Governments often reduce funding for 'soft' approaches to countering extremism, such as inoculation, especially under budgetary constraints. However, this study provides evidence that inoculation not only complements but also enhances the efficacy of more repressive measures, such as monitoring and removing extremist content. Therefore, rather than reducing support for inoculation strategies, they should be maintained and developed alongside these repressive policies. Our article also shifts the scientific discussion on the design of inoculation policies by demonstrating the importance of assessing the impact of different policies in conjunction. While previous literature has focused on the efficacy of individual policies in isolation (e.g., Ali et al., 2021), our findings reveal for the first time that the effects of policies differ significantly when assessed together, opening a new line of research inquiry into the combined effects and synergies between different approaches. Our analysis underscores that the scope of action is less crucial than the persistence and continuity of inoculation efforts. Targeting a smaller number of individuals consistently over a prolonged duration may be more impactful than focusing on a wider area

of the population for a limited number of periods. This offers a novel perspective on social norms and countering extremism policies.

Although we used synthetic data, the external validity of our models is enhanced by the calibration of the key parameters using data from existing empirical research on inoculation and radicalisation. The robustness of our models, validated through sensitivity testing and Monte Carlo simulations, ensures that the findings are credible and applicable. To enhance the realism of our simulations, we modelled both deplatforming and inoculation interventions as imperfect strategies for the curbing of online extremism. Deplatforming is often hindered by legislative limitations and extremist tactics that circumvent detection, resulting in uneven content removal. Inoculation strategies, while valuable, are restricted by their limited scope, temporary effects, and reduced impact over time. Our findings suggest that integrating these imperfect strategies can decrease the resources required to boost deplatforming effectiveness. These results not only reinforce the foundational principles of inoculation theory but also expand its utility in fostering long-term societal resilience against the proliferation of extremist content.

References

Adeyemi, D. (2022). Facebook content moderators in Kenya to receive 30-50% pay raise, following complaints. Techcabal. Retrieved from https://www.techcabal.com/

Ali, S., Saeed, M. H., Aldreabi, E., Blackburn, J., De Cristofaro, E., Zannettou, S., & Stringhini, G. (2021). Understanding the effect of deplatforming on social networks. In Proceedings of the 13th ACM Web Science Conference 2021 (pp. 187-195). ACM.

Banas, J. A., & Rains, S. A. (2010). A meta-analysis of research on inoculation theory. Communication Monographs, 77(3), 281-311.

Barton, G., Vergani, M. & Wahid, Y. (2022). Understanding Violent Extremism in Indonesia. In Countering Violent and Hateful Extremism in Indonesia: Islam, Gender and Civil Society (pp. 29-62).

Baruch, B., Ling, T., Warnes, R., & Hofman, J. (2018). Evaluation in an emerging field: Developing a measurement framework for the field of counter-violent-extremism. Evaluation, 24(4), 475-495.

Brouillette-Alarie, S., Hassan, G., Varela, W., Ousman, S., Kilinc, D., Savard, É. L., ... & Pickup, D. (2022). Systematic review on the outcomes of primary and secondary prevention programs in the field of violent radicalization. Journal for Deradicalization, (30), 117-168.

Corbeil, A., & Rohozinski, R. (2021). Managing risk: Terrorism, violent extremism, and anti-democratic tendencies in the digital space. In The Oxford Handbook of Cyber Security (pp. 163-180). Oxford University Press.

Dugan, L., & Fisher, D. (2023). Far-right and Jihadi terrorism within the United States: From September 11th to January 6th. Annual Review of Criminology, 6, 131-153.

Hassan, G., Brouillette-Alarie, S., Alava, S., Frau-Meigs, D., Lavoie, L., Fetiu, A., ... & Sieckelinck, S. (2018). Exposure to extremist online content could lead to violent

radicalisation: A systematic review of empirical evidence. International Journal of Developmental Science, 12(1-2), 71-88.

Jackson, A. D., & Patil, S. P. (2022). Phases of small worlds: a mean field formulation. Journal of Statistical Physics, 189(3), 40.

Koehler, D. (2014). The radical online: Individual radicalisation processes and the role of the Internet. The Journal for Deradicalisation, 1, 116-134.

Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. European Review of Social Psychology, 32(2), 348-384.

Lewandowsky, S., & Yesilada, M. (2021). Inoculating against the spread of Islamophobic and radical-Islamist disinformation. Cognitive Research: Principles and Implications, 6, 1-15.

McSwiney, J. (2021). Far-Right Recruitment and Mobilization on Facebook: The Case of Australia. In Rise of the far right: Technologies of recruitment and mobilization (pp. 23-40).

Mudde, C. (2019). The far right today. John Wiley & Sons.

Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. Journal of Applied Econometrics, 11, 619-632.

Roozenbeek, J., Van Der Linden, S., & Nygren, T. (2020). Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures.

Schumann, S., Clemmow, C., Rottweiler, B., & Gill, P. (2024). Distinct patterns of incidental exposure to and active selection of radicalizing information indicate varying levels of support for violent extremism. PLoS ONE, 19(2), e0293810.

Sickles, R. C., & Zelenyuk, V. (2019). Measurement of productivity and efficiency. Cambridge University Press.

Speckhard, A., & Ellenberg, M. (2020). Is internet recruitment enough to seduce a vulnerable individual into terrorism? Homeland Security Today, 8.

Steiger, M., Bharucha, T. J., Venkatagiri, S., Riedl, M. J., & Lease, M. (2021, May). The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1-14).

Stephens, W., Sieckelinck, S., & Boutellier, H. (2021). Preventing violent extremism: A review of the literature. Studies in Conflict & Terrorism, 44(4), 346-361.

Vu, A. V., Hutchings, A., & Anderson, R. (2023). No Easy Way Out: the Effectiveness of Deplatforming an Extremist Forum to Suppress Hate and Harassment. arXiv preprint arXiv:2304.07037.

Weimann, G. (2021). Motivational imbalance in jihadi online recruitment. In The Psychology of Extremism (pp. 280-303). Routledge.

Winter, C., Neumann, P., Meleagrou-Hitchens, A., Ranstorp, M., Vidino, L., & Fürst, J. (2020). Online extremism: Research trends in internet activism, radicalisation, and counter-strategies. International Journal of Conflict and Violence (IJCV), 14, 1-20.

Wolfowicz, M., Hasisi, B., & Weisburd, D. (2022). What are the effects of different elements of media on radicalisation outcomes? A systematic review. Campbell Systematic Reviews, 18(2), e1244.

Wolfowicz, M., Litmanovitz, Y., Weisburd, D., & Hasisi, B. (2020). A field-wide systematic review and meta-analysis of putative risk and protective factors for radicalisation outcomes. Journal of Quantitative Criminology, 36, 407-447.

Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data (2nd ed.). MIT Press.

APPENDIX 1. CALIBRATION

Simulation setup and timeline

We employed *NetLogo* version 6.3.0 to conduct agent-based simulations that span through one year measured in daily increments (i.e. $T = 365$). Each parametric configuration is replicated $m = 1, \dots, , M = 100$ times to robustly explore the behavioural space of the simulation. At time step $t = 0$ each user receives a demographic profile on the ground of which a susceptibility γ is computed. Each user receives the initial prior level of radicalisation B and is attributed a risk state, on the ground of which networks are generated through a simple preferential attachment protocol. Lastly, both inoculators and extremists are randomly injected within the structure. The model follows a discrete pacing $t = 0,1, \dots, T$. Following our calibration on user's activity of Appendix I, for clarity we equate this pacing to a daily frequency. At each time step $t = 1, \dots, T$, the new simulation step begins with the extremists deciding whether to share extremist content to the users (at a rate δ^E) they are linked to through the social network. Similarly, inoculators decide whether to provide the inoculation training. Both these actions follow a homogeneous Poisson process regulated by rates δ^R and δ^I, respectively. Whether an investigation is not already ongoing, enforcers may either start a new investigation (defined by $\delta^I \geq 0$) by pooling a subset of users of size $r^R \geq 0$ or keep investigating the pre-selected pool of suspects. If an extremist is identified within the pool, the extremist is deplatformed (i.e. removed from the system) at the conclusion of the investigation. Each user $i \in N$ updates her own radicalisation level B_i by performing the operations in Equation (1) - (3) conditional on which she will perform one of the following actions: *(a)* share the extremist content, *(b)* share the inoculation training she received (provided she received one), or *(c)* report extremist content she got exposed to (if any) to the authority.

In our model, we assume that a link exists between two users $i, k \in N$ if they are friends on Facebook. We further assume that links are weighted, reflecting the fact that there are different types of bonds between Facebook friends, some weaker and some stronger. The information enters the user's judgment with a persuasiveness magnitude $\phi(\epsilon) \geq 0$ that depends on the mode of contact: whether the extremist contacted the user (i.e., passive exposure to the radicalising content) $\phi = \phi^A \geq 0$, or the user contacted the extremist (i.e., active exposure to the radicalising content) , $\phi = \phi^I$, with $\phi^I \geq \phi^A$ (Schumann et al. 2024). The magnitude stemming from exposure is further filtered by the user's individual susceptibility γ_i and the effect of inoculation, provided she received one. Similarly, the persuasiveness of inoculation, $\iota \geq 0$, impacts the user's belief through two channels, depending on whether the policy is acquired by either forced or voluntary participation to training program. In the former case, the efficacy of the training is given by $\rho(\iota) = \rho^V$, whereas in the latter is given by $\rho(\iota) = \rho^F$, such that $\rho^V \geq \rho^F \geq 0$. Therefore, for any user $i \in N$, internal credibility c_i is obtained as

$$c_i(\epsilon, \iota) \equiv \gamma_i \times \phi(\epsilon) - (1 - \gamma_i) \times \rho(\iota)\,.$$

(1)

As second action, each user consults her social network and adjusts her own evaluation toward the radicalisation level emanating from it. The network influence is mediated by the user's *degree of conformism* $\kappa_i \geq 0$, a parameter[1] setting the speed and extent of adjustment of the user's belief to the realised radicalisation level of her friends. More precisely, for every period $t = 1, .., T$, given $\boldsymbol{B}$ be the matrix storing the radicalisation of each user, the *peer effect* $\alpha(\boldsymbol{B}, \boldsymbol{M}, k_i) \in [-1,1]$ is obtained by taking the weighted *deviation* from the previously realised radicalisation level of i and her friends. As third action, each user updates her own radicalisation level B_i based upon the *net credibility* of the belief β_i, defined as

$$\beta(\boldsymbol{B}, \boldsymbol{M}, \epsilon, \iota, i) = c_i(\epsilon, \iota) + \alpha(\boldsymbol{B}, \boldsymbol{M}, k_i),$$

(2)

and the history of information she received up to that decision node. More precisely, the posterior radicalisation is calculated as (see Lewandowsky et al., 2019)

$$B(\boldsymbol{B}, \boldsymbol{M}, \epsilon, \iota, i) = \frac{P(B|E)}{P(\sim B|E)} = \frac{P(B)}{P(\sim B)} \times 2^{\beta}.$$

(3)

Depending on the posterior realised radicalisation level B, users are partitioned in four classes which we collect under the label of *Risk State*. These states are: *Immune*, *Low-risk*, *Susceptible*, *High Risk* and *Extreme*. For convenience, in the rest of the paper, users within the latter two classes will be collectively referred to as *Vulnerable Users*. Conditional on her current *risk state*, each user picks the fourth and last action with some probability from the following menu: *(a)* share the extremist content, *(b)* share the inoculation training she received (provided she received one), or *(c)* report extremist content she got exposed to (if any) to the authority.

Realistically, in our setting enforcers' efficiency maps into the *time* required for identification and removal of extremists (see Appendix II for an analytical relationship between the monitoring radius r^R and removal time). In order to be effective, the removal of all extremists has to be *timely*, to limit exposure time of users to radical beliefs can spread uncontrolled from individual users to the broader population (see, for an empirical example, Hickey, 2023). As a second assumption, we assume that enforcers will take action only against extremists – that is, users who are radicalised and

[1] impact of friends' consensus is mediated by the closeness of the relationship as measured in terms of mind-likeliness, that is by means of the closeness of their belief to the belief of that specific user (up to a tolerance level).

systematically spread extremist content among users – and not against users who might share extremist content occasionally.

While in this work the relationship between an empirical online network and the social network is stylized and functional for capturing a parsimonious set of empirical features, a first-pass, unweighted mapping between empirical online social network data and computational objects can be easily established even in lack of information about personal online connections. Let $\boldsymbol{U}$ and $\boldsymbol{G}$ be two numbered lists, respectively given by all the empirical users of a generic online ecosystem of Facebook groups, and the online groups they belong to, respectively, and let the binary matrix $\boldsymbol{S}$ with element $s_{ij} \in \{0,1\}$ be the *social network* describing the membership structure existing between any online user $i \in \boldsymbol{U}$ and group $j \in \boldsymbol{S}$. We say that i is a member of group j if $s_{ij} = 1$, that is if a *link* between i and j exists. By construction, $\boldsymbol{S}$ is bipartite (i.e. partitioned in *at least* two sets of nodes, $\boldsymbol{U}$ and $\boldsymbol{C}$) as $s_{ij} = 0$ for $i, j \in \boldsymbol{U}$ and $i, j \in \boldsymbol{G}$. Let $\boldsymbol{G}^T$ be the matrix transpose of $\boldsymbol{G}$. Then, $\boldsymbol{M} \equiv \boldsymbol{G} \times \boldsymbol{G}^T$ is the one-mode network projection of $\boldsymbol{S}$, that is a matrix picking generic element $m_{ik} = 1$ if two users i and k are linked to *at least* one shared community. The equivalence between the empirical one-mode projection $\boldsymbol{M}$ and the artificial network is immediate.

We measure inoculation efficiency by assuming that inoculators target a random subset of the population with a radius of action given by $r^I \geq 0$, and a frequency of action given by $\delta^I \geq 0$, as measured in days of activity per month, such that $\delta^I = 1$ and $\delta^I = \infty$ indicate a daily activity, and no activity, respectively. In principle, inoculation bears two effects. A direct effect given by inoculation training received by users involved in the inoculation programs, and an indirect one, caused by interaction of inoculated and non-inoculated users. Differently from enforcers, inoculators perform only one action, that is they propagate inoculation.

Appendix 1.1. Baseline Calibrations

Users' Activity Monitorable Frequency. Using the data from a large conspiracy Facebook echo-chamber (N= 68,000 users) constructed by Bessi et al. (2015), we find that average user public activity (i.e. liking, sharing or writing public posts) across the echo-chamber is given by 95 items per year. We make the conservative assumption that users in our Facebook group equally split time between public and private engagement on the platform as well as conspiracy and non-conspiracy activities. This gives a total yearly activity of 380 monitorable items, roughly an item per day. Therefore, we align the

model's discrete pacing $t = 0,1, \dots, T$ to a daily frequency and claim that each user performs on average one monitorable activity per period.

Demographics. To anchor our study to a realistic population dynamic on social media platforms, we calibrate our ABM data generating process to reflect Australia's demographic profile on Meta platforms (Facebook, Instagram, Messenger), focusing on age and gender distributions of online users (Ramshaw, 2024). For additional attributes, we referred to demographic data from the Australian Bureau of Statistics 2022 for profile sampling (ABS, 2022). The age and gender distribution among online users, as per our calibration, shows variances across different age groups. For instance, the representation of female and male users aged 13-17 is 2.6% and 2.0%, respectively. This pattern continues across age groups, with the 18-24 age group comprising 9.4% females and 8.9% males, and so forth, indicating a gradual decrease in representation as age increases, with the 65+ age group consisting of 5.3% females and 3.7% males. Regarding education, the demographic spread across age and gender highlights significant educational attainment, especially in the 25-34 age group, where 50.9% of females and 38.3% of males have higher education. This trend gradually decreases with age, reflecting a diverse educational background across the population. Employment status, categorised by age, reveals that unemployment or lack of study is most prevalent among the 65+ age group, with 76.9% not engaged in employment or education, contrasting with lower percentages in younger age groups, such as 7.6% in the 18-24 age group. Marital status data indicates that 47% of individuals over 18 are married, suggesting a significant proportion of the adult population is in matrimonial relationships. Lastly, criminal history, differentiated by age and gender, shows that males have a higher incidence rate (2.482%) compared to females (0.798%) among the 13-65 age demographic, with a median age of 31 for these occurrences. This parametrisation provides a comprehensive demographic backdrop, employing statistical data to mirror the socio-demographic composition of Australia's online community, thereby enhancing the realism and applicability of our simulation to understand social media dynamics.

Susceptibility γ. Based on the theories and findings discussed in the Background section, we parameterised individual susceptibility to extremist content γ. Specifically, we considered a set of observable individual attributes with known real distributions in the population—age, gender, education, marital status, employment status, and criminal record. These attributes constitute risk or protective factors for radicalisation, as identified in Wolfowicz et al. (2020)'s meta-analysis. When a user is created and assigned their attributes, we calculate the odds ratios to determine the user's level of susceptibility, that is, their likelihood of adopting extremist views following exposure to extremist content (Table A.1). We assume a baseline probability of 0.5 for each user, implying users are agnostic on the radical opinion. We call "risk factor" (resp, "protective factor") a factor that increases (resp. decreases) the level of susceptibility away from the baseline.

Attribute	Factor Type	Effect Size	Odds Ratio	Definition
Age	Protective	-0.053	0.825	1 if age > 35; 0 otherwise
Education	Protective	-0.039	0.868	1 if bachelor's degree or above; 0 otherwise
Marital Status	Protective	-0.038	0.871	1 if married; 0 otherwise
Gender	Risk	0.082	1.347	1 if male; 0 if female
Employment Status	Risk	0.042	1.165	1 if unemployed or no study; 0 otherwise
Criminal History	Risk	0.331	3.397	1 if committed crime; 0 otherwise

Table A.1. Susceptibility factors used to calibrate γ.

Initial Radicalisation Score B. Each user is randomly attributed her initial radicalisation score according to the following protocol. Out of N users, 98% of them will receive their belief B according to a normal initial prior belief distribution with mass centred at 0.5 and standard deviation equal to 0.1. This is motivated by the fact that users belong to an agnostic online group and are potentially susceptible to the messaging of far-right extremists. The remaining 2% will receive their own belief from a uniform distribution with support $[0,0.5] \cup [0.95,1]$, thus implying that this 2% of users will form a mass uniformly spread within the *immune* and *extreme* states (see Table X).

Direct exposure ϕ. After being exposed to inoculation and/or extremist content, users assess the salience of the source, respectively given by $\iota \geq 0$ and $\epsilon \geq 0$. Drawing on Schumann et al. (2024), we categorise exposure to extremist content into two types: intentional ϕ^I and accidental, ϕ^A. We propose that the impact of exposure to extremist content halves across these scenarios, with intentional exposure having a greater effect than accidental exposure. Similarly, for inoculation training, forced participation ρ^F carries half of the effect of voluntary training ρ^V. The influence of exposure on users with a high level of susceptibility γ_i is notably more significant for extremist content and notably less for inoculation training.

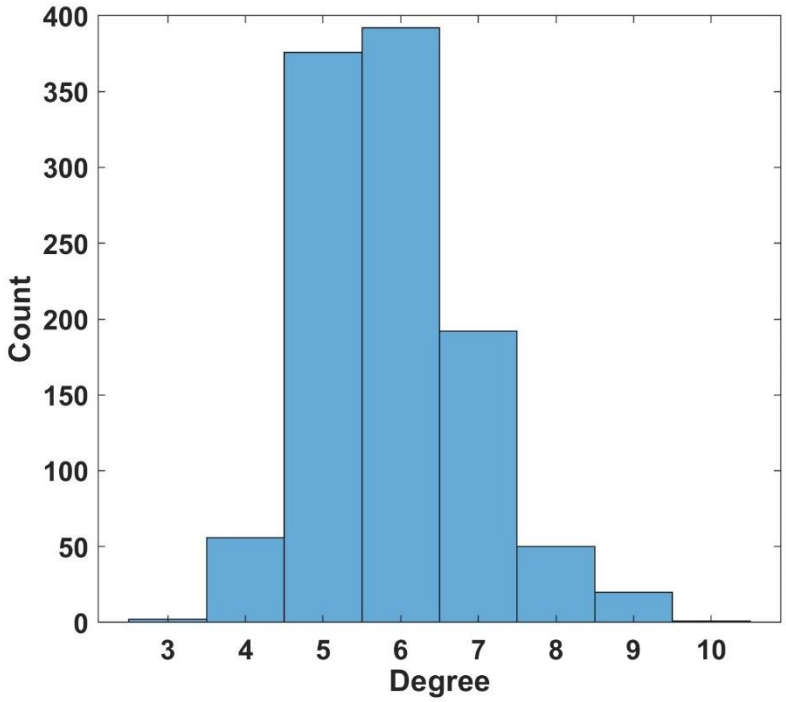

Figure A.1. Degree Distribution of Citizens emerging from the small-world network

Social Network M. The network structure of this work follows the calibration strategy proposed by Del Vicario et al. (2016). In that work, the authors calibrate an artificial ecosystem made of N=5000 agents to match information diffusion in a large empirical Facebook ecosystem. The authors find that the network structure generating information cascades that closely match empirical ones is a small-world network (see Watts and Strogatz, 1998) with small neighborhoods (equivalent to an average node degree of 8) and limited link rewiring (ranging from 0.01 to 0.2), implying a limited dispersion of the degree distribution.

In our work, the social network is created at the start of each simulation series via a configuration protocol (see Jackson, 2008) that attributes each individual a link profile. Links are created using a small-world protocol. To obtain a compatible degree structure given the small population, we distribute the N=1000 nodes in a 33x33 block structure with a clustering coefficient of $K = 2$. This procedure leads to the distribution reported in Figure A.1, where small dispersion centred around a degree mean of 6 is observed.

Network Exposure α. The conformity parameter κ_i regulates the incidence of peer effects on individual decision making. We design this parameter to a binary value, $\kappa_i = \kappa = 0.5$ for $B_i \in [0.05,0.95]$, and $\kappa = 0$ otherwise, to capture both the idea that change is relatively inertial and that very opinionated users (in either direction) will not be sensitive to network opinions, reflecting a core result of empirical literature (see, for a discussion and numerical estimates of preferential inertia, Bessi et al, 2015 and Vicario et. Al, 2016).

Enforcers Frequency δ^R. As explained in the main text, this parameter captures the length (measured in days per months) required to complete one investigation of the

collected pool of users. We proxy this parameter with a standard productivity measure, the average handling time, that is the time required to a moderator to perform her tasks on a single media item (Perrigo, 2022). Facebook employees are expected to maintain an average handling time of 50 seconds per item. This equals an average of 580 moderated items per day (e.g., Perrigo, 2022). Under the assumption that a moderator is required to assess one year and a half of activity for each moderated user in order to make a judgment on deplatforming, it takes roughly 10 days for a enforcer to investigate 1% of the user population.

Enforcers Radius r^R. Enforcement radius of action indicates the pool of users assessed in each investigation and captures the precision of repressive intervention within the network. This parameters is one of the main dimensions of investigation of our experiments. For this reason, to conduct our exploration, we will let r^R free to vary within a positive interval range.

Inoculators Frequency δ^I. With the enforcer radius r^R, inoculator frequency is a main dimension of investigation of this analysis and is allowed to span between $\delta^I = 10$, indicating three inoculation events per month and $\delta^I = 1$, indicating one inoculation event per day. In the analysis, we also include the case of no inoculation (equivalent to zero inoculation events).

Inoculators Radius r^I. Inoculator radius is conservatively calibrated to $r^I = 1$.

Inoculation salience ι. Based on previous experimental research findings, we assume that the average effect of inoculation is 0.04 (Lewandowsky and Yesilada, 2021).

Users' Actions. User's actions performed at the end of each period are context and belief-dependant. If a user encounters inoculation treatment at any time-step $t = 1, \dots, T$ with a belief score below 0.5, they have a 20% likelihood of sharing the inoculation content. When exposed to extremist content at time-step t, in the absence of enforcers intervention, and with a belief score of 0.5 or higher, there is a 20% chance of the user disseminating the extremist content, which decreases to 16% if they have previously received inoculation training. For those identified as 'extreme' and exposed to extremists at time-step (t), the model assumes a 100% likelihood of sharing the extremist content. Additionally, users in a *immune* state who are exposed to extremist content shared by their network at time-step *t* have a 5% probability of reporting the source of the extremist content to enforcery authorities.

Other parameters. All parameters are stored in Table A.2 below.

Variable	**Value/ Range**
Number of users N	1000
% of users with "immune" as initial status	1%

% of users with “extreme” as initial status	1%
% of influencers	1%
Mean of initial belief	0.5
Probability of share inoculation training to network	20%
Probability of share extremist content to network	20%
Probability of reporting extremist content	5%
Distribution of number of friends for user	$TruncNorm\ (5, 2.5, 0, 10)$
Distribution of link weight	$Norm\ (0.5, 0.5)$
Probability of active participants	10%
Number of extremists	1
Exposure to extremist content effect size	0.02
Frequency of extremists sharing extremist content	1 per day
Decay rate to extremist content	0.9
Inoculation effect size	0.04
Decay rate to inoculation training	0.9
Long term effect of training on individual susceptibility	Reduce S by I every 10 sessions
Tolerance for reporting before block a user	5

Table A.2. Parameter Calibration.

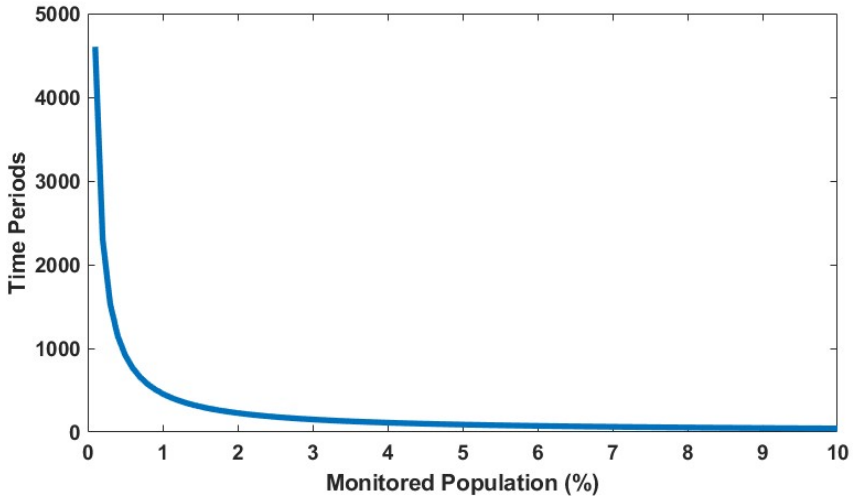

Figure A.2. The relationship between the enforcer radius r^R and the time required to identify and deplatform one single extremist with probability 99% from an online group made of $N = 1000$ users under a random identification protocol.

APPENDIX 2. MATHEMATICAL RESULTS

Appendix 2.1. The connection between Enforcer Precision and Time to Removal

In this short appendix we operationalize the enforcers' radius of action r^E, used in our model as a simple proxy for police's precision, into an empirically-relevant measure of police efficiency: the length of investigation. The mapping is useful to derive a factual measure of police efficiency that can be readily calibrated on data.

With no loss of generality, assume an online Facebook group of N=1000 nodes, one enforcer and one extremist and assume that $\delta^R = 1$ to remove notational cluttering. Each period $t > 0$, the enforcer randomly draws $r^R \geq 0$ users from the ecosystem (possibly, redrawing the same individual multiple times). Therefore, given p the probability of identifying and deplatforming the extremist in one single round, the probability not to identify and deplatform the extremist in one single round is given by

$$1 - p = \frac{N-1}{N} \times \frac{N-2}{N-1} \ldots \times \frac{N-r-1}{N-r} = \frac{N-r}{N}.$$

The probability to draw the extremist within the $t = \bar{t} \geq 0$ rounds is then given by

$$1 - \left(\frac{N-r}{N}\right)^{\bar{t} \times r^R}.$$

Hence, the 99% probability of catching the extremist within $t = \bar{t}$ periods is equal to

$$1 - \left(\frac{N-r}{N}\right)^{\bar{t}} \geq 0.99,$$

Which can be immediately rewritten as

$$(\bar{t} \times r^R) \log\left(\frac{N-r}{N}\right)^{\bar{t}} \leq \log(0.01)$$

Or

$$\bar{t} \leq \frac{\log(0.01)}{\log\big((N-r)/N\big)}$$

For example, with a radius of $r^R = 10$, the enforcer will catch the extremist with a probability of 99% after a *maximum of* $\bar{t} = 456$ draws. In Figure A.2 we report a graphical representation of the relationship between the radius r^R and the maximum number of periods required to guarantee removal of one extremist out of a population of $N = 1000$ users for the interval $r^R \in [1,100]$. From the example, the parameter r^E can be easily calibrated on an empirical measure of actual investigation times.

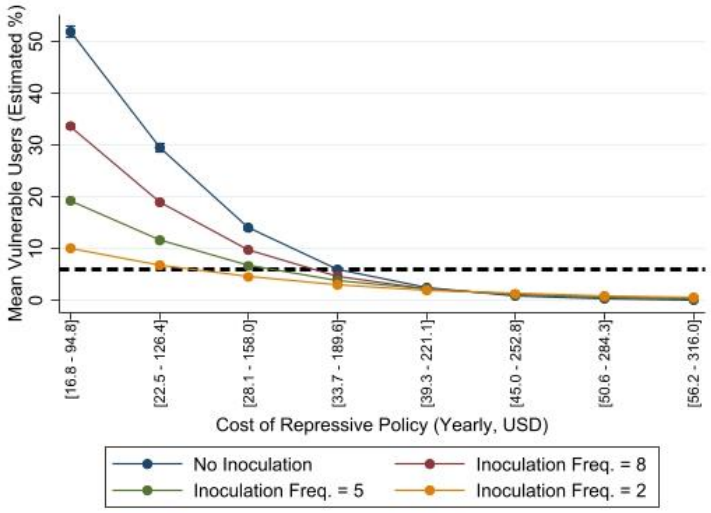

Figure A.3. Cost-benefit Analysis.

Appendix 2.2. Construction of the Benefit-Cost Analysis

We construct the cost-benefit analysis of Section 5.2 as follows. First, we densify the space of policy mixes (r^E, δ_I) by reducing the discrete step between policies to 1/3 of units. This way, we replace the original 11×11 policy mix space (r^E, δ_I) with a 110×110 space. Second, we use the estimates of Model 1 (Section 5.1) to compute the marginal effects of the mean fraction of vulnerable users for each of the $110 \times 110 = 12{,}100$

policy mixes. Third, we compute cost-equivalent policy mixes under three alternative scenarios as described in the main text. Through this methodology, it is possible to establish an cost-based equivalence between policies and assess the monetary value of inoculation. To provide an intuition, in Figure A.3 we plot the estimated fraction of vulnerable users as a function of the per-user cost of monitoring, under four alternative inoculation policies (respectively: no inoculation, $\delta^I = 2, \delta^I = 5$, and $\delta^I = 8$). In the figure, an investigation radius ranging from $r^R = 3$ to $r^R = 10$ is considered.

For example, from the figure we note that an yearly expenditure between \$33.7 and \$189.6, depending on whether the moderators are based in Kenya or the U.S., generates an expected mean fraction of vulnerable users of 8%. A similar fraction can be obtained by adopting either an inoculator frequency of $\delta^I = 5$ and spending for monitoring between \$28.1 and \$158.0 or an inoculator frequency of $\delta^I = 2$ and spending for monitoring between \$22.5 and \$126.4. This implies that the rational cost of an inoculation policy designed to be deployed every 5 days per month has to be no larger than \$5.6 or \$31.6, depending on the contractor's country, whereas the rational cost of an inoculation policy designed to be deployed every 2 days per month has to be no larger than \$11.20 and \$63.20. This translates into a per-inoculation cost of \$0.93 to \$26.3 in the first case and \$0.74 and \$4.21 in the second case.

The comparison of per-inoculation costs reveals an important consequence of structural synergies: c*eteris paribus,* policies are subject to decreasing returns: given two equivalent policy mixes, gains from expanding the use of one policy and, at the same time, reducing the use of the other one, reduce, as the expanding policy is incremented[2], suggesting the idea that heterogeneous mixes of policies fare better that the use of one strategy alone.

[2] Decreasing returns are visually noticeable from comparison of distance between the pair of curves characterized by no inoculation and an inoculation frequency of $\delta^I = 8$, and inoculation frequency $\delta^I = 5$ and $\delta^I = 2$, respectively.