Inoculation Against Online Extremism: An Agent-Based Simulation and Cost-Benefit Analysis

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

Abstract

Objectives. Extremism and its active control heavily impact the functioning of Western democracies and their budgets. Online environments provide a platform for extremist groups to spread their ideologies and recruit new members. Despite the efforts of tech companies and security agencies, there is growing frustration over their inability to curb online extremism through repressive policies such as the removal of extremists via deplatforming. There is an urgent need to devise better policies in this area. Prior studies have shown that inoculation—a preventative approach that involves exposing individuals to weakened forms of an argument so they can develop resistance to it—can significantly decrease both the willingness to share and agreement with online extremist content. No research has explored the effects of combining repressive and preventative policies. This research seeks to address this gap by identifying the existence of synergies between inoculation and deplatforming in the form of a net reduction in the resource load required to make the latter effective.

Methods. The present study develops an agent-based model of diffusion and control of online extremism informed by inoculation theory and calibrated using data from systematic reviews and empirical studies conducted in the field of terrorism research. 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. Agents decide to either embrace or refute it based upon their own characteristics, the influence exerted by extremists and the belief chosen by their peers. We introduce a standard top-down deplatforming intervention and a simple social-driven inoculation strategy finalised at controlling the spread of the extremist opinion. We develop a cost-benefit methodology to compare costs of various baskets of policies and realistically assess outcomes.

Results. First, we find strong evidence that inoculation is synergetic to deplatforming: for given levels of effectiveness, inoculation can *partially substitute* deplatforming. The mechanism is that inoculation stabilizes opinion formation through the network thus making the containment of early outbreaks of social contagion less arbitrary. Second, we construct a "possibility frontier" (Sickles and Zelenyuk, 2019) between the two policies. We identify the minimum amount of inoculation required to control the spread of extremism for each additional dollar spent on repression. Importantly, we demonstrate that a situation where both policies are simultaneously deployed dominates over the use of one policy in isolation. We argue that the cost-benefit analysis can help identifying the

optimal mix of policies that minimizes the cost of controlling the spread of online extremism.

Conclusions. There is strong ground for democratic societies to enforce efficient and transparent policies against online extremism. These measures should combine both repressive and preventive strategies, tailored to the level of susceptibility within the society to online extremism. Cost-benefit analysis highlights the efficiency advantages of having a comprehensive policy approach encompassing both types of measures.

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

1. Introduction

Extremism and its active control put a heavy toll on the functioning of western democracies and their budget. For example, in 2018-2019, North America had the largest percentage increase in the economic cost of terrorism, increasing by 44.9 per cent from the previous year, the equivalent of $165.9 million (Morgan et al., 2021). The region's deterioration was driven by the United States, which recorded an increase of $297 million, or 125 per cent from 2018, as deaths from terrorism rose from 16 to 35 (Morgan et al., 2021). The primary driver of this rise was an increase in far-right terrorism. More recently, far-right groups have also viewed the pandemic as an opportunity to fuel existing narratives with a rise in racist, anti-Semitic, Islamophobic or anti-immigrant hate speech.

Online environments provide a platform for the dissemination of extremist content to potentially susceptible audiences and the recruitment of new members into extremist organisations. Although radicalisation very rarely takes place exclusively online (Winter et al., 2020), individual accounts of terrorists indicate that online activities, including engaging with extremist propaganda and networking with other extremists, can significantly influence the formation of their identity and ideology (Koheler, 2014). Research suggests that exposure to online extremist content does not universally escalate support for extremist ideas (Hassan et al., 2020). Instead, it can increase support among individuals who share ideological affinities and a salient group identity with extremists (Drevon, 2016). This poses a specific risk with far-right extremism in Western contexts, where the audience potentially sympathetic to these views is larger compared to audiences for other forms of extremism like jihadist extremism. For this reason, a significant concern for security agencies in Western countries is the utilisation of the internet by far-right extremists for social interactions and political activism, including spreading propaganda and recruiting new members into extremist groups (Baele et al., 2023).

In reaction to government pressure to address violent extremist and terrorist activity online, four of the largest technology companies—Google, Facebook, Microsoft, and Twitter—launched the Global Internet Forum to Counter Terrorism (GIFCT) in June 2017 (Thorley & Saltman, 2023). The GIFCT's expressed purpose was to prevent terrorists from exploiting these companies' platforms and those of other members, many of which are either their subsidiaries or substantially smaller platforms. Despite the exponential growth of industry's monitoring investments (put a $ figure and maybe a note on indirect costs related to recent/hot/costly controversy on the precarity/wellbeing of African-resident contractors hired as online moderators), monitoring is inefficient: outcome documents of ad-hoc G7 Interior Ministers' meetings show substantial government frustrations with industry in general and the GIFCT members in particular because - as these policies are based upon high-intensity monitoring and ad-hoc removing of extremists from online platforms via deplatforming and/or arrest – they are costly and prone to inefficiencies (Corbeil & Rohozinski, 2021). It is then important to identify public-driven cost-effective strategies that can enhance and complement existing ones.

This article studies the effects of the introduction of a preventative policy called 'inoculation' on top of a common repressive policy - deplatforming – involving monitoring and removal of extremist content from online platforms. Preventative approaches like inoculation—often undertaken by civil society organisations funded by private actors or government agencies—encompass public education, training, outreach activities, and strategic communications targeted at audiences defined as 'at risk' of radicalisation (Barton et al., 2021). Such interventions are underpinned by inoculation theory, which suggests 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). We show that the introduction of an inoculation strategy on top of a deplatforming strategy may reduce the resource load to make the latter effective. Importantly, in this article, we acknowledge the imperfections of both repressive and preventative strategies. Repressive measures face challenges due to legislative constraints and tactics employed by extremists to evade online detection, leading to inconsistent removal of extremist content from social media. Likewise, preventative measures are constrained by their limited reach, temporary impact, and diminishing effect over time. This paper shows that combining an imperfect inoculation strategy with an imperfect repressive policy can reduce the resources needed to enhance the effectiveness of the latter.

2. Background

2.1. Definitions and scope

Scholars have long debated the definitions of extremism and radicalisation (Peels, 2023). In the context of contemporary Western democracies, extremist ideologies are usually characterised by the disregarding of the life, liberty, and human rights of others (Neumann, 2013). To define the concept of radicalisation, it is useful to distinguish between "behavioural radicalisation" (engagement in extremist actions) and "cognitive radicalisation" (support for extremist views), highlighting that extremist views do not always result in extremist actions (Vergani et al., 2020). In this study, we focus on cognitive radicalisation. 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. These ideas are often advocated by organisations that effectively use online platforms to recruit new members through spreading extremist propaganda and forging personal relationships with online users. As we know from personal accounts of former extremists and studies on online recruitment (Speckhard & Ellenberg, 2020; Weimann, 2021), recruiters often integrate themselves into thematic 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.

2.2. Exposure to extremist content online

Research shows that exposure to extremist content is not appealing to general viewers (Hassan et al., 2020). For example, Baines et al. (2010) and Sikorskaya (2017) explored reactions to Islamist propaganda, finding that such videos generally did not resonate with broad audiences. Rather, exposure to extremist content online might increase support for extremist views among people who share a salient identity and ideological leanings with the extremists: for example, online extremist content tended to elicit positive reactions in Muslim Kyrgyzstan's youth, trusted due to familiar sources, language, and Quranic quotes. Koehler (2014) interviewed eight former far-right extremists and found evidence of the Internet's role in radicalising them, providing a platform for communication and ideological exchange. Lee and Leets (2002) found varying persuasiveness of white supremacist material on US adolescents, influenced by pre-existing attitudes.

The idea that pre-existing attitudes influence the relationship between exposure to extremist content and radicalisation is further supported by research distinguishing between actively seeking vs accidental exposure to extremist content. Schumann et al. (2024) observed that the mode of exposure to extremist content could influence the level of support for extremism. Specifically, they noted that individuals with incidental exposure (high probability of incidental exposure, moderate probability of active exposure) showed a mean increase in support for extremism of 0.62 compared to those with no exposure. In contrast, those exposed through a combination of active and incidental exposure (high probability of both active and incidental exposure) exhibited a mean increase of 2.44 in support for extremism compared to the group with neither active nor incidental exposure. This suggests that users who have pre-existing attitudes that drives them towards extremist content are more likely to radicalise as a result of the exposure.

2.3. Inoculation against online extremism

Inoculation theory, conceptualised by McGuire in the 1960s, draws a parallel between the process of immunisation against diseases and the development of resistance to persuasive arguments (Lewandowsky & Van Der Linden, 2021). It proposes that exposure to a weakened form of an argument can stimulate a cognitive-motivational process similar to the body's production of antibodies, thereby increasing resistance to future persuasion attempts. This 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. This method is designed to strengthen an individual's ability to counteract future persuasive attacks by teaching them how to argue

against opposing viewpoints. 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. A notable meta-analysis by Banas and Rains (2010), reviewing 40 studies involving over 10,000 participants, found a medium effect size for inoculation interventions, highlighting their significant impact.

Two randomised controlled trials tested the effects of inoculation to counter cognitive radicalisation outcomes. Braddock (2019) used inoculation to examine its effectiveness in diminishing the persuasiveness of extremist content among American adults. The study's final sample comprised 357 participant, employing a 2 × 2 between-subjects experimental design to assess reactions to extremist left-wing vs. right-wing propaganda, considering the source of inoculation (researcher vs. former extremist). Findings indicated that inoculation increased counter-arguing against extremist messages ($t(355) = 2.99, p < .01$) and marginally increased anger ($t(355) = 1.95, p = .08$), reduced perceptions of the extremist group's credibility ($t(355) = 2.62, p < .01$), and decreased intentions to support the extremist group ($t(355) = 2.41, p < .01$). Structural equation modelling supported these results, although the direct effect of inoculation on support intentions was not significant, suggesting indirect effects via psychological reactance and perceptions of credibility.

Lewandowsky and Yesilada (2021) examined the effectiveness of inoculation techniques against cognitive radicalisation, specifically Islamophobic and radical-Islamist content. The study, employing a 2 × 2 between-subjects design with 591 UK participants, found that those who received inoculation exhibited significantly lower tendencies to perceive extremist content as reliable, to agree with extremist content, and to share such content, compared to the control group. The inoculation effect was quantified with modest effect sizes, notably in reducing the perceived reliability ($\eta^2 = 0.024$) and agreement with extremist content ($\eta^2 = 0.009$). According to this study, the mean support to Islamophobic and radical-Islamist extremist content for participants in the inoculation treatment group reported 4% less than the control group. This reinforces the idea that cognitive defenses can be strengthened through prior exposure to and refutation of misleading arguments, extending the applicability of inoculation theory beyond traditional domains and suggesting a broad-spectrum utility in combating various forms of extremism.

2.4. Risk factors of radicalisation

There is a vast literature in terrorism studies that looks at what factors predict an increased likelihood that an individual will radicalise (Vergani et al., 2020). Wolfowicz et al. (2020) conducted a systematic review and meta-analysis to evaluate risk and protective factors associated with radicalisation outcomes, distinguishing between cognitive and behavioural radicalisation outcomes. Through the analysis of 57 publications, the review identified 62 individual-level factors that predict cognitive radicalisation outcomes, and found that the most traditional criminogenic factors like low self-control, thrill-seeking, and specific attitudinal factors (e.g. possessing already

radical attitudes) exert more significant influence on cognitive radicalisation outcomes (Wolfowicz et al. 2020). The review also shows that some demographic factors appear to have significant – albeit small – relationships with cognitive radicalisation outcomes. Education exhibits a protective effect with a z score of -0.096 and a CI of -0.152 to -0.040, indicating higher levels of education decrease the likelihood of radical intentions. Age is also a protective factor, with a z score of -0.228 and a CI of -0.297 to -0.160, suggesting older individuals are less prone to radicalisation. Unemployment presents a z score of 0.061 and a CI of -0.016 to 0.139, implying a potential, albeit uncertain, risk factor for radicalisation due to the CI crossing zero. Male gender has a z score of 0.189 and a CI of 0.087 to 0.290, showing being male is associated with a higher risk of developing radical intentions. Criminal history shows a modest risk effect with a z score of 0.073 and a CI of 0.018 to 0.127, indicating a slight increase in radicalisation risk associated with antisocial personality traits (Wolfowicz et al. 2020).

3. STUDY

3.1. Purpose of the current study

In this article, we explore the potential of integrating two different approaches to mitigate the spread of extremism online: a preventative approach (that is, a simple inoculation intervention) and a standard repressive approach (that is, the removal of extremist users from the platform). We perform the exploration by designing and implementing an agent-based model (ABM), a technique[1] increasingly used in criminology to simulate complex scenarios in a virtual environment with autonomous agents following realistic, yet identifiable behavioural rules (Wooditch, 2021; Calderoni et al., 2021). The technique involves the construction of an artificial society mimicking a real one for aspects that are relevant for the phenomenon under study. In this work, the reference society is an online community made of individuals with a shared interest and political identity broadly aligned with 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. 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. Within this Facebook group, we study the diffusion of extremist beliefs and the effect of a mix of inoculation and repressive policies to control their diffusion.

[1] ABMs are grounded in the principle of parsimony, emphasising the simplification of complex systems to include only essential elements pertinent to the research question at hand. This approach not only enhances computational efficiency but also improves the interpretability of the model. While this technique is commonly used in fields like public health to study questions about inoculation against disease spread (e.g., Faucher et al., 2022), its adoption for the study of extremism and radicalisation is still at its infancy with rare exceptions (e.g., Pepys et al., 2020).

We proceed through two exercises. In the first exercise, we investigate the mechanism of action and effectiveness of these two structurally diverging policies and their interactions. Obtaining data on the efficacy of measures to combat online extremism is challenging due to various factors, including ethical concerns surrounding the design of randomised controlled trials, limited access to research data due to lack of transparency by social media platforms, and reluctance to participate in research among online users, particularly those with extremist ideologies. We overcome the lack of empirical observations through the data generating process of the ABM. 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.

This bring us to our second exercise, where 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[2]. 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 regulators and policy designers on efficient cost planning of inoculation policies: the *maximum* cost that a rational regulator *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.

This pricing exercise provides regulators with an operative principle that overcomes a major limit of inoculation policies: the elusiveness of cost-benefit analyses in this dimension. As inoculation policies in the field of countering online extremism are specific to the application at hand[3] and features of the target population, a direct "horse-race" between the cost-benefit ratios of repressive vs inoculation policies is not possible: within one same legal and program framework the dynamic nature of the implementation of these policies makes an ex-ante assessment of their cost structure challenging even for program designers and clients (Harris-Hogan, 2020).

---

[2] Operative references on cost-benefit analysis in the dimension of crime control include Chalfin, A.J. (2010), Dossetor, K. (2011) and Hornick, J.P., Paetsch, J.J., & Bertrand, L.D. (2000). Marsh, K. (2010) provides an extensive literature on economic evaluation of criminal justice interventions.

[3] Government around the world embed ad-hoc inoculation strategies within Violence Control Programs (VCP), which are resource, space and time dependant. For an example of the complex nature of VPCs with an application to Australia, we refer the reader to Chapter 5 of *Review of Australia's Counter-terrorism Machinery* (Department of the Prime Minister and Cabinet, 2015). For an evaluation of various inoculation strategies within the Australian VPC, we refer to *Allen, NSW Countering Violent Extremism Program Evaluation* (2019).

3.2. Model

We consider a scenario where far-right extremists try to radicalise others within one anti-immigration Facebook group. Importantly, in this study we focus on the process of cognitive radicalisation, that is, the development of support for extremist views. Our proposed model consists of four agent types: users, extremists, inoculators, and enforcers that interact along discrete time iterations, $t = 1, \dots, T$.

Notation. In the following we suppress time indices for readability and indicate scalar quantities (respectively, vectors/matrices) with normal (respectively, bold) font. Details on the behavioural assumptions and calibration are contained in Appendix I.

Online Users. The model is populated by $i = 1, \dots, N$ individuals inter-connected through different layers. They are all online users of Facebook and members of one anti-immigration Facebook group. Additional to the online group, users are linked to each other through personal friendship connections, thus establishing a social network. 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), based upon their own individual features and peer interaction. 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, users 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 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[4] $\boldsymbol{M}$. Specifically, 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

[4] 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.

different types of bonds between Facebook friends, some weaker and some stronger. To generate a network, we adopt a simple small-world protocol (see Jackson, 2008) using the calibration of Aiello et al. (2016) of real-world Facebook group networks.

The social network is one of the three channels of information flow accessed by users throughout the online group, the remaining two being direct exposure to extremists and/or inoculators.

In every period, users might receive extremist content through private messages on Facebook. The user will then individually assess the *internal credibility* $c_i \in [0,1]$ of the content as follows. Formally, let $\epsilon \geq 0$ be the *saliency* of the extremist's opinion. In psychology, a stimulus is salient when it attracts attention involuntarily: salient messages increase the prominence of some attributes while distracting users from others (Bordalo, Gennaioli and Shleifer, 2022). In a large longitudinal analysis of online communication and radicalisation dynamics, Schulze, Hohner and Rieger (2022) find that extremist narratives are opportunistic and mimetic to the public discourse, thus signalling an intention of their proponents to manipulate the message's saliency and maximize diffusion.

To focus on the interaction between individual decision-making and peer effect in information diffusion, we fix the level of salience across extremist messages, and assume that the opinion enters the user's judgment with a 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 $\phi^I \geq \phi^A$. This distinction reflects Schumann et al. (2024)'s findings on the effects of active and passive exposure to extremist content on the radicalisation process. Importantly, 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.

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

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

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[5] 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

$$\beta(\boldsymbol{B}, \boldsymbol{M}, \epsilon, \iota, i) = c_i(\epsilon, \iota) + \alpha(\boldsymbol{B}, \boldsymbol{M}, k_i), \tag{2}$$

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)

$$B(\boldsymbol{B}, \boldsymbol{M}, \epsilon, \iota, i) = \frac{P(B|E)}{P(\sim B|E)} = \frac{P(B)}{P(\sim B)} \times 2^{\beta}. \tag{3}$$

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 (See Table 1 for the mapping).

Risk state	*Immune*	*Low-risk*	*Susceptible*	*High Risk*	*Extreme*
Radicalisation State	Rejects extremist views		Uncertain	Accepts extremist views	
Radicalisation Score	< 0.05	<= 0.33	0.33 < B < 0.66	>= 0.66	0.95

Table 1. The mapping between the belief state and the radicalization score.

[5] 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).

Law Enforcers. These are online moderators broadly defined (e.g. contractors, police, antiterrorism), that is individual actors with full access to users' online activity that moderate the network and identify and deplatform extremists. We impose two realistic procedural rules to the behaviour of law enforcers. First, identification of extremists goes through two parallel channels: user reporting and direct monitoring. Importantly, 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 is scarce and ad-hoc to specific online contexts (Baruch, Ling and Hofman, 2018), an exponential trend of raising storage, machine and human-based costs (see ACMA, 2023), coped with a

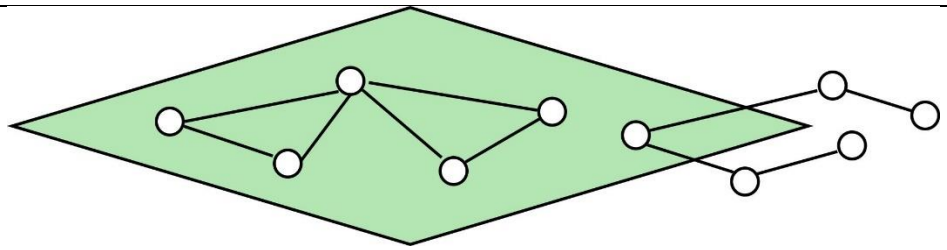

Figure 1. Ten individuals placed in a two-dimensional physical space characterized by a geographically-bounded inoculation policy (identified by the green polygon) and the structure of their social network.

progressive expansion of public budgets and stricter regulations in the dimension of online policing (Winter et al., 2020) signals that current practices are imperfect[6] due to various factors, including overloading. We map resource loading and monitoring efficiency into measurable, data-available quantities by assuming the following monitoring protocol: in every period t enforcers can only monitor a subset of nodes $r^R \geq 0$. We introduce inefficiency by assuming that the set is *randomly* picked across the user base of the whole 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.

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 law enforcers will take action only against extremists – that is, users who are radicalised and

[6] For measurements of efficiency of online moderation, as well as recent legislative advancements on mandatory online moderation we direct the reader to Schneider and Rizoiu (2023) .

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

Extremists. Extremists are individuals who spread extremist content among users. Extremists are assumed to be randomly embedded within the online group by means of which they gain exposure to users and – through Facebook friendship – they obtain the ability to message them privately. To keep the predictions of our model tractable, we impose that extremists are heterogeneous only in their positioning in the social network. In every period t, until removed by law enforcers[7], 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 delivering education-based interventions in offline settings such as schools, sport clubs, churches, etc) vehiculating inoculation policies to individuals that may belong to the anti-immigration Facebook group. After deployment, their position is fixed along the whole simulation. We parametrise inoculators as external to the network of the Facebook group because they represent real-world initiatives that affect dimensions of social interactions that are only partially projected in the social network considered in this work[8] (see Figure 1).

This modelling choice is adherent to real-world inoculation policies such as the recent mix of online and community-based measures implemented by New South Wales government of Australia in 2015 for counteracting violent extremism (see Harris-Hogan, 2020), but it is not directly capturing network-injected inoculation strategies such as groups run by debunkers or publicly-funded think tanks tasked with the goal of counter-informing users connected to these groups.

Similar to law enforcers, inoculation is imprecise and it has limited direct reach[9] of the user community. Symmetric to the law enforcer monitor efficiency, 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 law enforcers, inoculators perform only one action, that is they propagate inoculation. Each message carries a salience of $\iota \geq 0$.

---

[7] This is equivalent to a simple homogeneous Poisson process with a daily spread rate $\rho \geq 0$.

[9] For a discussion and examples on the relationship between the geographical scale of empirical inoculation policies and their social reach with an application to Australia, we refer the reader to Mehrton, A. (2019).

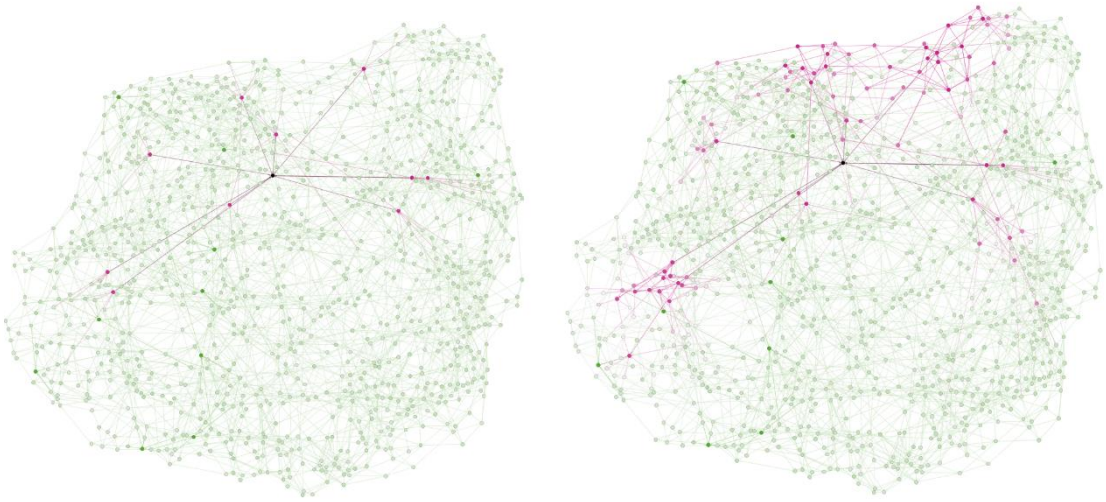

Figure 2. An example of belief dynamics on the network for a population of N=1000 users, one enforcer (not depicted), one inoculator (not depicted), one extremist and 10 users in extreme state. Nodes are colored according to the radicalisation level ranging from green (uncertain state) to red (extreme state). The extremist (at the center) is colored in black. *(Left Panel.)* Radicalisation level at t= 0. *(Right Panel.)* Radicalisation level at $t = T = 720$.

3.2.1. 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, law enforcers may either start a new investigation (with modulo 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.

4. ANALYTICAL STRATEGY AND HYPOTHESES

4.1 Assumptions and Baseline Calibrations

Let $(r^R, \delta^R, r^I, \delta^I)$ be a *policy mix*. To overcome the lack of empirical data on the interaction between inoculation and repression and isolate potential synergies, we simplify the definition of policy mix by conservatively maintaining a very small inoculation radius, $r^I = 1$, thus obtaining a workable lower bound on the efficacy of the inoculation policy. To make the model more tractable, following the data-driven calibration of moderators' productivity described in Appendix 1, we fix the investigation length throughout the simulations to $\delta^R = 1$ day per regulator (measured at full-time 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 a bag 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 $\delta^R = 10$ day. 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).

4.2. Experiments

4.2.1 Measuring Performance and Structural Synergies

To evaluate the effectiveness of the alternative policing strategies and test the existence of structural synergies, we take a parametric approach and run two sets of regressions on two performance measures. Let $n_m\ (t) \geq 0$ be the fraction of vulnerable users, that is users in high risk or extreme state in period t (corresponding to an individual belief $B \geq 0.65$) in simulation run $m = 1, \dots, 100$. 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 $T = 365$. The combination of these measures allows us to indirectly assess critical features of the diffusion dynamics: particularly, the maximum number of vulnerable agents observed within any simulation and the time required by regulators to curb the spread of disinformation. 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, 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 average effect of policy mixes (r^R, δ^I) across their full support. 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, which we refer to as $\bar{n}$ and $\bar{n}\,(T)$. To account for cross-simulation heterogeneity, we augment each the 121×2 performance averages by one standard deviation[10], respectively given by $\bar{\sigma}$ and $\bar{\sigma}\,(T)$ as obtained from the $M = 100$ simulations. Second, we split the configurations in three groups, *Policy Mix Set I*, *II* and *III*, according to the augmented performance above. These sets refer to mixes generating a combined fraction of individuals in vulnerable state above 5%, between 5% and 3% and below 3%, respectively. Third, to obtain the graphical description of effects, we map the three policy sets in the 11×11 policy mix space (r^E, δ_I) against the augmented average performance measures, $\bar{n} + \bar{\sigma}$ and $\bar{n}\,(T) + \bar{\sigma}\,(T)$.

Last, we explore the mechanism behind the functioning of each policy. To do so, we pick three mixes belonging to a common policy set and describe the temporal evolution of the fraction of vulnerable users $n_m\ (t)$ for each of the $m = 1, \dots, 100$ simulation across the $t = 1, \dots, 365$ periods. For clarity, we compute and compare the associated per-period average measures $\bar{n}(t)$ and $\bar{\sigma}(t)$.

[10] In untabled simulations we perform the analysis on the non-augmented performance measures (1) and (2) and find qualitatively similar results.

4.2.2 Cost-Benefit Analysis

To assess whether structural synergies can translate in economic synergies we conduct a cost-benefit analysis of policy mixes. We use the cost-benefit analysis to extrapolate the maximum value (hence, the rational cost) of inoculation policies as measured in terms of reduction of monitoring cost. We do so by mapping the cost structure of Section 4.1 into an analysis of the marginal effects of policies as extracted from the regression models of Section 4.2.1.

	Effect on $\bar{n}_m$	Effect on $n_m\ (T)$
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 2. *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.

4.2.3 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 $M = 100$ repetition). Second, the parametric space of critical parameters δ^R and r^I is explored in the neighbourhood of the chosen values to confirm the qualitative structure of results.

5. RESULTS

5.1 Results: Measuring Performance and Structural Synergies

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[11] 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 2, 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 each performance measure. Following the procedure of 4.2.1, we split the configurations in three groups, *Policy Mix Set I*, *II* and *III*, according to the average augmented performance measures factoring standard deviation.

	Effect on $\bar{n} + \bar{\sigma}$	Effect on $\bar{n}(T) + \bar{\sigma}(T)$
Policy Mix Set I	0.342***	0.166***
	-0.04	(0.02)
Policy Mix Set I	-0.303***	-0.139***
	-0.04	(0.02)
Policy Mix Set III	-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 3. Average difference in performance measures (augmented by one standard deviation) between the three sets of policy mixes. 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.

[11] The use of the inverse has neutral effects on the model structure at the basis of Table 2 and simplifies the interpretation of estimates.

Second, these sets refer to mixes generating a combined fraction of individuals in the extreme and high risk state above 5%, between 5% and 3% and below 3%, respectively. We note that performance between these three sets differ significantly. *Policy Mix Set I* obtains an augmented average and final fraction of vulnerable individuals of 34% and 16.6% ($p < 0.01$), respectively. For policies in *Policy Mix Set II,* the fractions reduce to

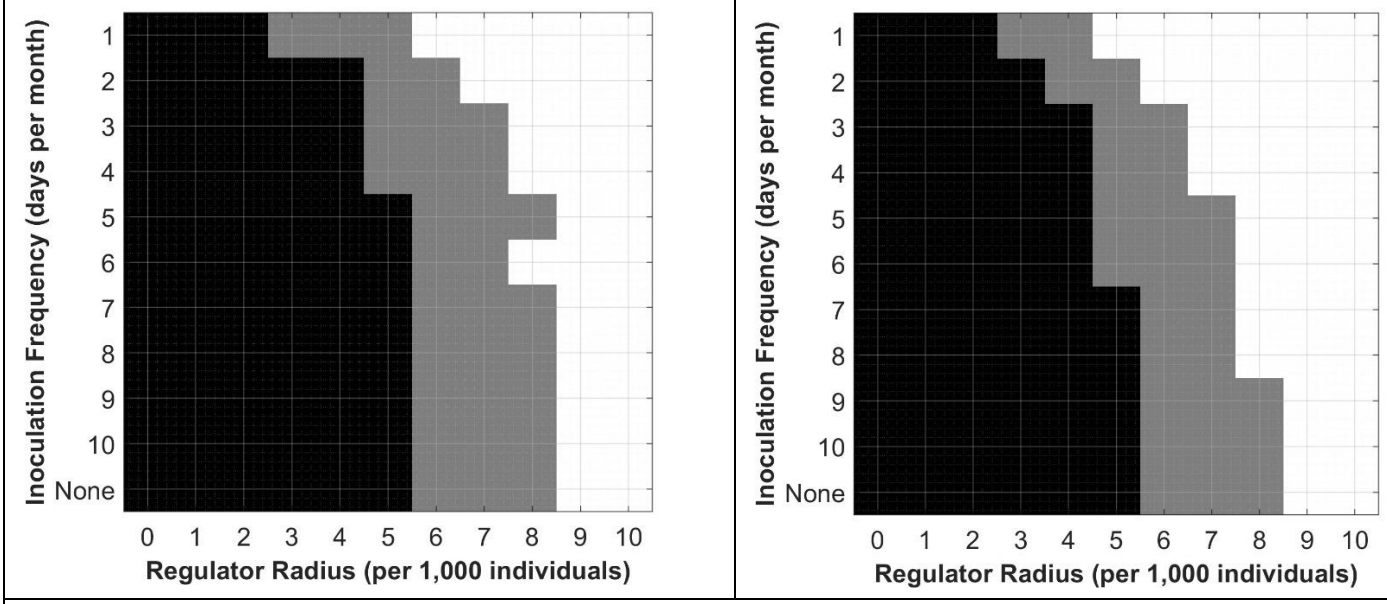

Figure 3. Possibility Frontier: the effect of various policy mixes on the combined fraction of vulnerable individuals (*Left Panel.*) Average mean fraction. (*Right Panel.*) Average end fraction.

3.9% and 2.7% ($p < 0.01$), respectively. Lastly, for policies in *Policy Mix Set III,* the amounts corresponds to 1.6% and 1.3% ($p < 0.01$), respectively.

Third, in Figure 3 we map the three policy sets in the 11×11 policy mix space (r^E, δ_I) against the augmented average performance measures, $\bar{n} + \bar{\sigma}$ and $\bar{n}(T) + \bar{\sigma}(T)$, respectively plotted in the left and right panel of the figure. For each pane, areas are coloured according to the policy set mix belongs to: black areas corresponds to mixes that fall within Policy Mix Set I, whereas grey and white areas corresponds to mixes in Policy Mix Sets II and III, respectively. From visual comparison 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

increases performance measures. Second, it appears that such substitution is possible only in the space defined by $r^R \geq 2$. Below such level of regulation intervention, *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*.

Next, in Table 4 we turn into exploring the mechanism of action of the two policies by measuring 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 simplicity, we focus on policies corresponding to the Policy Mix Set II. Critically, we remark that for comparable levels of average values, standard deviations are generally smaller[12] for policy mixes featuring a *higher level of inoculation* versus deplatforming.

To visualize the mechanism at the basis of the results of the variance analysis contained in Table 4, on the top panel of Figure 4 we plot the time series $n_m(t)$ for three configurations belonging to the policy mix set II, $(r^E, \delta^I) = (7, \text{none}), (6, 7), (4, 1)$. 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%. To emphasize the difference between the mechanism of action of the two policies, on the bottom panel we plot average mean fraction[13] of vulnerable users $\bar{n}\,(t)$ and the confidence interval computed on one standard deviation $\bar{\sigma}(t)$ computed at every time period $t = 1, . . ,365$ across the $M = 100$ simulations.

5.2 Results: Cost-Benefit Analysis

We use the estimates of the model in Section 5.1 to quantify the potential monetary saving stemming from the *joint* usage of deplatforming and inoculation. Using the productivity and cost structure of Section 4.1, we assess that the cost of a single Facebook contractor running an investigation 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.

In Figure 5 we plot the estimated fraction of vulnerable users as a function of the per-user cost of monitoring, corresponding to an investigation radius ranging from $r^R = 3$ to $r^R = 10$. In the picture, we consider four possible inoculation frequencies, spanning from no inoculation to $\delta^I = 8$ events per month. Through this methodology, it is possible to establish an cost-based equivalence between policies and assess the monetary value of inoculation. For example, from the figure we note that an yearly expenditure between $33.7 and $189.6, depending on the location of moderators, generates an expected

[12] In analysis available upon request we make this argument analytically precise by developing a large battery of Brown–Forsythe tests to compare the variance between policies achieving similar levels of effectiveness in means for the three sets of policy mixes.

[13] A similar analysis for the end period fraction of vulnerable users $\bar{n}(T)$ is available upon request.

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 is \$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 should be \$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[14], suggesting the idea that heterogeneous mixes of policies fare better that the use of one strategy alone.

[14] 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.

Vulnerable Users, Policy Mix Set II (Mean Across Period)

Inoculation Frequency	Mean Regulator 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 Regulator 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, Policy Mix Set II (Last Across Periods)

Inoculation Frequency	Mean Regulator 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 Regulator 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 4. Efficacy of policy mixes in Policy Mix Set II as defined in the main text.

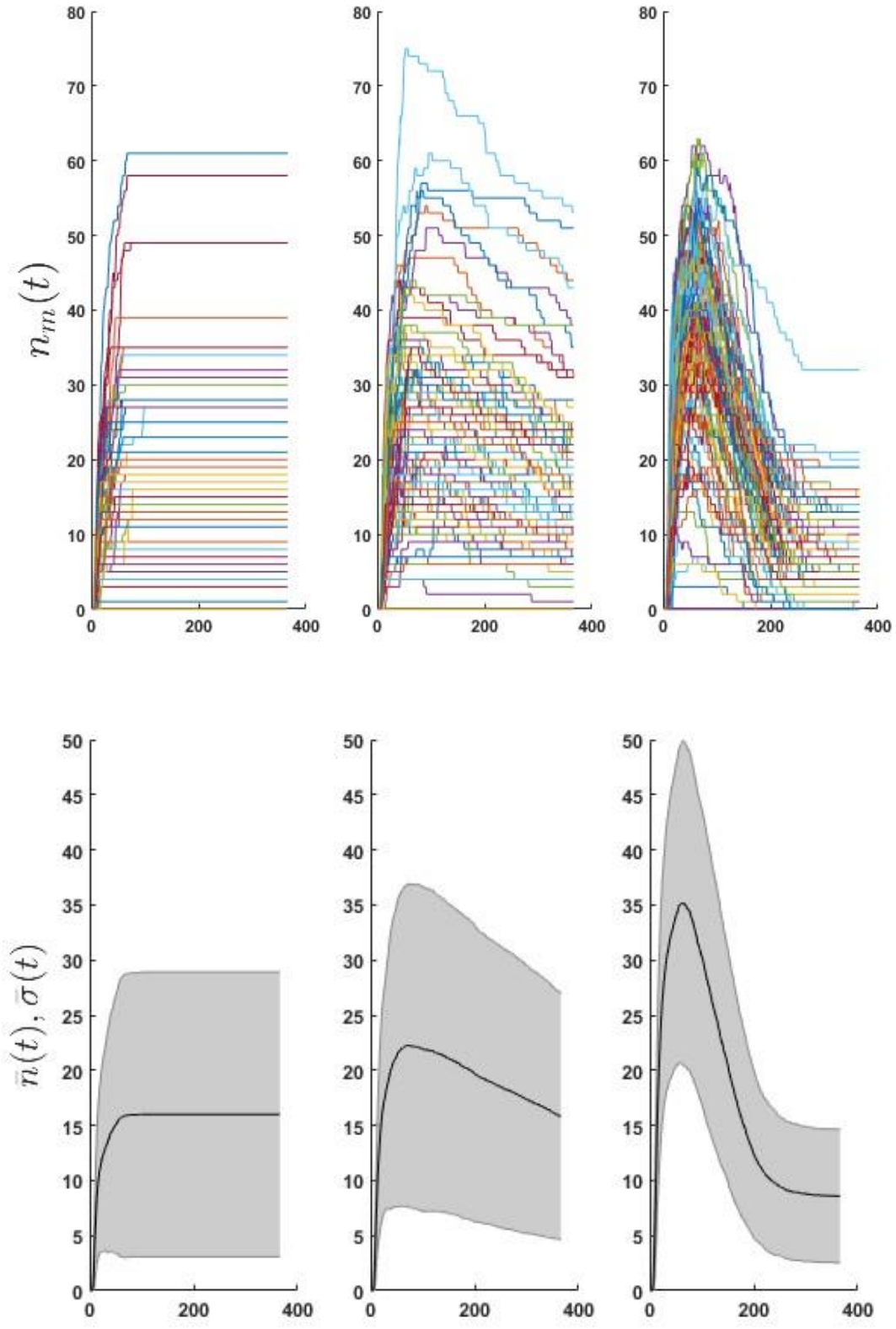

Figure 4. Time series for 3 equivalent policy mixes picked from the possibility frontier of average fraction of individuals in extreme and high risk state $(r^E, \delta^I) = (7, none), (6, 7), (4, 1)$.

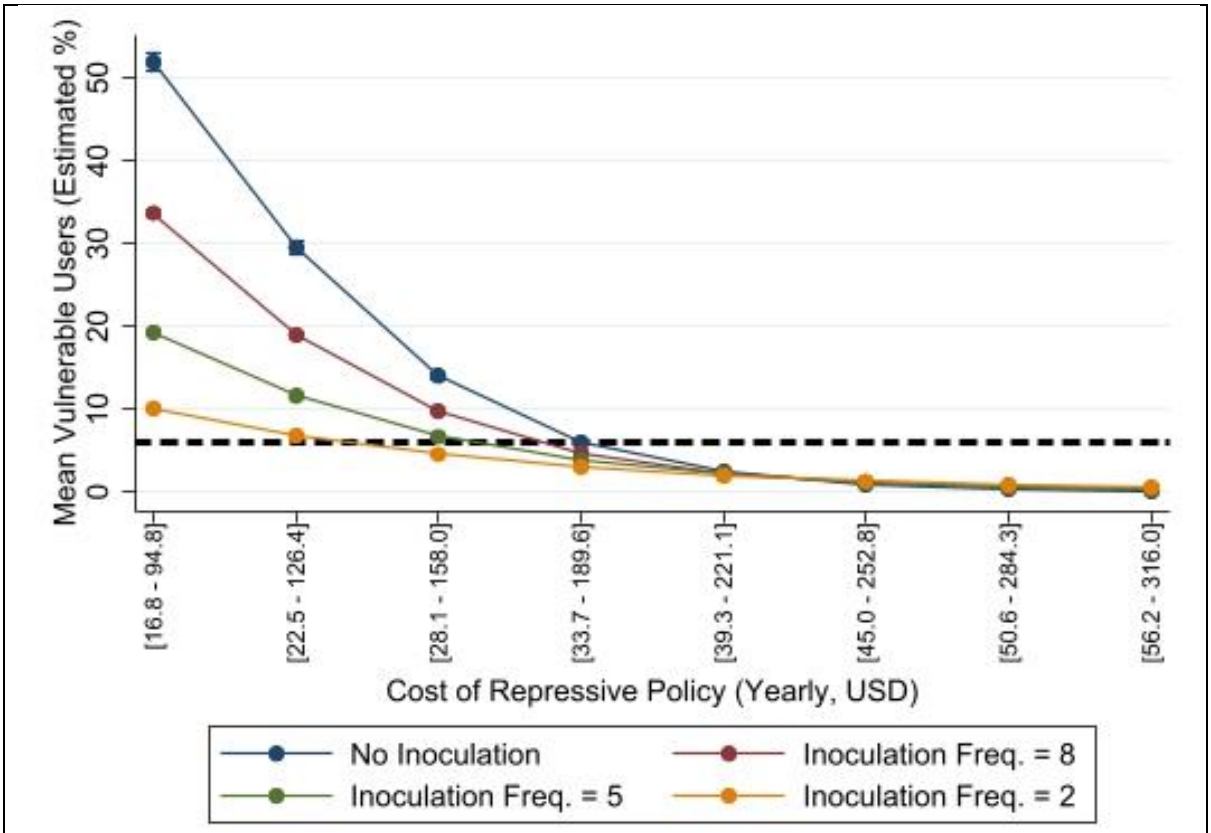

Figure 5. Cost-benefit Analysis.

6. Discussion (2300 words)

This study finds strong empirical support for the application of inoculation theory in the context of extremism, particularly in enhancing resistance to extremist narratives. According to Braddock (2019) and Lewandowsky and Yesilada (2021), inoculation can significantly reduce the persuasiveness and credibility of extremist content, which is corroborated by our findings that indicate reduced susceptibility under various

inoculation frequencies. Our study extends this by demonstrating economic and structural synergies when inoculation strategies are combined with monitoring policies. This synergistic effect suggests that not only do these strategies operate effectively in isolation but their integration can amplify overall policy efficacy. The economic viability of these combined strategies, evidenced through an analytical cost-benefit analysis, further underscores the practical applicability of inoculation theory in countering extremism within diverse socio-political contexts.

The application of inoculation strategies in our models highlights the predictability and resilience of societal responses to extremist ideologies. 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 predictable response is crucial in managing the dynamics of extremist ideologies effectively. The robustness of our models, validated through sensitivity testing and Monte Carlo simulations, ensures that the findings are credible and applicable. 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.

Placeholder In doing so, this research seeks to assess the efficiency of reallocating funds from repressive measures like police monitoring and online surveillance to preventative inoculation strategies. Repressive measures pose challenges to democratic societies, including potential backlash and perceived infringements on individual freedoms (Kattelman, 2019; Dugan & Fisher, 2023). In the long term, such strategies may inadvertently erode trust in democratic institutions (Spalek, 2010). Conversely, preventative approaches, often spearheaded by civil society organisations, align more closely with democratic principles by promoting engagement, empowerment, and resilience within communities (Barton et al., 2021). These strategies not only uphold individual liberties but also foster a participatory culture that strengthens societal resilience against extremism and yield wider positive effects on societies (Dalgaard-Nielsen & Schack, 2016). Hence, this study's significance lies in providing a data-driven foundation for policymakers to optimise resource allocation towards more democratic and sustainable counter-extremism strategies.

7. Conclusions (89 words)

References

ABS (Australian Bureau of Statistics) (2022) Education and Work, Australia, available at https://www.abs.gov.au/statistics/people/education/education-and-work-australia/may-2022 (retrieved 23 February 2024)

ACMA (Australian Communications and Media Authority), Telcos and law enforcement: Monitoring industry performance 2022–23

Adeyemi, Daniel. " Facebook content moderators in Kenya to receive 30-50% pay raise, following complaints." Techcabal, Originally published: March 4, 2022.

Alrababah, A., Marble, W., Mousa, S., & Siegel, A. A. (2021). Can exposure to celebrities reduce prejudice? The effect of Mohamed Salah on islamophobic behaviors and attitudes. American Political Science Review, 115(4), 1111-1128.

Baele, S. J., Brace, L., & Coan, T. G. (2023). Uncovering the far-right online ecosystem: An analytical framework and research agenda. Studies in Conflict & Terrorism, 46(9), 1599-1623.

Barton, G., Vergani, M., & Wahid, Y. (Eds.). (2021). Countering violent and hateful extremism in Indonesia: Islam, gender and civil society. Springer Nature.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. Salience. Annual Review of Economics 14 (2022): 521-544.

Calderoni, F., Campedelli, G. M., Szekely, A., Paolucci, M., & Andrighetto, G. (2021). Recruitment into organized crime: An agent-based approach testing the impact of different policies. Journal of Quantitative Criminology, 1-41.

Chalfin, A.J. (2010). From impact analysis to cost-benefit analysis: Methodological issues in the joint estimation of costs and benefits. Cost-benefit analysis and crime control, Roman, .K., Dunworth, T., & Marsh,. (eds), 167-181. The Urban Institute Press, Washington.

Cook, J., Maibach, E., van der Linden, S., & Lewandowsky, S. (2018). The consensus handbook. George Mason University. https://doi.org/10.13021/G8MM6P

Dalgaard-Nielsen, A., & Schack, P. (2016). Community resilience to militant Islamism: Who and what?: An explorative study of resilience in three Danish communities. Democracy and Security, 12(4), 309-327.

Drevon, J. (2016). Embracing Salafi jihadism in Egypt and mobilizing in the Syrian jihad. Middle East Critique, 25, 321-339. Doi:10.1080/19436149.2016.1206272

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.

Faucher, B., Assab, R., Roux, J., Levy-Bruhl, D., Tran Kiem, C., Cauchemez, S., ... & Poletto, C. (2022). Agent-based modelling of reactive vaccination of workplaces and schools against COVID-19. Nature communications, 13(1), 1414.

Hornick, J.P., Paetsch, J.J., & Bertrand, L.D. (2000). A manual on conducting economic analysis of crime prevention programs. Ottawa, ON: National Crime Prevention Centre.

Harris-Hogan, Shandon. "How to evaluate a program working with terrorists? Understanding Australia's countering violent extremism early intervention program." Journal of policing, intelligence and counter terrorism 15.2 (2020): 97-116.

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.

Hickey, Daniel, et al. "Auditing Elon Musk's impact on hate speech and bots." Proceedings of the international AAAI conference on web and social media. Vol. 17. 2023.

Kattelman, K. T. (2020). Assessing success of the global war on terror: Terrorist attack frequency and the backlash effect. Dynamics of Asymmetric Conflict, 13(1), 67-86.

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.

Marsh, K. (2010). Economic evaluation of criminal justice interventions: A methodological review of the recent literature. Cost-benefit analysis and crime control, Roman, .K., Dunworth, T., & Marsh, K. (eds), 1-28. The Urban Institute Press, Washington

Miller-Idriss, C. (2022). Hate in the homeland: The new global far right.

Neumann, P. R. (2013). The trouble with radicalisation. International affairs, 89(4), 873-893.

Peels, R. (2023). What Is It to Explain Extremism?. Terrorism and Political Violence, 1-18.

Pepys, R., Bowles, R., & Bouhana, N. (2020). A Simulation Model of the Radicalisation Process Based on the IVEE Theoretical Framework. Journal of Artificial Societies and Social Simulation, 23(3).

Ramshaw, A. (2024) Social Media Statistics for Australia (Updated January 2024), available at https://www.genroe.com/blog/social-media-statistics-australia/13492 (retrieved 23 February 2024)

Rottweiler, B., Gill, P., & Bouhana, N. (2022). Individual and environmental explanations for violent extremist intentions: a German nationally representative survey study. Justice quarterly, 39(4), 825-846.

Schneider, P. J., & Rizoiu, M. A. (2023). The effectiveness of moderating harmful online content. Proceedings of the National Academy of Sciences, 120(34), e2307360120.

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.

Spalek, B. (2010). Community policing, trust, and Muslim communities in relation to "New Terrorism". Politics & Policy, 38(4), 789-815.

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

Van der Linden, S., Maibach, E., Cook, J., Leiserowitz, A., & Lewandowsky, S. (2017). Inoculating against misinformation.. Science (New York, N.Y.), 358(6367), 1141–1142. https://doi.org/10.1126/science.aar4533

Vergani, M. (2018). How is terrorism changing us?. Basingstoke: Palgrave Macmillan.

Vergani, M., & Tacchi, E. M. (2016). When Catholics turn right: the effects of the Islamic terrorism threat on the fragmented Catholic Italian voters. Journal of Ethnic and Migration Studies, 42(11), 1885-1903.

Vergani, M., Iqbal, M., Ilbahar, E., & Barton, G. (2020). The three Ps of radicalisation: Push, pull and personal. A systematic scoping review of the scientific evidence about radicalisation into violent extremism. Studies in Conflict & Terrorism, 43(10), 854-854.

Ware, J. (2020). Fighting back: The Atomwaffen Division, countering violent extremism, and the evolving crackdown on far-right terrorism in America. Journal for Deradicalisation, (25), 74-116.

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.

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

Wooditch, A. (2021). The benefits of patrol officers using unallocated time for everyday crime prevention. Journal of quantitative criminology, 1-25.

Wooldridge, J. M. 2010. Econometric Analysis of Cross Section and Panel Data. 2nd ed. Cambridge, MA: MIT Press.

Morgan, T., Adams, O., & Hammond, D. Global Terrorism Index 2020: the shifting landscape of terrorism. Counterterrorism Yearbook 2021.

Perrigo, Billy. "Inside Facebook's African Sweatshop." TIME, Updated: February 17, 2022, 10:12 AM EST, Originally published: February 14, 2022, 7:30 AM EST.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, *393*(6684), 440-442.).

Thorley, T. G., & Saltman, E. (2023). GIFCT Tech Trials: Combining Behavioural Signals to Surface Terrorist and Violent Extremist Content Online. Studies in Conflict & Terrorism, 1-26.

Corbeil, A., & Rohozinski, R. (2021). Managing Risk: Terrorism, Violent Extremism, and Anti-Democratic Tendencies in the Digital Space. The Oxford Handbook of Cyber Security, 163.

Ross, C., Orr, E. S., Sisic, M., Arseneault, J. M., Simmering, M. G., & Orr, R. R. (2009). Personality and motivations associated with Facebook use. *Computers in human behavior*, *25*(2), 578-586.

APPENDIX 1. CALIBRATION

In this section we present the calibration strategy. A baseline setup is discussed and then further specialized for the two exercises performed in the main text.

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 $\boldsymbol{\phi}$. 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.

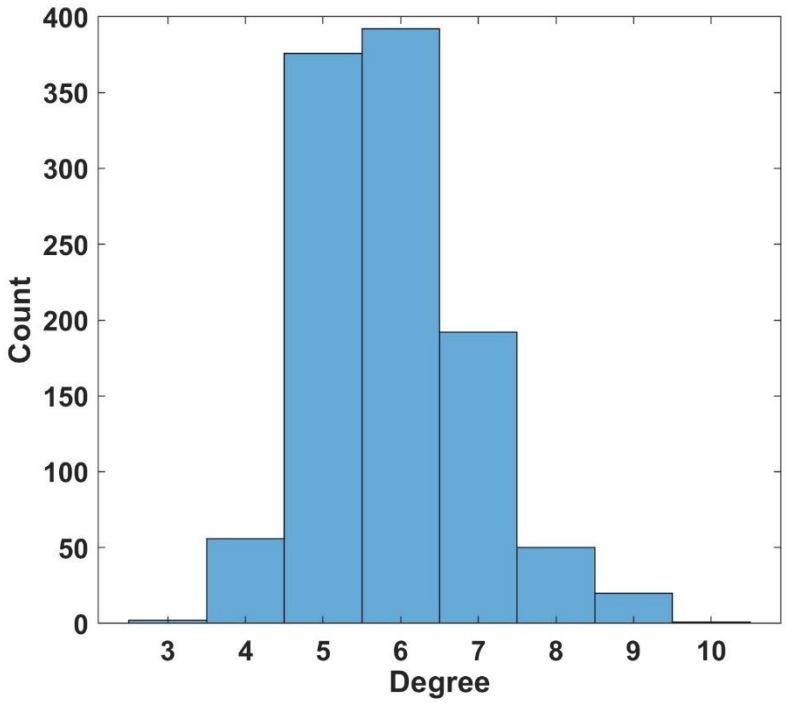

Figure A.1. Degree Distribution of Citizens emerging from the small-world network

Social Network $\boldsymbol{M}$. 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 distribution 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).

Regulators 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 (TIME, February 12th, 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., TIME, February 12th, 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 regulator to investigate 1% of the user population.

Regulators Radius r^R. Law 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 regulator 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 regulatory 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.

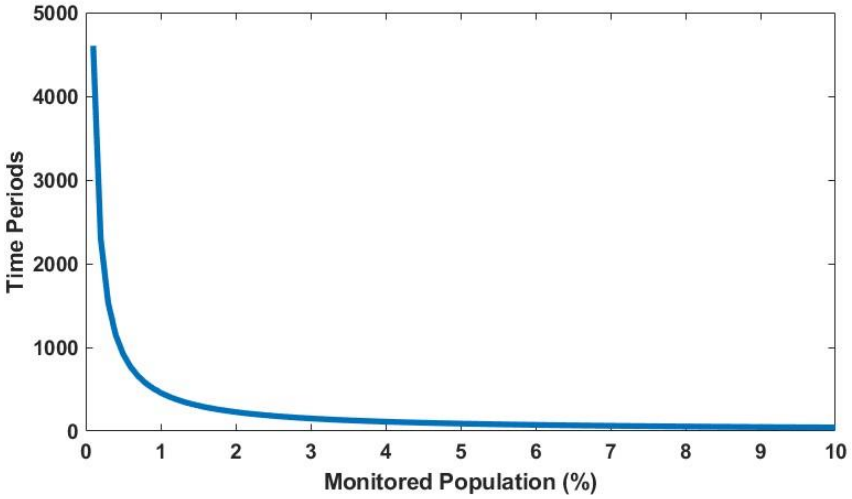

Figure A.2. The relationship between the regulator radius r^R and the time required to identify and deplatform one single extremist with probability 99% from an online group made of $N = 1000$ users under a random identification protocol.

APPENDIX 2. MATHEMATICAL RESULTS

The connection between Regulator 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 regulator and one extremist and assume that $\delta^R = 1$ to remove notational cluttering. Each period $t > 0$, the regulator 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 regulator 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.

OLD PART

Agent-Level characteristics of Model

Based on the theories and findings discussed in the Background section, we parameterised individual susceptibility or vulnerability to extremist content. 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 X). We assume a baseline probability of 0.5 for each user. Each risk factor increases the level of susceptibility, while each protective factor decreases it by a corresponding ratio.

Table X.

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

To anchor our case study to a realistic population dynamic on social media platforms, we calibrated our data 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.

Situational-Level characteristics

We employed NetLogo version 6.3.0 for modelling to conduct agent-based simulations that span 2 years (730 days), incorporating daily time-steps. Each parameter permutation was iterated 100 times to robustly explore the behaviour space of the simulation, focusing on the impact of extremist content over time within an online network. The susceptibility of each user to extremist content ranges from an immune state, where individuals outright reject extremist content, to an extreme state, where individuals are fully convinced by such extremist content. Each user receives a radicalisation score to describe these states, ranging from less than 0.05, signifying rejection of extremist views, to scores above 0.95, indicating extreme conviction in extremist views. The following table illustrates the possible states and the state transition of users in the network.

Table X.

Risk state	*Immune*	*Low-risk*	*Susceptible*	*High Risk*	*Extreme*
Radicalisation State	Rejects extremist views		Uncertain	Accepts extremist views	
Radicalisation Score	< 0.05	<= 0.33	0.33 < B < 0.66	>= 0.66	0.95

In our simulation, the initial setup designates most users as having an initial mean belief score of 0.5. This is motivated by the fact that they all belong to an anti-immigration Facebook group and are potentially susceptible to the messaging of far-right extremists.

Among 1000 users, only 1% are categorised as either immune or in an extreme state. The formation of initial network links incorporates several assumptions to mimic real-world social networks: users who are 'immune' or 'low risk' tend to follow other users that are 'immune' or 'low risk', adhering to a truncated normal distribution for the number of followings. These 'immune' or 'low risk' users also establish reciprocal connections, or friendships, with other 'immune' or 'low risk' users, again following a truncated normal distribution for the number of such relationships. We posited that some users would be 'influencers', that is, with a broader reach within the network, and accumulate followers based on a truncated normal distribution. Conversely, users with 'high risk' or 'extreme' views preferentially form friendships with other users characterised by 'high risk' or 'extreme' views according to a truncated normal distribution for the number of friends. Additionally, 'high risk' or 'extreme' views follow extremists, with the number of followings also determined by a truncated normal distribution. The network's structure is designed to remain constant throughout the duration of the simulation, providing a stable framework within which to observe the dynamics of information spread and belief formation. This approach allows for the controlled manipulation of specific parameters critical to our study.

The simulation begins with the extremists sharing extremist content. At each time step, if a user is exposed to extremist content, an impact score is calculated based on a range of parameters (i.e., the individual's susceptibility (protective and risk factors), information exposure methods, network's opinion, and whether they received inoculation treatment), followed by a radicalisation score update. Besides that, each user is assumed to perform any of the following social network activities: share inoculation training, share extremist content, and report extremists to the regulatory authority.

Operational rules within the simulation are defined to reflect various scenarios. If a user encounters inoculation treatment at any time-step (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 regulatory authorities.

After being exposed to either inoculation or extremist content, users assess the credibility of the source by gathering opinions from their connections within the network. Drawing on Schumann et al. (2024), we categorise exposure to extremist content into three types: active, accidental, and word-of-mouth (i.e., via a non-extremist user without direct extremist links). We propose that the impact of exposure to extremist content diminishes by half across these scenarios, with active exposure having a greater effect than accidental exposure, and accidental exposure having a greater effect than indirect exposure through another user. Similarly, for inoculation training, voluntary participation

yields a greater effect size compared to mandatory training, which in turn has a greater effect than training shared via word-of-mouth within the network. The influence of exposure on users with a high level of susceptibility is notably more significant for extremist content and notably less for inoculation training. If a user is not exposed to either extremist content or inoculation training at timestep t, the impact is calculated by recalling impact at timestep t-1 with discount by their respective decay rate. At timestep t = 0, the impact is 0.

Next, the model takes into consideration network opinions. The opinion is considered only for the sharing activities in the users network. We calculated a belief consensus score by the weighted average of current belief score in the users network (who they follow) based on the closeness of the relationship, whether they are influencers, and mind-likeliness (if the difference of their belief is within a tolerance level) for those who shared training or extremist content at timestep t. The difference between belief consensus score in the network and the current belief of the user itself determines the magnitude for change in belief. The output of the model, individual radicalisation score, is calculated and updated based on a Bayesian function. The posterior score is calculated by the prior score multiplied by a likelihood function of information assessment and consensus of network opinion.

$$B = \frac{P(B|E)}{P(\sim B|E)} = \frac{P(B)}{P(\sim B)} \times 2^{\beta}$$

β = perceived credibility of information

= network opinions + information assessment

= O + (PR x S) – (PI x (1 – S))

The following table provides an overview of the parameters manipulated for the purpose of our study.

Table X.

Variable	*Value/ Range*
Number of users	1000
Number of users with “immune” as initial status	1%
Number of users with “extreme” as initial status	1%
Number 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 number of followings for user	TruncNorm (5, 2.5, 0,1 0)
Distribution of number of followers for influencer user	TruncNorm (150, 75, 100, 200)
Distribution of number of followings to extremists	TruncNorm (X/2, X/4, 0, X)
Distribution of link weight	Norm (0.5, 0.5)
Probability of active participants	10%
Number of extremists (X)	[1, 2]
Exposure to extremist content effect size	[0.02, 0.04, 0.06]
Frequency of extremists sharing extremist content	1 per day
Decay rate to extremist content	0.9
Number of extremists	[0, 1, 3, 5, 7 ,10]
Inoculation effect size (I)	0.04
Frequency of inoculators provide training	1 per month (every 30 days)
Decay rate to inoculation training	0.9
Radius for inoculation policy area	4
Long term effect of training on individual susceptibility	Reduce S by I every 10 sessions
Number of enforcers	[0, 1]
Radius for enforces patrolling area	5
Frequency of enforcers patrolling	1 per week (every 7 days)
Tolerance for reporting before block a user	5

Model validation

In our study, we enhance the validity of the model through three complementary approaches. Firstly, the model is grounded in empirical data patterns, using Wolfowicz et

al. (2020)'s systematic review to identify key demographic factors that act as protective and vulnerability indicators in cognitive radicalisation. Additionally, we incorporate findings from Lewandowsky & Yesilada (2021) on the efficacy of inoculation against extremist content. Secondly, our model draws on psycho-social theories such as inoculation theory (Lewandowsky & Van Der Linden, 2021), which has been empirically established to offer resistance to persuasion across various domains, including health communication and countering online extremism (Banas and Rains 2010; Braddock, 2019; Lewandowsky & Yesilada, 2021), and social identity theory, which we use to model different scenarios underpinning the effects of exposure to extremist content (Vergani, 2018; Marble et al., 2021). Thirdly, we validate our model using "stylised facts" (Pepys et al. 2020), a method effective when quantitative validation data are scarce. This involves matching model outputs against established characteristics of the phenomenon under study. For instance, our model posits that agents, despite sharing a general anti-immigration stance, exhibit varied susceptibilities to far-right extremism, and it suggests a positively skewed distribution of extremist propensities, indicating a very small fraction of the population is significantly prone to extremist content. This design ensures the model reflects the dynamics of real-world extremism.