Extremism and Threats to Liberal Democracy

Policy Implications & Applications

A Research Agenda for
Liberal Democratic Resilience

ERA
Institute

02

Authors

Dr. Matthias J. Becker, NYU and University of Cambridge

Dr. Benjamin Folit-Weinberg, Ohio State University

Citation recommendation

Becker, M. J., & Folit-Weinberg, B. (2026). Extremism and Threats to Liberal Democracy: Policy Implications & Applications. Berlin: The Zera Institute.

Design credit

Design concept, layout, and typographic production by Simpelplus (Berlin)

Funding

This research was funded by the **Senatsverwaltung für Kultur und sozialen Zusammenhalt – Abteilung Kultur** (Senate Department for Culture and Social Cohesion – Department of Culture, Berlin).

Senate Department
for Culture
and Social Cohesion
BERLIN

Table of Contents

Part One

1. Purpose and Scope 5

2. Who This Report Is For 6

3. The Central Challenge 7

4. Policy Relevance Across Four Domains 8

5. Strategic Payoff 9

Part Two

6. Empirical Foundation: What We Observed 11

7. Four Hypotheses for Future Investigation 13

8. Why These Questions Matter for Policy 16

9. The Missing Research Infrastructure 18

10. Conclusion: From Observation to Liberal Democratic Resilience 22

Part One

From Diagnosis to Intervention: Why This Second Report

1. Purpose and Scope

The first report documented **how online radicalization unfolds within mainstream digital environments** through the **co-production of meaning between influencers, platforms, and audiences.** It demonstrated that illiberal and extremist discourse is not confined to fringe actors or isolated platforms but emerges through **participatory dynamics**—most notably within **comment sections**—where narratives are intensified, normalized, and rendered actionable. Antisemitism was identified as a **structural driver within these dynamics**, functioning as a connective discursive grammar linking conspiracy thinking, institutional delegitimization, and the normalization of violence across ideological contexts.

This second report builds directly on those empirical findings. Its purpose is to translate these findings into an **intervention-oriented research agenda**. Rather than treating extremism as a static outcome or a fixed ideological position, we follow the conclusion of the first report by conceptualizing radicalization as a **dynamic, interaction-driven process**—one that unfolds through **repeated exposure, linguistic habituation, moral framing, and escalation** within digital discourse ecosystems.

The central question guiding this report is therefore not only how online radicalization occurs, but **where and when it can be disrupted**. What points of leverage exist for policymakers, platform governance actors, educators, security practitioners, and developers of monitoring technologies? Which mechanisms appear particularly intervention-sensitive, and which require longer-term structural responses?

This report advances a set of empirically grounded hypotheses that identify such leverage points. These hypotheses are derived from the first report's systematic analysis of more than 10,000 user comments across platforms, events, and national contexts. The present study was not designed to test them, since doing so would require research infrastructure that does not currently exist. Making that gap visible—and arguing for the necessity of its closure—is a core objective of this report.

Before proceeding to the specific suggestions, it is worth reiterating several of the implications of the first report. Researchers and other actors concerned with radicalization should focus on **mainstream discourse**, not only on explicitly extremist spaces; on **posts and comment sections**, rather than isolated posts or individual actors; and on **illiberal discourse and processes of discursive closure**, rather than ideological orientation alone.

If, as the first report suggests, one need not go to extremist spaces to be radicalized, and if, as the first report suggests, commenters co-produce discourse with influencers, then **commenters themselves possess substantial capacity to shape discursive trajectories**—potentially amplifying, stabilizing, and normalizing illiberal frames.

A further implication, developed later in this report, is that because commenters actively co-produce online discourse, comment sections themselves constitute a critical leverage point for influence operations—including those conducted by foreign actors—with direct implications for democratic resilience and national security.

This insight brings into view four interrelated levels of future research and policy-relevant intervention:

- **Security/National Security & Risk Assessment:** Early-warning system indicators; improved detection of coordinated or foreign influence operations
- **Education & Prevention:** Empirically grounded insights into how radical frames normalize violence; near real-time foundations for counter-radicalization and resilience-oriented educational content
- **Social Media Policy & Regulation:** Evidence-based criteria for platform governance; identification of amplification dynamics warranting intervention
- **Applied AI Development:** Translation of qualitative discourse insights into context-aware monitoring and diagnostic systems, with an emphasis on prevention rather than purely reactive enforcement

In addition to these hypothesis-driven research priorities, the report outlines one deliberately limited exploratory study design that illustrates how selected hypotheses—particularly those concerning algorithmic amplification and moderation displacement—can be investigated using a constrained, proof-of-concept approach. This design is not intended to resolve the structural gaps identified in this report, but to function as a proof of concept demonstrating what becomes empirically visible once distribution-side dynamics are examined alongside discourse-level analysis.

2. Who This Report Is For

This report is written for decision-makers and practitioners concerned with democratic resilience in digital environments, including:

- **Policymakers and regulators**, particularly those involved in platform governance, transparency regimes, and duty-of-care standards
- **Security and risk assessment actors** responsible for identifying early warning signals of radicalization, coordinated influence operations, or escalation toward violence
- **Platform governance and trust-and-safety teams** seeking evidence-based criteria for prioritization and intervention
- **Educators and prevention practitioners** tasked with developing media literacy, pattern literacy, and counter-radicalization initiatives
- **Researchers and AI developers** working at the intersection of discourse analysis, monitoring systems, and ethical AI deployment

Rather than offering prescriptive solutions, the report provides actionable research priorities that can inform policy, regulation, education, security operations, and applied AI development.

3. The Central Challenge

The empirical findings of the first report revealed recurring patterns in online radicalization. However, systematically testing the hypotheses these patterns suggest requires research infrastructure that does not currently exist.

No single institution presently combines:

- sustained multi-year funding
- cross-platform data access
- privacy-protected longitudinal tracking
- integrated expertise in discourse analysis, data science, security studies, and platform research.

As a result, **critical policy-relevant questions remain unanswered**—not because they are unimportant, but because no ecosystem is currently capable of answering them at scale.

This report therefore pursues three objectives: to document empirically observed radicalization dynamics in a form usable for **policy reasoning**; to identify the key research questions that must be addressed in order to **move from diagnosis to prevention**; and to clarify why **existing research capacities are insufficient** to do so at scale. The exploratory algorithm-baseline study described later in the report should be read as a limited bridge design that renders certain distribution-side mechanisms partially observable, without substituting for the longitudinal, cross-platform infrastructure that effective prevention ultimately requires.

4. Policy Relevance Across Four Domains

Defining Early Warning in Digital Radicalization Research

In this report, early warning does not refer to the prediction of individual acts of violence, the automated identification of “extremist users,” or the attribution of intent. Rather, early warning is understood as the structured, ongoing observation of discursive dynamics within large-scale digital environments.

Early-warning signals emerge at the level of aggregate patterns, including the consolidation of dominant discourse regimes, shifts from plural or contested framing toward moral absolutism, and the increasing density and normalization of illiberal or anti-democratic repertoires following triggering events. These signals are probabilistic and contextual, designed to inform prioritization, human assessment, and preventive intervention rather than enforcement or attribution.

Within this framework, computational methods support early warning by structuring attention across large volumes of discourse and identifying moments of escalation or consolidation, while expert analysts provide contextual interpretation and judgement.

As noted above, the research agenda outlined in this report is directly relevant to four policy domains:

Security and Risk Assessment
Improving early-warning capabilities by identifying indicators of escalation, coordination, and discourse-driven violence risk—particularly in the immediate aftermath of triggering events. This includes influence campaigns, potentially by foreign actors.

Policy and Regulation
Providing empirical foundations for platform governance, including criteria for identifying amplification dynamics that warrant intervention and for assessing whether moderation reduces or merely displaces extremist discourse.

Education and Prevention
Informing evidence-based approaches to media literacy and prevention by clarifying how radical frames normalize, how discursive thresholds erode, and when intervention is most effective.

Applied AI Development
Translating qualitative insights into context-aware monitoring systems that prioritize prevention and human-in-the-loop decision-making over reactive enforcement.

5. Strategic Payoff

By linking empirical analysis to intervention-oriented hypotheses, this report produces **actionable knowledge** capable of informing strategies to **protect liberal democratic norms in digital environments**.

Not all hypotheses outlined here can or should be tested simultaneously. The report therefore presents a research agenda, whose components can be pursued independently or sequentially depending on mandate, capacity, and strategic priority.

Of particular importance is the hypothesis concerning early commenter influence, which suggests that targeted interventions during the earliest phase of a discourse event—particularly when early comments achieve high visibility—may prevent broader escalation at relatively low cost. This represents a potentially high-impact, low-resource entry point for policy and platform action.

Part Two

From Empirical Observation to Policy-Relevant Hypotheses

6. Empirical Foundation: What We Observed

Across all observed patterns, antisemitism functions as a cross-cutting escalation grammar, particularly in identity-salient events, amplifying conspiracy formation, moral inversion, and violence legitimation.

This report builds on the empirical findings of the first study, which analyzed over 10,000 user comments posted within approximately one week following three major violent or politically charged events in 2025 across YouTube, X (Twitter), TikTok, and Instagram. The aim of that analysis was not to predict violence or measure long-term behavioral change, but to identify discursive patterns that reliably emerge in the immediate aftermath of triggering events and that appear structurally relevant for radicalization processes.

Across platforms, cases, and national contexts, five empirical patterns were particularly salient. Together, they motivate the hypotheses formulated in Section 7.

6.1 Co-Production Dynamics in Comment Ecosystems

The analysis consistently showed that online radicalization is not produced by influencers alone, nor by isolated extremist users. Instead, it emerges through **co-production between influencer framing, platform affordances, and audience participation.**

Influencer posts establish an initial interpretive frame. Commenters then draw selectively from a shared illiberal toolkit—including enemy construction, conspiratorial scaffolding, moral absolutism, and the normalization or justification of violence. Through repetition, alignment, and contestation, comment sections frequently transform implicit or ambiguous cues into explicit ideological positions. In this process, meaning is not merely echoed but sharpened, stabilized, and normalized.

6.2 Platform-Specific Variation Without Clear Causal Attribution

Distinct patterns of escalation were observed across platforms. Long-form comment environments (e.g., YouTube) facilitated elaborated narrative consolidation; short-form and high-velocity platforms (e.g., X, TikTok) favored sloganization, moral signaling, and rapid polarization.

However, the present study cannot isolate whether these differences are driven primarily by:

- platform design
- audience self-selection
- moderation regimes, or
- algorithmic amplification dynamics.

The observed variation is therefore descriptive rather than causal, underscoring the need for systematic comparative testing.

Importantly, this **uncertainty itself has relevance to policy**: it underscores how limited both platform-internal and regulatory understanding remains regarding which specific design features, moderation practices, or audience dynamics drive radicalization outcomes.

6.3 Contextual Modulation: Germany and the United States

The same illiberal repertoires appeared across national contexts, but with different surface forms.

In Germany, legal constraints and historical memory incentivized coded language, moral inversion, and Holocaust-related reframing. In the United States, discourse was more frequently explicit, ideologically plural, and embedded in mainstream political culture under First Amendment protections.

Crucially, these differences did not reflect different underlying mechanisms. Rather, they demonstrated how a shared discursive toolkit adapts strategically to legal, cultural, and historical conditions.

6.4 Early Comment Dynamics and Discursive Direction

Across multiple cases, threads in which early comments attracted high engagement and deployed illiberal frames appeared more likely to develop sustained escalation within the observed one-week window.

While this pattern was not tested systematically, it suggests that temporal positioning within comment threads may play a disproportionate role in shaping downstream discourse norms—an observation with potentially high policy relevance.

6.5 Event Characteristics and Discourse Persistence

Events combining political violence with identity-salient targets—particularly those involving antisemitic dimensions—generated more persistent and internally coherent illiberal discourse than policy controversies or abstract debates.

Although the one-week observation window limits claims about longer-term persistence, the density and internal consistency of discourse during this period varied markedly by event type, indicating that not all “discourse shocks” are equal in radicalization potential.

Interim Conclusion

Taken together, these **observations suggest that online radicalization is interaction-driven rather than content-driven**, structurally shaped by **platforms and context**, and **temporally sensitive with identifiable escalation windows**. They also make clear that current research designs are insufficient to answer the most policy-relevant questions raised by these patterns. The following section, therefore, formulates a set of testable hypotheses derived directly from these empirical observations—each tied to concrete policy applications and research gaps.

7. Four Hypotheses for Future Investigation

The empirical patterns described above generate a set of **intervention-relevant hypotheses**. These hypotheses are grounded in observation but **were not tested** in the present study. Doing so would require longitudinal data, cross-platform tracking, or platform-internal information that is currently unavailable to independent researchers.

Nevertheless, articulating these hypotheses is essential for policy, security, education, and AI development, as they identify where targeted intervention could plausibly disrupt radicalization dynamics.

7.1 Linguistic Habituation and Progressive Adoption of Illiberal Repertoires

Observed pattern:
Within the one-week post-event observation window, discourse environments characterized by dense illiberal framing exhibited increasing repetition and stabilization of the same rhetorical patterns.

Hypothesis:
Repeated exposure to illiberal discourse increases the likelihood that users will adopt similar discursive repertoires over extended periods (12–18 months), even in the absence of explicit recruitment or directive messaging.

Testing requirements:
Longitudinal cross-platform tracking of individual user discourse evolution; systematic coding of discourse patterns over time; identification of exposure thresholds, time lags, and escalation patterns; privacy-protected research infrastructure capable of tracking users across platforms without compromising anonymity.

Why this matters:
If confirmed, this would support exposure-based risk models relevant to **early-warning systems, education and prevention**, and **AI-assisted differentiation** between organic radicalization and automated behavior.

7.2 Moderation Displacement Rather Than Reduction

Observed pattern:
Platforms with stricter moderation regimes showed fewer explicit extremist statements but more coded, ironic, or ambiguous forms of illiberal discourse.

Hypothesis:
Stronger moderation reduces explicit extremism but may increase implicit or coded forms, indicating displacement rather than elimination of radicalizing discourse.

Testing requirements:
Cross-platform comparison using moderation policy data and enforcement metrics; controlled baseline exposure studies examining the content distribution patterns encountered by new accounts under different engagement conditions; user migration tracking across platforms; controls for audience self-selection effects; standardized coding frameworks applicable across different linguistic forms (explicit, coded, ironic).

Why this matters:
Testing this hypothesis is central to evaluating whether current moderation strategies reduce harm or merely transform its linguistic surface—directly affecting regulatory design and AI detection priorities.

7.3 Event Type Predicts Duration and Intensity of Discursive Escalation

Observed pattern:
Within the one-week window, identity-salient violent events generated denser and more internally coherent illiberal discourse than policy controversies.

Hypothesis:
Events combining political violence with identity-based symbolism generate longer-lasting and more escalation-prone discourse cycles (extending beyond one week) than non-violent or abstract political events.

Testing requirements:
Extended monitoring windows (7–14+ days minimum); systematic comparison across diverse event types; measurement of sustained engagement, recruitment indicators, and escalation patterns; capacity for rapid research deployment following triggering events.

Why this matters:
If validated, this would inform event-based moderation protocols, security risk windows, and the timing of counter-messaging interventions.

7.4 Early Commenter Influence on Thread-Level Radicalization

Observed pattern:
Threads in which early, high-engagement comments deployed illiberal frames appeared more likely to develop escalatory discourse within the observed period.

Hypothesis:
Early high-engagement extremist comments significantly increase the likelihood of subsequent thread-level radicalization.

Testing requirements:
Temporal sequencing analysis across large samples (thousands of threads); focused analysis of the first 10–20 comments per thread; comparison by engagement level and user type; quasi-experimental controls for confounding variables. Unlike H1-H3, this hypothesis may be testable using existing comment data without requiring longitudinal or cross-platform infrastructure, though it still requires sufficient sample size, temporal analysis capacity, and ideally, platform partnership for validation.

Why this matters:
This hypothesis offers high policy return at low cost. Unlike H1–H3, it may be testable using existing comment data and could justify prioritizing early comment moderation, seeding detection, and bystander intervention strategies.

Taken together, these hypotheses do not point to single interventions or isolated tools. Rather, they describe an interaction-driven system in which radicalization unfolds through temporally sensitive, platform-mediated, and audience-amplified dynamics. Addressing such dynamics requires an analytical ecosystem rather than a singular solution—one that integrates empirical research, computational monitoring, expert interpretation, and policy application.

Within this framework, **early warning** refers not to prediction or enforcement, but to the capacity **to identify discursive escalation as it unfolds**, to distinguish **transient outrage from sustained normalization**, and to support **contextualized human assessment and preventive intervention**. The following section translates this ecosystem-oriented perspective into concrete policy implications across four domains.

8. Why These Questions Matter for Policy

The hypotheses outlined above are not abstract research concerns. Each corresponds to concrete policy dilemmas currently faced by governments, platforms, educators, and security practitioners. What is missing is not awareness of the problem, but empirically grounded guidance on where intervention is most effective.

This section translates the hypotheses into policy-relevant implications across four domains.

Each of the subsections below corresponds directly to one or more of the hypotheses outlined in Section 7, translating them into policy-relevant implications across distinct domains.

8.1 National Security and Risk Assessment

Current challenge:
Security agencies increasingly recognize that violent acts rarely emerge without prior online signaling. Yet they lack reliable methods to distinguish:

- organic radicalization from coordinated influence operations
- short-term outrage from durable ideological consolidation
- and high-risk discourse from background noise

These indicators are probabilistic rather than predictive and are not intended to support deterministic forecasting or pre-emptive enforcement, but to inform time-bound human assessment and prioritization.

What testing would enable:

- would clarify whether linguistic adoption patterns can serve as early indicators of individual radicalization trajectories, distinct from bot activity or one-off expressions.
- would identify which event types reliably produce extended escalation windows, enabling prioritization of monitoring resources.
- would allow detection of coordinated “seeding” strategies through early comment activity.

Policy payoff:
Rather than reacting after violence occurs, security actors could deploy probabilistic early-warning indicators within defined temporal windows (particularly during the first week following identity-salient events), supporting human decision-making for analytical prioritization rather than operational targeting, without automated enforcement.

One essential implication of the first report is that, because they co-produce online discourse, **commenters play a crucial role in radicalization**. This provides a major lever of power that can be abused through influence campaigns, including by foreign powers, which can actively erode trust in institutions, normalize violence, and actively degrade liberal democracy. These hypotheses thus have a potentially important national security implication as well.

8.2 Platform Regulation and Governance

Current challenge:
Regulatory debates—particularly in the EU and the United States—are largely evidence-poor. Platforms are asked to mitigate “systemic risk” without clear, testable criteria for:

- which features amplify radicalization
- when moderation reduces harm versus displaces it
- and where intervention produces the highest return

What testing would enable:

- would directly inform whether stricter moderation regimes reduce extremism or merely transform its linguistic form.
- would provide empirical justification for prioritizing **early comment moderation**, rather than relying solely on post-level takedowns.
- would support event-based escalation protocols rather than uniform enforcement.

Policy payoff:
These findings could operationalize provisions such as DSA Article 34 (systemic risk assessment) and Article 40 (researcher access), grounding regulation in observable mechanisms rather than platform self-reporting.

8.3 Education, Prevention, and Democratic Resilience

Current challenge:
Educational and prevention initiatives often target extremist ideology directly but lack insight into how normalization occurs before explicit extremism appears.

What testing would enable:

- would identify vulnerable stages in discursive adoption, allowing interventions before ideological hardening.
- would support cross-platform literacy, teaching recognition of coded and implicit forms rather than only explicit hate.
- would inform bystander intervention strategies focused on early norm-setting in comment spaces.

Policy payoff:
Education could shift from reactive content rebuttal toward **pattern literacy**—equipping users to recognize escalation dynamics, moral framing traps, and normalization processes across platforms.

Moreover, since the specific expressions of illiberalism, overt and coded, evolve far more rapidly than most educators can keep up with, active monitoring of these expressions would allow for educators to address them in educational settings as they arise on social media.

8.4 AI Development and Monitoring Technologies

Current challenge:
AI systems struggle with irony, context, and coded meaning. Most existing tools prioritize the detection of explicit content and operate downstream of escalation.

What testing would enable:

- would provide training data for distinguishing human adoption timelines from automated behavior.
- would improve the detection of moderation-evasion strategies.
- would enable **real-time thread-level risk scoring** during early discourse phases.

Policy payoff:
AI systems could shift from reactive enforcement to early-warning and prioritization tools, supporting human moderators rather than replacing them. Crucially, this approach avoids overclaiming automation and preserves proportionality and accountability.

From a policy perspective, the relevance of AI-assisted approaches lies not in standalone detection systems, but in their integration into modular analytical pipelines. Within such pipelines, supervised models trained on expert-annotated data provide relatively stable reference layers, while unsupervised and context-aware components support the identification of emerging narratives, shifts in rhetorical framing, and changes in discourse intensity. These computational layers do not replace expert judgement; they structure attention and make large-scale discursive dynamics observable, enabling timely human assessment and proportionate response.

9. The Missing Research Infrastructure

The preceding chapters demonstrate that online radicalization follows identifiable, recurrent patterns across platforms, events, and national contexts. They also show that these patterns generate empirically grounded hypotheses with direct relevance for platform governance, security assessment, education, and AI-assisted monitoring. However, the present study also reveals a critical structural problem: the key hypotheses identified in this report cannot currently be tested systematically.

This limitation does not reflect a lack of analytical clarity or empirical insight. Rather, it reflects the absence of a research infrastructure capable of studying radicalization as a longitudinal, cross-platform, interaction-driven process. Without such infrastructure, policy interventions remain reactive, moderation efforts remain symptom-focused, and preventive strategies lack empirical grounding.

The absence of a shared research infrastructure has direct implications for early-warning capacity. Without sustained, cross-platform, and longitudinal analytical systems, early-warning signals remain fragmented, episodic, and difficult to interpret. Effective early warning depends on the ability to observe discursive dynamics across time, platforms, and events, to distinguish short-term outrage from sustained escalation, and to contextualize computational signals through expert analysis. Infrastructure, in this sense, is not an auxiliary concern but a precondition for translating empirical observation into preventive policy action.

9.1 What Is Currently Missing

Several forms of research capacity are structurally absent or severely underdeveloped.

At present, policymakers lack systematic, empirically grounded intelligence about how harmful discourse emerges, mutates, and consolidates within mainstream digital environments over time (see also Becker, 2025).

Longitudinal tracking capacity.
Current studies—including the present one—are largely limited to short observation windows. There is no sustained infrastructure for tracking how individuals' discourse evolves over months or years, how exposure accumulates, or when escalation thresholds are crossed.

Cross-platform comparability.
Radicalization dynamics unfold across multiple platforms with distinct affordances, moderation regimes, and audience cultures. Yet most research remains platform-specific, preventing systematic comparison or detection of displacement effects. Exploratory baseline studies—such as controlled observations of content exposure for newly created platform accounts—can provide valuable snapshots of algorithmic distribution dynamics. However, by design, such studies cannot capture cumulative exposure effects, user adaptation, or longer-term radicalization trajectories. They therefore complement, but cannot substitute for, longitudinal and cross-platform research infrastructure.

Temporal sequencing analysis.
There is no standardized capacity to analyze how discourse develops within threads over time—particularly the role of early commenters, engagement ordering, and amplification dynamics during the first hours or days of a discourse event.

Integrated qualitative-computational frameworks.
While computational approaches excel at scale and speed, they remain weak at detecting implicit, coded, or context-dependent hate. Conversely, qualitative discourse analysis offers precision but lacks scalability. Infrastructure that integrates both remains exceptional rather than standard.

Independent validation and transparency.
Platform-internal research is not externally auditable, while independent researchers lack access to moderation data, recommendation logic, or engagement metrics necessary for robust causal analysis.

9.2 Why Existing Actors Cannot Fill These Gaps Alone

While Section 9.1 identified the research capacities that are currently missing, the problem is not merely technical. No single existing sector is structurally positioned to assemble these capacities independently.

Universities possess methodological expertise and normative independence, but typically lack sustained funding, platform access, and long-term computational infrastructure.

Platforms hold the relevant data and technical capacity, but face conflicts of interest, limited incentives for transparency, and weak mechanisms for external validation.

Security agencies have resources and operational urgency, but generally lack discourse-analytic expertise, academic accountability, and cross-national comparative frameworks.

Civil society organizations provide monitoring and rapid response, but lack the scale, stability, and analytical integration required for hypothesis testing and longitudinal research.

As a result, existing efforts remain fragmented. Each actor contributes partial insight, but none can independently generate the evidence base required for preventive, proportionate, and rights-respecting intervention.

9.3 Consequences of Inaction

The absence of an integrated research ecosystem has direct policy consequences. Recent platform-wide moderation changes following high-profile attacks—implemented rapidly but without empirical grounding—have at times displaced extremist discourse into more coded forms rather than reducing its prevalence, illustrating the costs of acting without validated evidence.

In short, liberal democratic resilience is undermined not by a lack of concern but by a lack of coordinated research capacity.

9.4 Implications for Policy

Addressing these gaps will require sustained, cross-sector collaboration, regulated research access, and multi-year investment. The present report does not prescribe a single institutional model or implementation pathway. Instead, it establishes the empirical necessity of building one.

While no single project can resolve the structural gaps identified above, targeted investment in hypothesis-driven research can materially advance understanding of where and when intervention is most effective, generating empirical findings that directly inform subsequent policy, platform governance, and security decision-making.

As a concrete and manageable first step, policymakers and funders could support the creation of a dedicated, interdisciplinary research team tasked with undertaking a multi-year program of work aligned with the research priorities outlined in this report. Such a team would bring together expertise in discourse analysis, data science, platform research, security studies, and applied AI, and would operate under clear ethical, legal, and transparency safeguards.

The mandate of this team would not be predictive enforcement or operational targeting, but the systematic study of discursive escalation dynamics across platforms and events, including longitudinal tracking, temporal sequencing analysis, and cross-platform comparison. Crucially, this would require regulated research access to platform data, coupled with independent oversight and mechanisms for external validation.

Establishing such a team would not, on its own, resolve the structural deficits identified in this report. However, it would represent a critical step toward building the analytical capacity required for evidence-based prevention, proportionate policy intervention, and credible early-warning frameworks grounded in empirical understanding rather than reactive response.

Without such initial investment, efforts to counter online radicalization will remain fragmented and episodic. With it, liberal democratic societies gain the opportunity to move from post hoc reaction toward anticipatory, accountable, and rights-respecting governance of digital discourse environments.

10. Conclusion: From Observation to Liberal Democratic Resilience

The analysis presented across both reports leads to a clear overarching conclusion: online radicalization is not random, episodic, or confined to extremist fringes. It follows recognizable patterns that unfold within mainstream digital environments through interaction, repetition, and normalization. These processes are shaped by platform architectures, influencer framing, and—crucially—audience participation.

The first report documented how illiberal and antisemitic discourse is co-produced in comment sections, how implicit meanings are stabilized through interaction, and how violence becomes morally defensible long before it is explicitly advocated. The second report has translated these findings into a forward-looking research agenda, identifying where intervention may be possible, what remains empirically unknown, and why existing institutional arrangements are insufficient to address these challenges.

Taken together, the findings suggest three core implications for liberal democratic resilience:

First, **liberal democratic erosion is a discursive process**. It does not require immediate institutional collapse or overt authoritarian takeover. It unfolds through gradual shifts in what can be said, implied, justified, and normalized within everyday political communication. Monitoring democratic health, therefore, requires attention not only to actions and outcomes, but to the communicative environments in which political meaning is produced.

Second, **effective prevention depends on understanding dynamics, not just content**. Isolated posts, keywords, or actors rarely explain escalation on their own. Radicalization emerges through temporal sequences, interactional reinforcement, and exposure effects that remain largely invisible to current moderation, regulatory, and security frameworks.

Early warning, as articulated across this report, is therefore best understood as a capacity rather than a tool: the capacity to **monitor discursive escalation as it unfolds**, to **interpret probabilistic signals** within their political and social context, and to **intervene** at moments when **normalization and consolidation are still reversible**. Computational methods enable scale and temporal sensitivity, but liberal democratic resilience ultimately depends on the institutions, expertise, and coordination required to act on these signals responsibly.

Third, **evidence-based intervention depends on infrastructure**. Exploratory designs can illuminate discrete mechanisms, but without sustained, cross-platform, and longitudinal research capacity, democratic societies remain structurally reactive—intervening only after escalation has already consolidated.

This report does not claim to offer final answers. Instead, it clarifies the questions that must now be addressed and the conditions under which they can be answered responsibly. Building the research ecosystem outlined here is not a technical luxury or an academic ambition. It is a prerequisite for developing proportionate, transparent, and effective responses to online radicalization that respect democratic values while addressing genuine risks.

Liberal democracies do not fail only through force. They weaken when exclusion becomes routine, when violence becomes morally legible and accepted, and when conspiracy replaces institutional trust. **These processes are visible, traceable, and—under the right conditions—can be interrupted**. Doing so will require not only resources, but coordinated mandates across regulators, research institutions, platforms, and security actors—none of whom can address these dynamics in isolation.

The patterns are known.
The hypotheses are clear.

What remains is the collective decision to build the capacity required to act on them.