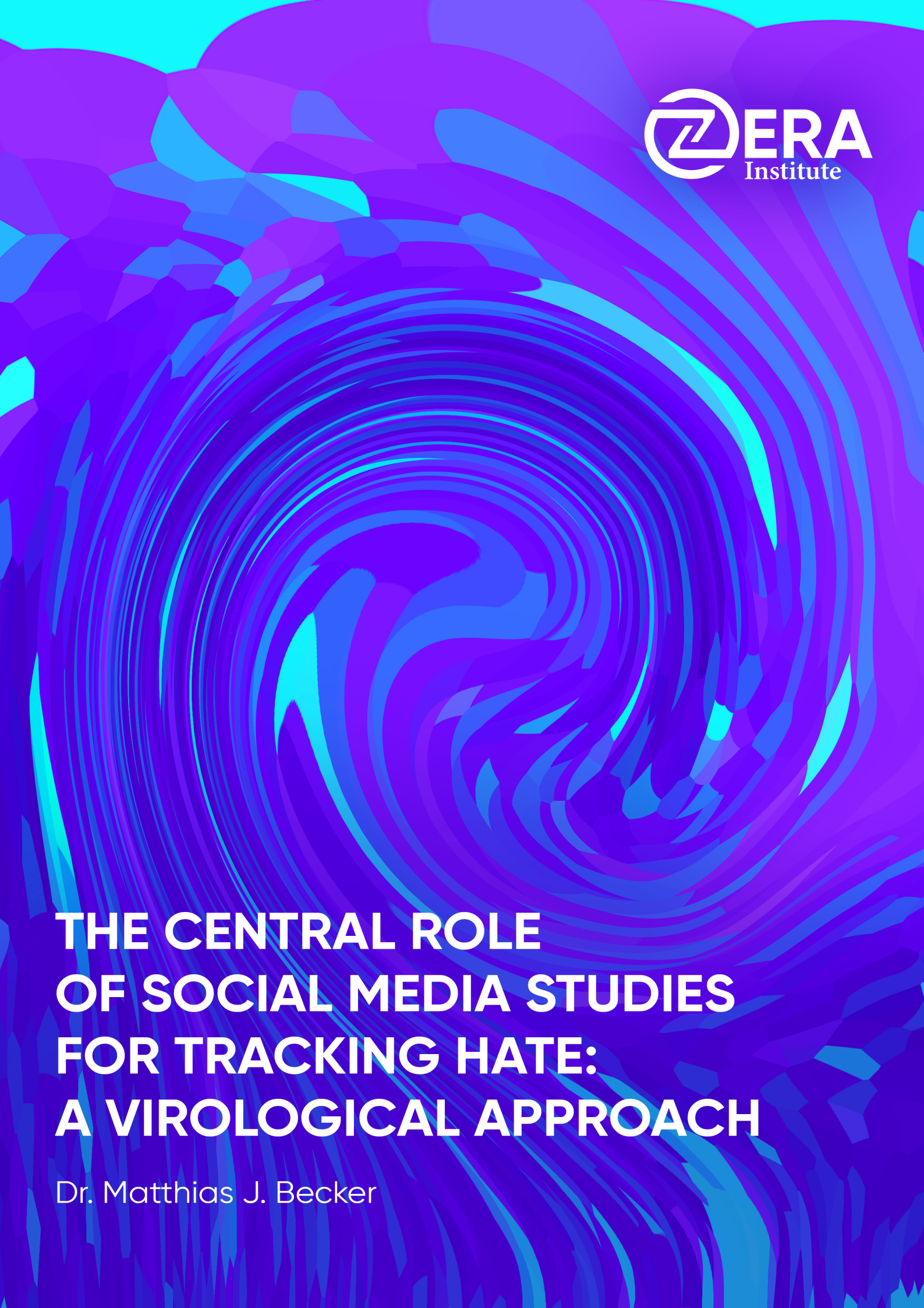

# THE CENTRAL ROLE OF SOCIAL MEDIA STUDIES FOR TRACKING HATE: A VIROLOGICAL APPROACH

Dr. Matthias J. Becker

Zera Institute e.V.
Molkenmarkt 2
10179 Berlin

info@zera-institute.com
www.zera-institute.com

# The Central Role of Social Media Studies for Tackling Hate: A Virological Approach

*Dr. Matthias J. Becker*
*Lead, Decoding Antisemitism*

## Abstract

Social media platforms have become the central marketplace of ideas in contemporary democracies—yet they remain a black box. Without systematic tools to deconstruct the dynamics unfolding within these digital spaces or track their evolution in real-time, we lack the fundamental intelligence necessary to understand the societal and political trends shaping our societies. This article argues for the urgent integration of social media studies—combining discourse analysis with actor-focused research—into every attempt to comprehend polarization, radicalization, and the erosion of democratic norms.

Adopting a **virological approach**, this article treats hate ideologies—antisemitism, racism, misogyny, conspiracy theories, and anti-democratic resentment—as adaptive pathogens that mutate, spread, and exploit vulnerabilities within the digitally restructured public sphere. Just as virologists must decode viral mechanisms at the molecular level before developing vaccines, researchers must systematically map how hate discourse manifests, transforms, and circulates across platforms through strategic ambiguity, algorithmic amplification, and participatory radicalization. This requires a transdisciplinary synthesis: computational social science to detect patterns at scale, qualitative discourse analysis to interpret implicit meanings and cultural codes, and AI-enhanced methodologies to track real-time evolution.

The stakes are existential for democratic societies. Social media platforms have fundamentally restructured political communication, creating favorable "political-cultural opportunity structures" for extremist narratives while simultaneously dismantling traditional gatekeepers and truth-arbiters. New intermediaries—influencers, algorithms, billionaire platform owners—now shape public discourse without institutional accountability, amplifying transgressive content that generates engagement regardless of social harm. The result is accelerating polarization, normalization of previously marginal ideologies, and erosion of the shared epistemic foundations necessary for democratic deliberation.

Yet policymakers have largely neglected this quasi-unregulated "town square," allowing it to influence societal consciousness in uncontrolled and increasingly dangerous ways.

Zera Institute e.V.
Molkenmarkt 2
10179 Berlin

info@zera-institute.com
www.zera-institute.com

Without rigorous social media studies that can penetrate the black box of digital discourse—revealing who says what to whom, how messages transform as they circulate, which actors and algorithms amplify which narratives, and what offline consequences follow—we

operate blind. This article presents a roadmap from research to action: building advanced mixed-method infrastructures, understanding phenomena and trends through systematic monitoring, translating knowledge into policy and education, and developing context-sensitive interventions. Only through sustained investment in social media studies as a central pillar of democratic governance can we hope to counter the viral spread of hatred and preserve the integrity of the public sphere.

# Introduction

In recent years, we have witnessed an alarming resurgence of hate-based ideologies across Western democracies—antisemitism, racism, misogyny, conspiracy thinking, and anti-democratic resentment—accompanied by a broader revival of authoritarian movements and systematic disinformation campaigns. These phenomena represent not merely temporary spikes in prejudice but rather the reemergence and transformation of age-old hatreds, now amplified by digital communication technologies. What makes this particularly concerning is the accelerating societal polarization and radicalization these dynamics produce, fundamentally threatening the cohesion of democratic societies.

This transformation reflects what scholars have termed a fundamental shift in "political-cultural opportunity structures"—the conditions under which hate speech, conspiracy theories, and anti-democratic ideologies can spread and gain legitimacy.1 The digitally restructured public sphere, coupled with what has been described as an "epistemic crisis" stemming from eroded trust in traditional gatekeepers and truth-arbiters such as mainstream media and democratic institutions, has created highly favorable conditions for the dissemination of extremist narratives across the ideological spectrum.2 These dynamics are further exacerbated by what some characterize as a "global wave of autocratization," wherein democratic norms and institutions face systematic erosion worldwide.3 What makes this revival particularly concerning is the widespread moral disorientation it reveals: segments of civil society that traditionally champion progressive values have, often unwittingly, aligned themselves with movements that embody the very authoritarianism, sexism, racism, and homophobia they claim to oppose.

Until recently, many believed that higher education and critical thinking would serve as bulwarks against extremist narratives. Yet the activities observed at universities across North America and Europe challenge this assumption. These demonstrations, praised by authoritarian regimes hostile to liberal democracy, raise fundamental questions about how educated individuals come to embrace ideologies fundamentally at odds with democratic principles.

Zera Institute e.V.
Molkenmarkt 2
10179 Berlin

info@zera-institute.com
www.zera-institute.com

Several explanations warrant consideration. These individuals may be genuinely uninformed, having never seriously examined the belief systems they appear to support. Alternatively, the prevailing disorientation and uncertainty across the West may be at play. Throughout history, such collective anxiety has frequently prompted people to embrace, despite their better judgment, positions that offer seemingly comprehensive solutions to complex problems. The relativist thinking that dominates contemporary discourse further compounds this tendency, creating confusion about fundamental values and pushing individuals toward ideological extremes.

However, even these factors cannot fully account for the current erosion of moral and intellectual clarity. In reality, the phenomenon emerges from an amalgamation of these elements, powerfully reinforced by social media platforms that have fostered an environment defined by materialism, hyper-individualism, and the relentless pursuit of self-validation. Social media platforms bring together all the essential elements for a perfect storm characterized by extreme feelings of aggression, fear, and isolation.

The recent surge in antisemitism, racism, misogyny, and conspiracy-driven disinformation, coupled with widespread rejection of democratic values, stems from multiple, interconnected factors. These phenomena can be partially attributed to malicious actors who deliberately manipulate public discourse to serve their own agendas. Numerous reports have substantiated this phenomenon, documenting the orchestration of public opinion by state and non-state actors seeking to destabilize democratic societies.4

Furthermore, the nature of digital communication involves both bottom-up and top-down dynamics that amplify hate and disinformation. While social media enables horizontal, bottom-up flows of information that bypass traditional gatekeepers, these dynamics intersect with new vertical, top-down mechanisms structured by opaque algorithmic curation. New gatekeepers—influencers, billionaire media moguls like Elon Musk, and platform algorithms—operate without the legal and institutional oversight standards that circumscribe established media. Algorithms privilege attention-generating, transgressive, polarizing, and emotional content, favoring commercial traffic while amplifying voices that promote disinformation and hatred in what has been termed a "post-factual age."5 Social media platforms' exploitation of these algorithmic systems, ostensibly for commercial gain, nevertheless exacerbates trends detrimental to fostering constructive, empathetic dialogue among online communities.6

These issues notwithstanding, the most significant factor remains the inherent nature of online communication itself—characterized by anonymity, mutual reinforcement among users, and the rapid dissemination of extremist ideologies. What has emerged is a complex and far-reaching toxicity unparalleled in human history. Operating under its own set of rules, the interactive web is shaping the consciousness of society at large.

Both bottom-up and top-down dynamics work together to expand and legitimize digital hate, fostering what has been characterized as a communicative authoritarian "politics of

transgression" online and in society at large.7 This politics of transgression operates across all forms of hatred and anti-democratic discourse—from antisemitism and racism to misogyny and conspiracy theories—gradually shifting the boundaries of socially acceptable speech. Through this process, extremist framings become normalized as legitimate opinion, and hateful ideologies migrate from fringe subcultures into mainstream publics.

Largely neglected by policymakers, this quasi-unregulated "town square," as Cal Newport aptly describes it, is influencing societal thought in an uncontrolled manner.8 Anything short of a serious assessment of this concatenation of events virtually guarantees harmful consequences for democratic stability.

Against this backdrop, linguistic and computational analyses become indispensable tools. As with any social contagion, combating hate-based ideologies—whether antisemitism, racism, misogyny, or conspiracy thinking—requires knowledge about their contemporary expressions and transmission mechanisms. While the dynamics described in this article apply across multiple forms of hatred and disinformation, we focus primarily on antisemitism as a paradigmatic case for understanding broader patterns of digital hate and radicalization.

Antisemitism provides a particularly instructive lens due to its discursive versatility and its function as a connective ideology that links otherwise distinct forms of hatred and illiberal resentment. Moreover, the current resurgence of antisemitism has been characterized as part of a "global antisemitic wave" unfolding both online and offline, connected to broader crises in democratic governance and the rise of authoritarian movements worldwide.9 This wave exemplifies how hate ideologies operate within favorable political-cultural opportunity structures created by the digitally restructured public sphere.

Research on antisemitism has identified two partly competing and partly complementary dynamics through which hate speech expands in digital environments. First, social media creates favorable conditions for transgressive dynamics, gradually shifting the boundaries of socially acceptable speech—allowing unfiltered hate, conspiracy myths, and extremist narratives to gain space and legitimacy. Second, hate ideologies broaden their reach by transforming over time and adapting to social codes of acceptability, circulating in camouflaged forms that maintain plausible deniability.10 Both dynamics—transgression and transformation—operate simultaneously across different forms of hate speech, from explicit racism to coded antisemitism to conspiracy-driven anti-democratic rhetoric.

As Rob Williams from the University of Southern California (USC) Shoah Foundation emphasizes regarding antisemitism, "we have to name it." Collecting knowledge about its various forms constitutes "the first steps we must take in the fight against antisemitism."11 Yet the methodological and analytical approaches developed for understanding antisemitism offer broader applicability: the same techniques for decoding implicit bias, tracking algorithmic amplification, and identifying radicalization pathways prove essential

Zera Institute e.V.
Molkenmarkt 2
10179 Berlin

info@zera-institute.com
www.zera-institute.com

for confronting racism, misogyny, conspiracy theories, and anti-democratic disinformation campaigns. If we seek to understand and combat the most prevalent manifestations of Jew-hatred, we must focus on what has become by far the most important venue for political opinion formation and debate culture: social media.

We still know far too little about the current state of antisemitic discourse in the online world. This applies to more conventional platforms like YouTube and Facebook as well as to Instagram, TikTok, Reddit, and various fringe platforms like 4chan. These digital spaces

play a pivotal role in shaping the political perspectives of young individuals, yet they remain largely unexplored, functioning essentially as a black box.

Research and civil society initiatives attempting to address this problem deserve recognition. The Decoding Antisemitism project12 is one such effort. While this project has attempted to excavate the myriad forms of open and coded antisemitism on the web,13 it has even been constrained by technological, administrative, and financial barriers. Consequently, most initiatives focus on specific platforms during particular measurement periods and employ specific research designs. They remain far from achieving a comprehensive understanding of the online ecosystem comparable to how social scientists describe phenomena in the offline world.

Recent empirical work has begun to document the scale and complexity of this challenge. Analysis of YouTube comment sections following the October 7, 2023 attacks revealed that antisemitic discourse surged to 36-38% of comments on major UK news outlets—nearly double the pre-crisis baseline of 15-25%.14 Similarly, reactions to the Washington, D.C. museum shooting in May 2025 showed an average of 43% antisemitic content across eight major English-language news channels, with some outlets reaching 66%.15 These figures illustrate not merely isolated incidents but a systematic normalization of antisemitic expression in mainstream digital spaces, contributing to broader patterns of societal polarization and radicalization.

The multitude of expressions—denial, trivialization, justification, or celebration of antisemitic violence, as well as stereotypes and conspiracy narratives across diverse platforms and online environments—are not adequately understood. Such understanding constitutes a fundamental requirement for any effective counterstrategy, whether initiated by educational or political actors.

## The Spectrum of Antisemitic Expression

Understanding antisemitism in digital spaces requires recognizing its discursive versatility. Antisemitic communication operates along a continuum from explicit incitement to implicit and coded expressions whose interpretation depends on contextual and cultural knowledge. This spectrum exists because antisemitism thrives on what scholars have termed "communication latency"—the persistence of antisemitic meaning in coded or camouflaged forms that remain socially intelligible while retaining deniability.16 Such latent

Zera Institute e.V.
Molkenmarkt 2
10179 Berlin

info@zera-institute.com
www.zera-institute.com

forms rely on irony, ellipsis, and intertextual cues rather than openly hostile statements, making antisemitic meaning context-dependent and discursively mediated. Digital platforms magnify this dynamic through what can be described as a "politics of transgression": implicit expressions test and expand the boundaries of acceptable speech through repetition and algorithmic reinforcement while maintaining plausible deniability.17

These questions become even more urgent when considering cases of physical violence, including lethal attacks. Historically, incidents such as those in Pittsburgh, Christchurch, and Halle have been linked to the online environments of perpetrators, where hate speech, conspiracy theories, and calls to violence were continuously propagated. More recently, the Washington, D.C. shooting of May 2025—in which 30-year-old Elias Rodriguez

murdered two Jewish embassy staff members, Yaron Lischinsky and Sarah Milgrim, at the Capital Jewish Museum while shouting "free, free Palestine"—further demonstrated how online radicalization pathways precede offline violence. Rodriguez, who reportedly claimed he "did it for Gaza," had a background in radical anti-Israel activism. When individuals are consistently exposed to reality-distorting discourse, the likelihood of their beliefs being radicalized increases substantially. This underscores how language and discourse play pivotal roles in shaping attitudes, which ultimately can manifest in violent actions.

## The Multimodal Challenge

Contemporary antisemitism does not confine itself to text alone. It increasingly manifests through multimodal assemblages where images, memes, emojis, and videos carry as much semantic weight as words. Following the October 7 attacks, emojis such as paragliders (🪂), Palestinian flags (🇵🇸), and watermelons (🍉) functioned as proxies for solidarity with Hamas or approval of violence. The watermelon emoji performs a particularly complex semiotic function: it recontextualizes the colors of the Palestinian flag while simultaneously implying that expressions of solidarity require coded forms due to supposed censorship—an insinuation that aligns with conspiratorial notions of Jewish control over public discourse.18

These multimodal signals evade keyword detection systems while remaining readily legible to in-group audiences. Memes repurpose popular formats to insert antisemitic analogies, portraying Jews or Israel as deceptive, greedy, or manipulative. Profile images featuring swastikas or Hitler portraits signal antisemitic alignment without requiring textual content. This complexity introduces higher-order challenges for computational systems: models must align visual, textual, and symbolic cues while integrating cultural literacy to interpret meaning accurately.

Comprehending these manifestations and the interconnections between stimuli, discourse events (both offline incidents and online occurrences with the potential to inflame antisemitism), and the subsequent waves of distortion, conspiracy narratives, and

Zera Institute e.V.
Molkenmarkt 2
10179 Berlin

info@zera-institute.com
www.zera-institute.com

disinformation can enable us to reconstruct user motivations and, by extension, understand offline behaviors. This illuminates what appeals to particular audiences and what their underlying needs and concerns entail. This kind of social media research represents the most effective—and likely the only—means of generating actionable insights into societal sentiments.

## A Virological Approach: Roadmap from Research to Action

Understanding hate ideologies in the digital age requires us to adopt a fundamentally new perspective. Like virologists studying pathogens, social media and hate researchers must examine how antisemitism, racism, misogyny, conspiracy thinking, and anti-democratic resentment exist, morph, and spread through contemporary communication networks. Just as virologists develop vaccines by understanding viral mechanisms at the molecular level, we must decode the structure, transmission patterns, and mutation dynamics of hate discourse to create effective remedies.

Recent research has revealed a critical transmission mechanism: strategic ambiguity functions as a catalyst for what can be termed "cascading radicalization" across digital platforms.19 This process follows a three-phase model whereby ambiguous elite discourse is reframed by digital intermediaries and then collapses into explicit hate speech in audience participation—a sequence summarized as "ambiguity → reframing → collapse."20 Understanding these transmission dynamics is essential for combating all forms of hate speech, from antisemitism and racism to misogyny and conspiracy-driven disinformation.

This virological approach demands sophisticated analytical tools that can only emerge from the systematic integration of humanities, social sciences, and data science. The challenge before us is not merely interdisciplinary—it requires a transdisciplinary synthesis where computer science, quantitative social research, and qualitative interpretive methods work in concert throughout every phase of investigation. While this article draws primarily on antisemitism research to illustrate methodological principles, the framework applies equally to understanding and combating other forms of hate speech and disinformation.

### Step 1: Advanced Research Infrastructure—Mixed Methods and LLM-Enhanced Methodologies

The foundation of effective intervention begins with robust research capabilities that transcend traditional disciplinary boundaries. Contemporary computational social science has typically combined data science with quantitative methods, but this framework overlooks the indispensable role that qualitative, interpretive research plays throughout the research process—an oversight that produces significant blind spots and biases.21

Zera Institute e.V.
Molkenmarkt 2
10179 Berlin

info@zera-institute.com
www.zera-institute.com

A comprehensive approach integrates three complementary domains: **computer science** (technical infrastructure for large-scale data collection, algorithmic processing, and machine learning), **quantitative social science** (theory-driven frameworks, statistical modeling, and rigorous hypothesis testing), and **qualitative social science** (exploratory approaches for understanding emerging phenomena, developing nuanced theoretical frameworks, and providing contextual understanding necessary for meaningful interpretation). This integration produces a computational social science capable of both statistical analysis and interpretive depth—one that can explore new phenomena while testing hypotheses at scale, interpret complex cultural meanings while processing millions of data points, and establish evaluation criteria that reflect genuine understanding rather than mere pattern matching.

Recent developments demonstrate the necessity of systematic integration across five research stages:21 **(1) Concept definition and operationalization** translates theoretical constructs like the IHRA working definition into linguistically testable indicators that account for co-text, context, and multimodal expression.22 **(2) Data collection and sampling** balances breadth of coverage with depth of contextual understanding—the Decoding Antisemitism project has annotated over 300,000 comments, yet this remains modest compared to billions circulating online.23 **(3) Annotation and model training** represents the critical juncture where human interpretation guides machine learning. Expert annotation integrates three knowledge domains: language knowledge (linguistic structures, pragmatic meanings), co-text knowledge (immediate discourse context), and world knowledge (historical understanding of hate tropes and cultural codes).24 Critically, annotation applies a **conservative attribution principle**: when statements permit multiple interpretations—one prejudiced and one non-prejudiced—the statement is coded as non-prejudiced unless substantial evidence establishes the prejudiced interpretation as more probable, protecting against over-interpretation while acknowledging potential underestimation of actual hate prevalence.25

**(4) Computational analysis** leverages recent advances in large language models through prompt engineering and chain-of-thought reasoning. Emerging retrieval-augmented generation (RAG) architectures couple LLMs with external knowledge bases—annotated corpora, lexicons, stereotype taxonomies—reducing hallucination by integrating relevant evidence. Looking ahead, context-engineered architectures could provide models with conversational co-text, historical event context, named-entity resolution, and situational metadata, approximating human expert reasoning.26 **(5) Model evaluation and validation** ensures automated systems produce meaningful insights rather than reproducing biases. Preliminary experiments with recent LLMs (Gemini-2.5-Flash, Llama-3.3-70B, MoonshotAI Kimi-K2) achieved F1-scores above 0.88 on antisemitism detection tasks,27 yet significant limitations remain: fine-tuned BERT models miss approximately one-third of cases,28 antisemitism remains statistically rare (1-2% of datasets), and data scarcity is exacerbated by platform restrictions and resource-intensive expert annotation.29

Zera Institute e.V.
Molkenmarkt 2
10179 Berlin

info@zera-institute.com
www.zera-institute.com

## Step 2: Understanding Phenomena and Trends

Armed with these sophisticated mixed-method tools, researchers can move beyond static description to dynamic understanding. This phase focuses on identifying how hate narratives—antisemitic, racist, misogynistic, conspiratorial, and anti-democratic—emerge, evolve, and proliferate across digital networks. Key objectives include mapping the mutation of hate tropes as they adapt to contemporary discourse, tracking transmission pathways across platforms and communities, identifying superspreader accounts and amplification mechanisms, detecting early warning signals of escalating radicalization and societal polarization, and understanding the interaction between offline events and online discourse intensification.

Recent research has identified a systematic three-phase pattern in how hate speech and disinformation spread through digital ecosystems:30 **(1) Primary discourse** involves elite figures articulating statements characterized by strategic ambiguity—rhetoric permitting multiple interpretations that appeal to divergent audiences while maintaining plausible deniability.31 **(2) Secondary discourse** occurs when digital intermediaries (YouTubers, TikTokers, podcasters, influencers) reframe elite messaging, typically reducing ambiguity by sharpening language and positioning it within explicit ideological frameworks. Unlike traditional journalists, these creators operate within attention economies that reward provocation and emotional intensity, often shaping downstream interpretation more powerfully than original statements. **(3) Tertiary discourse** unfolds in comment sections where audiences actively co-create meaning through participatory engagement. Here, ambiguity collapses: implicit meanings become explicit, political critique blurs into conspiracy theory, and extreme positions gain legitimacy through social proof rather than authoritative endorsement.

This three-phase circulation enables what researchers term "cascading radicalization"—ambiguous elite rhetoric becoming progressively more explicit and extreme through semantic crystallization (formulations becoming endorsements), emotional intensification (measured rhetoric becoming charged), and normalization of extremity (shocking statements becoming everyday discourse).32 Hate ideologies function as living, adaptive systems that respond to interventions, exploit platform vulnerabilities, and continuously evolve new expressions to evade detection through orthographic manipulation, compound neologisms, and rhetorical indirection.33 The aftermath of October 7, 2023 exemplified this dynamic through open celebration of violence, while reactions to the Washington museum shooting in May 2025 relied more heavily on conspiracy narratives, denial, and coded dismissal—illustrating how different triggering events activate different patterns within the broader spectrum of hate communication.34

Understanding these dynamics in real time provides the intelligence necessary for effective response across all categories of hate speech and disinformation.

Zera Institute e.V.
Molkenmarkt 2
10179 Berlin

info@zera-institute.com
www.zera-institute.com

## Step 3: Knowledge Transfer into Policy, Law, and Education

Research insights remain academic exercises unless translated into actionable frameworks for societal actors. This critical phase involves multiple dimensions of knowledge transfer:

**Policy Development** requires providing policymakers with evidence-based recommendations for platform regulation, content moderation standards, and legislative frameworks that balance free expression with protection from incitement. The European Union's Digital Services Act represents a first attempt to mandate greater transparency in content moderation and algorithmic processes, yet effective implementation requires collaboration with academic experts and access to high-quality datasets.35

**Legal Application** means equipping law enforcement and judicial systems with the knowledge to identify, investigate, and prosecute hate crimes while understanding the online radicalization pathways that precede offline violence. Legal frameworks must account for cross-jurisdictional differences: what constitutes protected speech in the United States may qualify as group defamation or hate speech under EU or German jurisprudence.

**Educational Intervention** involves developing curricula and pedagogical approaches informed by actual understanding of how young people encounter and internalize antisemitic narratives online. This includes creating media literacy programs that specifically address the rhetorical strategies and emotional appeals that make extremist content compelling—teaching students to recognize communication latency, coded language, and multimodal signaling. Such educational efforts are essential for countering radicalization and reducing societal polarization among younger generations.

**Platform Accountability** entails engaging technology companies with specific, evidence-based insights about how their algorithmic systems amplify harmful content and providing concrete recommendations for architectural changes that would reduce this amplification without requiring impossible levels of content moderation. Current moderation approaches often focus on slurs or direct threats while ignoring implicit or multimodal forms of hate speech, and enforcement remains inconsistent across languages and regions.36

Platform dynamics introduce additional complexity through what has been termed "context collapse"—the flattening of diverse audiences into a single public sphere.37 Statements crafted for one interpretive community may be decoded very differently by others when

they circulate across platforms. Social media users engage in "impression management" across collapsed contexts, potentially increasing reliance on ambiguous formulations that permit multiple readings. Algorithms further shape circulation patterns: content that generates engagement (including controversy) receives amplification regardless of whether that engagement reflects accurate interpretation. Understanding these dynamics

Zera Institute e.V.
Molkenmarkt 2
10179 Berlin

info@zera-institute.com
www.zera-institute.com

is essential for developing effective content moderation and platform governance strategies that address not merely individual harmful posts but the systemic amplification mechanisms that enable hate speech to spread.

A persistent challenge lies in the tension between over-blocking and under-blocking. When content is removed too aggressively, legitimate political critique or neutral references to Jewishness risk being censored. Research has shown that widely used moderation services systematically inflate toxicity scores for comments containing words like "Jew" or "Israel" regardless of stance.38 Conversely, when antisemitic content remains online, it risks normalization and causes emotional harm to Jewish users. Automated moderation systems must therefore achieve context-sensitive classification that distinguishes between, for example, "Israel is committing genocide" as contested political critique versus antisemitic Holocaust inversion.

## Step 4: Building Prerequisites for Effective Prevention and Intervention

The ultimate goal is not merely to understand antisemitism but to prevent its spread and intervene effectively when it emerges. This requires several interrelated capabilities:

**Early Warning Systems** deploy the computational tools developed in Step 1 to monitor online environments continuously, detecting concerning trends before they escalate into offline harm. Whatever unfolds during the next pandemic, global economic crisis, or climate-related resource shortage often simmers beforehand in the depths of the anonymous web. With appropriate analytical tools, we can detect early warning signs.

**Rapid Response Capabilities** create institutional structures that can quickly mobilize counter-narratives, fact-checking, and strategic communications when antisemitic disinformation campaigns emerge. This requires understanding not only what narratives circulate but also which superspreader accounts and amplification mechanisms drive their dissemination.

**Community Resilience** uses research insights to strengthen the psychological and social resources of communities targeted by hate, helping them recognize manipulation tactics and maintain cohesion in the face of attacks. Jewish communities expect protection from online antisemitism, yet poorly designed interventions risk re-traumatization or further marginalization.

**Rehabilitation Pathways** develops intervention strategies for individuals showing signs of radicalization, informed by understanding of what draws people into extremist ideologies

and what might draw them back out. This requires mapping not merely explicit expressions but the gradual radicalization process through which distorted ideas solidify into fixed belief systems. Effective de-radicalization programs must address the societal polarization that creates fertile ground for extremist recruitment.

Zera Institute e.V.
Molkenmarkt 2
10179 Berlin

info@zera-institute.com
www.zera-institute.com

This roadmap represents a comprehensive response to hate ideologies that matches the sophistication of the threat. Just as modern medicine relies on molecular biology to develop targeted therapies, modern democracy requires computational social science to develop targeted interventions against hate in all its forms—antisemitism, racism, misogyny, conspiracy thinking, and anti-democratic disinformation. These viruses of hatred have found new vectors of transmission in social media; we must become the virologists who decode their mechanisms and develop the remedies our societies desperately need.

While social media has emerged as one of the most significant destabilizers of our time, it also presents a tremendous opportunity. Over the past century, our societies have made varying degrees of effort to eradicate the deeply ingrained hatred that has permeated cultures for millennia. This enduring challenge prompts many to doubt whether this ideology of hate can ever be completely eradicated.

Traditionally, surveys were our primary means of gauging the prevalence of antisemitism in society. However, they pale in comparison to the insights gleaned from thorough analysis of social media discourses. Social media interactions capture voluntary expressions, dialogues among thousands of individuals, and the entire spectrum of human interaction. Understanding online discourse equates to understanding the formation of attitudes within our society—a level of insight we have never before possessed in human history.

The thorough identification of predominant communication patterns online extends beyond documenting the present; it can illuminate the future. Understanding the emergence of trends within our society and deciphering how hatred manifests, through which channels, and employing what messaging strategies can reveal where future harm is likely to occur.

To grasp the forthcoming direction of our society—encompassing both online trends and their offline ramifications—rigorous interdisciplinary research into antisemitism, racism, misogyny, conspiracy thinking, and other hate-based ideologies threatening our free societies, democratic political systems, and social cohesion is imperative. This research must utilize all available methodological approaches. Social scientists and data science experts must engage in ongoing dialogue and mutual learning. While political interventions or platform regulations alone cannot solve these threats to democracy, neither can artificial intelligence address them adequately in isolation. Only through collaborative efforts can we discern, in real time, the evolution of online debates, the resonance and dissemination of specific extremist narratives and disinformation campaigns, and the individuals or groups who wield influence in shaping the interpretation of critical issues.

This text serves not only as an assessment of hate-related social media studies but also

as a plea. As a society, we must acknowledge that the traditional methods of upholding social peace, embodied in our political, legal, educational, and security institutions, are inadequate for navigating this new digital domain. We must dedicate our efforts not merely to exploring these dynamics with a few spotlights, but rather to comprehensively understanding the phenomenon to identify both the symptoms, their underlying causes,

Zera Institute e.V.
Molkenmarkt 2
10179 Berlin

info@zera-institute.com
www.zera-institute.com

and the potential consequences. This will enable us to initiate a truly well-informed dialogue with the drifting segments of society. For this, we need robust research and fruitful collaborations between academia, the political arena, and civil society—developing methodologies that work across all forms of hate speech, from antisemitism and racism to misogyny and anti-democratic disinformation. Only by decoding the discourses that shape our digital age can we begin to rebuild the moral grammar of democratic society.

## Footnotes

1. Lars Rensmann, "Politischer Antisemitismus im postfaktischen Zeitalter: Formen und Ursachen in Demokratien des 21. Jahrhunderts" (Baden-Baden: Nomos, 2025); Heiko Beyer et al., "Antisemitismus in der Gesamtgesellschaft von Nordrhein-Westfalen im Jahr 2024," Antisemitismusbeauftragte des Landes Nordrhein-Westfalen (2024). ↩
2. Gabriele Cosentino, "Social Media and the Post-Truth World Order: The Global Dynamics of Disinformation" (Cham: Palgrave Macmillan, 2020). ↩
3. Anna Lührmann and Staffan I. Lindberg, "A Third Wave of Autocratization is Here: What is New About it?" *Democratization* 26, no. 7 (2019): 1095–1113. ↩
4. Institute for Strategic Dialogue, "Information Laundromat Detects Banned Russian Propaganda Across Hundreds of Websites," accessed October 2024, https://www.isdglobal.org/isd-in-the-news/information-laundromat-detects-banned-russian-propaganda-across-hundreds-of-websites; Institute for Strategic Dialogue, "Russia-Ukraine War," accessed October 2024, https://www.isdglobal.org/tag/russia-ukraine-war; Institute for Strategic Dialogue, "China, Russia and Iran Are Exploiting the Israel-Hamas Conflict for Their Advantage," accessed October 2024, https://www.isdglobal.org/isd-in-the-news/china-russia-and-iran-are-exploiting-the-israel-hamas-conflict-for-their-advantage; see also Deen Freelon and Chris Wells, "Disinformation as Political Communication," *Frontiers in Political Science* 4 (2022), https://doi.org/10.3389/fpos.2022.885362. ↩
5. "Facebook Whistleblower Frances Haugen Testified That the Company's Algorithms Are Dangerous—Here's How They Can Manipulate You," *The Conversation*, October 5, 2021, https://theconversation.com/facebook-whistleblower-frances-haugen-testified-that-the-companys-algorithms-are-dangerous-heres-how-they-can-manipulate-you-169420. ↩
6. Ibid. ↩
7. Lars Rensmann, "The Noisy Counter-Revolution: Understanding the Cultural Conditions and Dynamics of Populist Politics in Europe in the Digital Age," *Politics and Governance* 5, no. 4 (2017): 123–135. ↩
8. Cal Newport, "The Real Problem with Twitter," accessed October 2024, https://calnewport.com/the-real-problem-with-twitter. ↩

Zera Institute e.V.
Molkenmarkt 2
10179 Berlin

info@zera-institute.com
www.zera-institute.com

9. Matthias J. Becker and Lars Rensmann, "The Dynamics of Digital Antisemitism: Anti-Jewish Narratives in the Aftermath of October 7th," in *Antisemitism Online: An Ancient Hatred in the Modern World*, ed. Todd Pittinsky (Oxford: Oxford University Press, in press). ↩
10. Ibid. ↩
11. USC Shoah Foundation, "Facing Antisemitism," accessed October 2024, https://sfi.usc.edu/video/facing-antisemitism. ↩
12. Decoding Antisemitism, accessed October 2024, https://decoding-antisemitism.eu. ↩
13. Matthias J. Becker et al., eds., *Decoding Antisemitism: An AI-Driven Study on Hate Speech and Imagery Online* (Cham: Springer, 2024), https://doi.org/10.1007/978-3-031-49237-2. ↩
14. Matthias J. Becker et al., *Celebrating Terror: Antisemitism Online After the Hamas Attacks on Israel. Preliminary Results I* (Berlin: Technische Universität Berlin, Center for Research on Antisemitism, 2023), https://doi.org/10.14279/depositonce-19143. ↩
15. Matthias J. Becker, Jordan Blatter, and Oksana Stanevich, "Decoding Antisemitism Online: Linguistic and Multimodal Challenges in the Age of AI" (working paper, 2025). ↩
16. Werner Bergmann and Rainer Erb, "Kommunikationslatenz, Moral und öffentliche Meinung. Theoretische Überlegungen zum Antisemitismus in der Bundesrepublik Deutschland," *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 38 (1986): 223–246. ↩
17. Becker and Rensmann, "The Dynamics of Digital Antisemitism." ↩
18. Becker, Blatter, and Stanevich, "Decoding Antisemitism Online." ↩
19. Parker Bach, Carolyn E. Schmitt, and Shannon C. McGregor, "Let Me Be Perfectly Unclear: Strategic Ambiguity in Political Communication," *Communication Theory* 35, no. 2 (2025): 96–106; Matthias J. Becker, Gabrielle Beacken, and Liora Sabra, "How Political Ambiguity Transforms Across Digital Platforms: Mechanisms of Discourse Radicalization" (working paper, 2025). ↩
20. Ibid. ↩
21. Katharina Soemer, Daniela Gurschow, and Steffen Egert, "Social Science and AI Joining Forces: Towards New Approaches for Computational Social Science" (paper presented at the interdisciplinary conference on computational methods, 2025). ↩ ↩2
22. International Holocaust Remembrance Alliance, "Working Definition of Antisemitism" (2016), https://holocaustremembrance.com/resources/working-definition-antisemitism; Becker et al., *Decoding Antisemitism: An AI-Driven Study*. ↩
23. Becker, Blatter, and Stanevich, "Decoding Antisemitism Online." ↩
24. Ibid. ↩
25. Matthias J. Becker and Hagen Troschke, "Decoding Implicit Hate Speech: The Example of Antisemitism," in *Challenges and Perspectives of Hate Speech Research*, ed. C. Strippel et al., *Digital Communication Research* 12 (2023): 335–352, https://doi.org/10.48541/dcr.v12.20. ↩

Zera Institute e.V.
Molkenmarkt 2
10179 Berlin

info@zera-institute.com
www.zera-institute.com

26. Becker, Blatter, and Stanevich, "Decoding Antisemitism Online." ↩
27. Jordan Blatter and Blue Square Alliance Against Hate, "Testing Next-Generation AI to Better Detect Antisemitism," Blue Square Alliance Command Center Insights (2025), https://www.bluesquarealliance.org/command-center-insights/testing-next-generation-ai-to-better-detect-antisemitism. ↩
28. Milena Pustet and Helena Mihaljević, "Automated Detection of Antisemitic Texts: Is Context All We Need?" in *Decoding Antisemitism: An AI-driven Study on Hate Speech and Imagery Online. Discourse Report 6*, ed. Matthias J. Becker et al. (Berlin: Technische Universität Berlin, Center for Research on Antisemitism, 2024), https://decoding-antisemitism.eu/publications/sixth-discourse-report. ↩
29. Becker, Blatter, and Stanevich, "Decoding Antisemitism Online"; Eva Steffen, Milena Pustet, and Helena Mihaljević, "Algorithms Against Antisemitism? Towards the Automated Detection of Antisemitic Content Online," in *Antisemitism in Online Communication: Transdisciplinary Approaches to Hate Speech in the Twenty-First Century*, ed. Matthias J. Becker et al. (London: Open Book Publishers, 2024), https://doi.org/10.11647/OBP.0406.08. ↩
30. Becker, Beacken, and Sabra, "How Political Ambiguity Transforms Across Digital Platforms." ↩
31. Bach, Schmitt, and McGregor, "Let Me Be Perfectly Unclear." ↩
32. Becker, Blatter, and Stanevich, "Decoding Antisemitism Online." ↩
33. Ibid. ↩
34. Becker and Rensmann, "The Dynamics of Digital Antisemitism." ↩
35. Jash Patel, Heer Mehta, and Jeremy Blackburn, "Evaluating Large Language Models for Detecting Antisemitism," arXiv preprint arXiv:2509.18293 (2025), accepted to EMNLP 2025 Main Conference. ↩
36. Ibid. ↩
37. danah boyd, "Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications," in *A Networked Self: Identity, Community, and Culture on Social Network Sites*, ed. Zizi Papacharissi (New York: Routledge, 2010), 39–58. ↩
38. Alexis Chapelan et al., *Decoding Antisemitism: An AI-driven Study on Hate Speech and Imagery Online. Discourse Report 5* (Berlin: Technische Universität Berlin, Center for Research on Antisemitism, 2023), https://doi.org/10.14279/depositonce-17105; Lucas Dixon et al., "Measuring and Mitigating Unintended Bias in Text Classification," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)* (2018), 67–73. ↩