Governing Disinformation Through Preventive Digital Policing: Technology, Legitimacy, and Social Implications in Indonesia

Andhika Raka Nurizki[1], Chairul Muriman Setyabudi[2], Solikhah Yuliatiningtyas[3]
[1,2,3]Universitas Indonesia, Depok, Indonesia
Email: andhika.raka@ui.ac.id

Abstract

The rapid expansion of digital technologies and social media has intensified the spread of disinformation and hate speech, creating significant social and governance challenges. While existing studies largely focus on platform governance, algorithmic regulation, or coercive law enforcement, limited attention has been given to preventive digital policing as a form of technology-enabled social governance. This study examines how preventive digital policing is implemented by Indonesia's Criminal Investigation Agency (Bareskrim Polri) in addressing disinformation and hate speech, with particular emphasis on legitimacy and public trust. Using an exploratory case study approach, this research relies on secondary data, including official institutional documents, policy reports, and peer-reviewed academic literature. The data were analyzed through thematic analysis to identify patterns in technological adoption, preventive strategies, and institutional challenges. The findings show that Indonesia's digital policing strategy prioritizes early detection through open-source intelligence (OSINT), multi-platform monitoring, and basic artificial intelligence tools, combined with non-coercive interventions such as digital warnings, online mediation, and public education initiatives. These measures emphasize persuasion and early intervention rather than punitive enforcement. However, the study also identifies key challenges related to technological limitations, regulatory ambiguity, institutional capacity, and public perceptions of surveillance and control. The findings suggest that the effectiveness of preventive digital policing depends less on technological sophistication than on transparency, procedural fairness, and societal acceptance. This study contributes to the science and society literature by reconceptualizing digital policing against disinformation as a legitimacy-dependent preventive governance mechanism, highlighting the interaction between technology, state authority, and social trust in a Global South context.

Keywords: *Preventive Digital Policing, Disinformation Governance, Technology and Society, Police Legitimacy, Indonesia.*

A. INTRODUCTION

The rapid expansion of digital technologies and the ubiquity of social media platforms have fundamentally reshaped the architecture of contemporary societies, transforming the fundamental ways in which information is produced, circulated, and consumed. While these developments have undeniably facilitated broader access to information and enhanced public participation, they have simultaneously created fertile ground for the intensification of disinformation and hate speech. These phenomena generate profound social, political, and governance challenges that threaten the stability of democratic discourse.

The digital landscape in Indonesia has undergone a seismic shift over the past decade, evolving into one of the world's most active and dense online ecosystems.

According to the latest data from the Indonesian Internet Service Providers Association (APJII, 2024), internet penetration in the archipelago has reached a staggering 221.5 million users. This vast connectivity, while a testament to rapid technological advancement, has fundamentally altered the sociological fabric of the nation. Social media serves as the primary gateway for this massive population, acting not just as a tool for communication, but as the dominant infrastructure for information consumption, political discourse, and social interaction.

However, this massive connectivity harbors a significant "dark side" that poses an existential threat to public order and democratic health. The democratization of information production has led to a decentralization of authority, where traditional gatekeepers are bypassed by algorithmically driven platforms. Reports by MAFINDO (Masyarakat Anti Fitnah Indonesia, 2024) highlight an alarming trend: thousands of unique hoaxes and disinformation threads are identified annually, with a notable and dangerous "spike" occurring during political cycles, such as the 2019 and 2024 General Elections. These hoaxes are not random; they are often strategically crafted to exploit existing social fissures, utilizing what scholars describe as the "economy of emotions."

Empirical evidence from global studies reinforces the severity of this crisis. Research by Vosoughi, Roy, and Aral (2018) and Cinelli et al. (2020) has consistently demonstrated that false and emotionally charged content, particularly that which triggers fear, anger, or disgust, spreads with 70% greater velocity and reaches a deeper audience than verified, factual information. In the Indonesian context, this is exacerbated by "algorithmic silos" or echo chambers, where platform algorithms prioritize "viral potential" and user engagement over accuracy or truth. This creates a feedback loop where disinformation is amplified by the very technology meant to connect the public, leading to what is termed an "infodemic."

The consequences of this digital crisis are not confined to the virtual world; they translate directly into tangible, real-world harm and physical violence. Disinformation in Indonesia has evolved into a potent catalyst for horizontal violence and social destabilization. Case studies of social conflict in regions such as Papua demonstrate how a single viral hoax can trigger localized riots and widespread unrest. Similarly, during the 2019 and 2024 elections, coordinated disinformation campaigns were used to delegitimize electoral processes, leading to civil disturbances and deep-seated social polarization (Stewart, McCarty, & Bryson, 2020).

This polarization is often "affective," meaning it is not just a disagreement over policy, but a deep-seated animosity toward those with differing views, fueled by partisan sorting and inflammatory digital content (Tornberg, 2022). Furthermore, the persistent exposure to manipulated information contributes to the erosion of trust in public institutions. When the line between truth and falsehood becomes permanently blurred, the public's confidence in the government, the judiciary, and law enforcement agencies like the Police is significantly undermined. This breakdown in trust creates a "security vacuum" that further complicates the state's ability to govern effectively in the digital age. Thus, the digital crisis in Indonesia is not merely a technical problem of "fake news," but a complex socio-political challenge that requires a fundamental

shift in how the state conceptualizes security and guardianship in a hyper-connected society.

To analyze the emergence and governance of digital harms in the Indonesian context, this study utilizes the intersection of Routine Activity Theory (RAT) and Situational Crime Prevention (SCP). These criminology frameworks, traditionally applied to physical spaces, provide a robust analytical lens for understanding how Bareskrim Polri has reconfigured its operational logic to address the borderless nature of social media.

According to the foundational work of Cohen and Felson (1979), the occurrence of a crime or in this context, the dissemination of harmful disinformation is not merely the result of social pathology but requires the convergence in time and space of three essential elements: a motivated offender, a suitable target, and the absence of a capable guardian. In the digital realm, this convergence is facilitated by the unique affordances of social media platforms. The motivated offenders are actors who produce or amplify hoaxes and hate speech, often driven by political, ideological, or economic motives (Bradshaw & Howard, 2019). The suitable targets are the vast pool of Indonesian internet users, reaching 221 million by 2024, many of whom are susceptible to emotionally charged content due to varying levels of digital literacy (Vosoughi, Roy, & Aral, 2018).

The critical component in Bareskrim's digital transformation is the effort to establish a "capable guardian" within a virtual space that was previously perceived as a lawless frontier. As Leukfeldt and Yar (2016) argue, the internet removes physical barriers, allowing offenders to strike targets remotely. By integrating Open-Source Intelligence (OSINT) and automated monitoring, Bareskrim Polri acts as a digital guardian that "observes" the digital drift of information, aiming to disrupt the opportunity for disinformation to reach its target.

Complementary to RAT is the framework of Situational Crime Prevention (SCP), which shifts the focus from the offender's disposition to the environment in which the crime occurs. SCP, as theorized by Newman, Clarke, and Shoham (2021), emphasizes the reduction of criminal opportunities through five main strategies: increasing the effort, increasing the risks, reducing the rewards, reducing provocations, and removing excuses.

The implementation of the Virtual Police and the Virtual Alert system by Bareskrim Polri serves as a prime empirical example of SCP in action. By identifying potentially illegal content, such as hate speech or hoaxes, and issuing a direct, formal warning to the user via Direct Message (DM) or public comments, the police effectively increase the perceived risk of detection. This intervention functions as a "digital nudge" or a "soft policing" tactic.

Instead of immediate criminalization under the ITE Law, which often results in protracted legal battles and public controversy, the Virtual Alert aims to *remove excuses* for the offender by informing them that their content violates specific regulations. This preventive mechanism seeks to de-escalate the viral momentum of a post before it can trigger real-world social conflict. This approach aligns with the logic of "target

hardening," where the digital discourse environment is made less conducive to the spread of disinformation through the visible presence of state authority. By prioritizing "perceived risk" over "actual punishment," Bareskrim Polri attempts to govern the information space through a more efficient, technology-enabled preventive paradigm (Ferguson, 2017; Montasari, Carpenter, & Masys, 2023).

Indonesia serves as a critical laboratory for observing the evolution of modern law enforcement in the face of digital disruption. As the nation grappled with a persistent "infodemic," Bareskrim Polri (the Criminal Investigation Agency) recognized that traditional, physically bound policing methods were insufficient to curb the velocity of online harms. Consequently, the agency has undergone a structural and philosophical metamorphosis, shifting from a reactive "arrest-and-prosecute" stance to a proactive, preventive digital paradigm. This transformation is not merely technological but represents a fundamental shift in the "science of policing" toward maintaining social order in a hyper-connected society.

The first pillar of this transformation is the institutionalization of Open-Source Intelligence (OSINT) and data analytics. Bareskrim has moved beyond simple monitoring to a sophisticated mapping of the "digital drift", the process by which information migrates across platforms, often gaining toxicity as it moves (Goldsmith & Brewer, 2015). By utilizing metadata analysis and social network mapping, the agency can identify "account clusters" and coordinated inauthentic behavior (CIB).

Evidence from recent operations suggests that this capability allows the agency to pinpoint the "patient zero" of a viral hoax before it achieves critical mass. This data-driven approach allows the police to understand the architecture of disinformation networks, distinguishing between organic public frustration and manufactured digital campaigns designed to trigger horizontal conflict. This shift aligns with the "New Visibility" of policing, where the state uses the same digital tools as the public to assert its role as a guardian (Goldsmith, 2010).

Perhaps the most visible manifestation of this transformation is the Virtual Police initiative. Since its inception, the unit has issued thousands of "Virtual Alerts"—direct, standardized warnings sent via social media platforms to users identified as spreading potentially illegal content. This mechanism functions as a "digital nudge," a form of soft policing that prioritizes correction over criminalization.

A relevant case study in the efficacy of this approach can be observed during recent regional elections. Bareskrim records indicate that early interventions by the Virtual Police led to a significant reduction in hate speech momentum. In many instances, users who received a "Virtual Alert" opted for the voluntary deletion of their inflammatory posts, thereby neutralizing the threat of social escalation without the need for a physical arrest. This demonstrates a shift toward Situational Crime Prevention, where the police "harden" the digital environment by making the risk of detection immediate and personal (Newman, Clarke, & Shoham, 2021).

The third key piece of evidence in this transformation is the increased reliance on Restorative Justice and online mediation. For years, the ITE Law (Electronic Information and Transactions Law) was criticized for being a "rubber article" used for

excessive criminalization. In response, Bareskrim has established new Standard Operating Procedures (SOPs) that prioritize mediation, especially for cases of minor defamation or non-systemic hate speech.

Data from Bareskrim indicates a steady rise in the percentage of digital disputes resolved through restorative mechanisms rather than formal prosecution. This reflects a sophisticated understanding of Police Legitimacy; by acting as a mediator rather than a coercive force, the police attempt to build trust and social cohesion (Tyler, 2006; Bradford et al., 2017). This strategic shift suggests that the future of digital governance lies in the ability of state institutions to foster dialog and education, effectively transforming the police from "watchmen" into "trusted digital mediators" within the Indonesian information ecosystem (Montasari, Carpenter, & Masys, 2023).

Despite the significant technological advancements integrated into Indonesia's security infrastructure, the transition to a digital policing paradigm is not without its fundamental contradictions. The findings of this study suggest that the operational success of these tools is frequently bottlenecked by what can be termed the Legitimacy-Technology Paradox. This paradox occurs when an increase in technological surveillance and intervention capability does not lead to a corresponding increase in social order, but instead results in heightened public suspicion and resistance.

The core of this challenge lies in the sociological principles of Police Legitimacy and Procedural Justice. As argued by Tyler (2006) and Bradford et al. (2017), the effectiveness of law enforcement is not derived solely from the state's capacity to exercise power or deploy sophisticated technology; rather, it is deeply rooted in the public's perception of fairness. People are far more likely to comply with legal directives and engage in self-regulation when they believe that the authorities are acting with transparency, neutrality, and respect for individual rights. In the digital sphere, where the line between "protection" and "surveillance" is increasingly thin, these perceptions of legitimacy become the primary filter through which the public receives police interventions.

In the Indonesian context, this gap in legitimacy is exacerbated by the legal framework governing digital interactions most notably the ITE Law (Electronic Information and Transactions Law). While Bareskrim Polri frames initiatives like the Virtual Police as educational and preventive, the underlying law is often criticized for containing "multi-tafsir" or highly ambiguous articles (Afisa et al., 2024). When the criteria for what constitutes "dangerous disinformation" or "hate speech" remain subjective or poorly defined, the public may view a "Virtual Alert" not as a helpful nudge from a "capable guardian," but as a targeted instrument of state surveillance designed to suppress dissent.

This perception leads to a significant "chilling effect" on digital discourse. Instead of fostering a more literate and responsible digital society, intrusive policing, if perceived as illegitimate, may simply drive users to more encrypted, less transparent platforms or silence legitimate democratic expression altogether. As Chan and Bennett Moses (2016) observe, when big data tools are deployed in environments

with low institutional trust, they risk being seen as tools for "digital social control" rather than public safety.

Against this background, the central research question of this study emerges: How is preventive digital policing implemented to govern disinformation and hate speech in Indonesia, and how does legitimacy shape its effectiveness as a technology-enabled governance strategy?

Accordingly, the objective of this study is to move beyond a purely technical evaluation of OSINT and AI tools. Instead, it seeks to examine the implementation of Bareskrim's digital practices through an interdisciplinary lens, analyzing the complex interaction between technological practices, institutional legitimacy, and social trust. By investigating these dynamics, the research aims to provide a critical contribution to the "science of policing" and the broader field of digital governance, specifically highlighting the unique challenges faced by Global South democracies in navigating the promises and perils of digital transformation (Montasari, Carpenter, & Masys, 2023).

B. METHOD

This study employs an exploratory case study approach to examine the digital transformation within the Criminal Investigation Agency of the Indonesian National Police (Bareskrim Polri). This approach is particularly suitable for investigating contemporary issues in their real-life context, emphasizing the "how" and "why" behind emerging patterns of technology-driven policing. The exploratory design enables an in-depth examination of internal procedures, regulatory directives, and the transition from traditional investigative methods to a preventive, digitally-oriented policing paradigm.

Data for this study were collected through interviews with key personnel at Bareskrim Polri in Jakarta, including the Head of Cyber Operations (Kasubag Operasional Cyber) and the Head of the Cyber Forensics Laboratory Unit (Kasubnit Lab Forensik Cyber). In addition, the study involved a review of direct reports and database records within the Polri information system. The research was conducted over a two-and-a-half-year period, from early 2023 to mid-2025, at Bareskrim Polri in Jakarta. To ensure a comprehensive analysis while maintaining academic rigor, the research relies entirely on secondary data. The data corpus includes:

1. Official Institutional Documents**:** Reports and official publications from Bareskrim Polri and related government agencies.
2. Academic Literature**:** Peer-reviewed journal articles, scholarly books, and scientific publications regarding cybercrime, digital policing, and state policy.
3. Policy and Legal Frameworks**:** National regulations such as the Electronic Information and Transactions (ITE) Law, internal Standard Operating Procedures (SOPs), and Circular Letters from the Chief of Police.

The collected data underwent thematic analysis to identify, analyze, and report patterns (themes) within the data. This process involved:

1. Categorization**:** Organizing information based on the core research questions: the forms of digital transformation, the integration of technology in prevention, and the resulting challenges.
2. Theoretical Triangulation**:** Validating findings by filtering them through multiple theoretical lenses, including Routine Activity Theory (RAT), Situational Crime Prevention (SCP), and Digital Policing Theory.
3. Synthesis**:** Integrating institutional data with social theories to evaluate the effectiveness and legitimacy of the strategies implemented.

The reliability of the findings is supported by the use of authoritative secondary sources and established academic frameworks. By grounding the analysis in well-documented institutional reports and verified scientific literature, the study minimizes subjective bias. Furthermore, the appropriateness of the exploratory case study method is aligned with established research standards for investigating evolving institutional shifts where the boundaries between the phenomenon and context are not clearly evident (Yin, 2017).

C. RESULTS AND DISCUSSION

1. The Digital Transformation of Bareskrim Polri in Governing Disinformation

The research findings indicate that the digital transformation within Indonesia's Criminal Investigation Agency (Bareskrim Polri) represents a strategic adaptation to the rapid evolution of digital crimes and communication patterns over the last decade. This transformation is characterized by a fundamental shift in the policing paradigm moving from traditional, reactive law enforcement toward a *preventive digital policing* model that prioritizes data-driven insights, early intervention, and multi-stakeholder collaboration. This shift is essential given that digital technologies have profoundly reshaped how information is produced and consumed, while simultaneously intensifying the spread of harmful content.

The first major discovery of this research is the development of a sophisticated technical infrastructure designed to act as a "capable guardian" in digital spaces. Bareskrim has established a multi-layered technological framework to monitor the vast landscape of the Indonesian internet, which in 2024 reached over 221 million users (APJII 2024).

a. *Open-Source Intelligence (OSINT) and Data Analytics***:** OSINT has become the cornerstone of digital policing within Bareskrim. By utilizing metadata analysis, social network mapping, and account cluster identification, the agency can track the origin and flow of disinformation in real-time. This system allows authorities to see conversation patterns and viral trends before they escalate into physical conflict. The development of this OSINT system aligns with international standards observed in Europe, the United States, and Australia for the early detection of digital threats (Crawford and Hutchinson 2016)
b. *Specialized Cyber Units (Dittipidsiber)***:** Bareskrim, through the Directorate of Cyber Crimes (Dittipidsiber), has established a dedicated analysis unit

specifically for the structured monitoring of disinformation, hate speech, and cyber threats. This represents a significant evolution from traditional policing models, providing a centralized hub for digital intelligence that did not exist previously.

c. *Implementation of Artificial Intelligence (AI)*: To manage the massive volume of social media data, Bareskrim utilizes AI-driven tools for automated content classification. Technologies such as keyword-based filtering, sentiment analysis, and pattern recognition are employed to select and prioritize potentially harmful content. While effective for large-scale selection, findings indicate these tools still face limitations in understanding linguistic nuances, irony, and cultural sarcasm.

d. *Multi-Platform Monitoring*: Recognizing that disinformation is inherently cross-platform—often moving between different social media sites to hide its origins and broaden its reach—Bareskrim conducts surveillance across various platforms. This monitoring is tailored to the unique "affordances" of each platform, such as visibility and spreadability.

The shift toward prevention is manifested in several non-coercive digital policing initiatives designed to reduce the need for repressive legal action and de-escalate social tensions.

a. *Virtual Police and Virtual Alert*: One of the most significant innovations is the *Virtual Police* initiative. When a user is identified as spreading potentially illegal disinformation or hate speech, the Virtual Police issues a "Virtual Alert", a direct, educational warning sent via digital channels. This serves as a "digital nudge" to inform the user of potential legal violations before formal prosecution is initiated, reflecting a shift toward preventive education.

b. *Online Mediation and Restorative Justice*: In line with global trends in modern policing, Bareskrim increasingly adopts a restorative approach to digital crimes. This involves resolving digital disputes, particularly those involving minor hate speech or defamation, through online mediation rather than immediate criminalization. This "soft policing" approach aims to reduce social tension and minimize excessive criminality.

c. *Public Reporting Portal (patrolisiber.id)*: To increase transparency and community involvement, Bareskrim operates *patrolisiber.id*. This platform allows the public to report problematic content quickly, effectively turning citizens into active participants in the digital security ecosystem and increasing the agency's responsiveness.

d. *Early Warning Systems*: The integration of OSINT and AI allows for the identification of potential threats before they reach a peak viral state. This "early warning" strategy is consistent with the principles of Situational Crime Prevention (SCP), which emphasizes reducing opportunities for criminal actions.

Despite significant technological progress, the implementation of preventive digital policing faces a complex set of internal and external challenges.

Table 1. Key Challenges in Bareskrim Polri's Digital Transformation

Category	Specific Findings and Data Points
Technical & Algorithmic	Rapid evolution of platform algorithms outpaces monitoring tools. AI faces difficulty in accurately interpreting local dialects, irony, and sociocultural contexts. Data security and privacy risks remain a persistent concern.
Human Resources (SDM)	Critical shortage of high-competency data analysts, digital forensic experts, and OSINT specialists. High personnel workload and existing competency gaps hinder development.
Legal & Regulatory	Ambiguity in the ITE Law regarding hate speech definitions, which can be multi-interpretable and lead to public controversy. Lack of regulatory alignment between national laws and the global policies of social media platforms.
Institutional Coordination	Coordination hurdles between Bareskrim, government agencies (Komdigi), and international platform giants due to differing data access policies.
Public Legitimacy	Public perception of "over-surveillance". Some view *Virtual Police* as a tool to control opinions or "scare" the digital society rather than as an educational measure.

Source: Data synthesized and processed by the author from various institutional reports and academic literature (2025).

The research identified several indicators of how these digital strategies are performing in the field, highlighting both successes and inherent limitations.

a. Reduction in Viral Momentum: In many cases, the swift action of the *Virtual Police* in providing clarifications and warnings has led to a noticeable shrinkage in the viral momentum of harmful content. By moving quickly, the police can neutralize hoaxes before they become entrenched.
b. Speed of Response: The integration of digital reporting and OSINT has increased Bareskrim's ability to act on problematic content before it escalates into physical social conflict.
c. Limitations - The "Algorithmic Noise": Effectiveness is often dampened by "algorithmic noise," where social media algorithms prioritize emotionally charged, controversial content. This often results in official police corrections being drowned out by viral misinformation.
d. The Literacy Gap: Low levels of digital literacy among a large portion of the population remain a primary driver of hoax re-circulation. Interventions are often temporary if the public lacks the skills to independently verify information.

A critical part of the transformation involves strengthening internal literacy and expertise.

a. Internal Digital Literacy: Bareskrim has developed internal SOPs, content classification guides, and OSINT modules for investigators. This is based on the

understanding that the success of digital policing depends heavily on the human ability to understand, operate, and evaluate the technology.

b. Shift in Organizational Culture: The move toward preventive measures indicates a change in organizational culture from purely reactive enforcement to a more proactive, educational, and collaborative stance.

In summary, the results demonstrate that Bareskrim Polri has successfully established a technological and operational foundation for Preventive Digital Policing. The agency has moved toward a more human-centric, dialog-based approach through initiatives like the *Virtual Police* and online mediation. However, the ultimate success of this transformation is not purely a technological matter; it is deeply tied to the institution's ability to navigate regulatory ambiguity, bridge the human resource gap, and build lasting legitimacy in the eyes of a skeptical digital public.

2. Reconceptualizing Digital Policing as Preventive Governance

The implementation of preventive digital policing by Indonesia's Criminal Investigation Agency (Bareskrim Polri) represents a critical evolution in the intersection of technology, state authority, and social governance. This discussion synthesizes the research findings with established theoretical frameworks to address the primary research question: How is preventive digital policing implemented to govern disinformation in Indonesia, and how does legitimacy shape its effectiveness?

The transition of Bareskrim Polri toward a technology-enabled preventive model can be critically analyzed through the lens of Routine Activity Theory (RAT) and Situational Crime Prevention (SCP). According to Cohen and Felson (1979), the occurrence of "crimes" in digital spaces—such as the viral spread of hate speech or coordinated disinformation—requires the convergence of a motivated offender, a suitable target (vulnerable public), and the absence of a capable guardian.

This study finds that Bareskrim's adoption of Open-Source Intelligence (OSINT) and automated monitoring serves as a structural attempt to re-insert the "capable guardian" into a borderless digital environment. Unlike traditional policing that relies on physical presence, digital guardianship here is exercised through "data presence." The findings regarding Virtual Police and Virtual Alerts directly align with SCP principles by increasing the "perceived risk" and "effort" for potential offenders (Newman, Clarke and Shoham, 2021). By issuing a digital warning before a crime is fully consummated, the state is effectively "hardening the target" of the public discourse.

However, a critical analysis suggests that this guardianship is uniquely constrained by the affordances of social media. As Evans et al. (2017) argue, technology is not neutral; its design influences behavior. The *spreadability* and *anonymity* of platforms often outpace the speed of police detection. Therefore, the "novelty" found in Indonesia's approach is the attempt to use the platform's own affordances—direct messaging and public tagging—to execute policing actions. This represents a shift from "policing the person" to "policing the information flow".

A central theme of this research, as stated in the introduction, is that technological capability alone does not guarantee effectiveness. The findings regarding public skepticism toward "over-surveillance" highlight what this study terms the Legitimacy-Technology Paradox. Based on Police Legitimacy Theory (Tyler, 2006; Bradford et al., 2017), the effectiveness of state intervention depends on the public's perception of procedural fairness and transparency.

In the Indonesian context, the use of OSINT and AI to select content for "Virtual Alerts" is often shrouded in institutional opacity. When the criteria for what constitutes "dangerous disinformation" remain ambiguous—compounded by the multi-interpretable nature of the ITE Law—the public begins to view preventive policing as a form of "technology-enabled social control" rather than "public protection." The findings suggest that when legitimacy is low, the "Virtual Alert" does not function as an educational nudge but as a "chilling effect" on digital expression. This validates the gap identified in the literature: in Global South contexts, state-led digital interventions are often viewed through the lens of political authority rather than neutral governance (Chan and Bennett Moses, 2016).

The research results identified significant hurdles in human resources and regulatory ambiguity. Analyzing these through Policy Implementation Theory (Van Meter and Van Horn, 1975; Hill and Hupe, 2001), it is evident that the "disposition of implementers" and "resource availability" are the weakest links in Bareskrim's digital transformation.

The "novelty" of this finding lies in the realization that Indonesia has acquired "First World" technology (AI, OSINT) but operates it within a "Global South" institutional and regulatory framework. The shortage of data scientists and the reliance on general investigators to make complex socio-linguistic judgments on hate speech leads to "implementation slippage." As noted in the results, the AI struggle with sarcasm and local irony means the human-in-the-loop is frequently overwhelmed. This gap between the "high-tech" image and the "limited-capacity" reality creates a friction that undermines the governance of disinformation.

This study contributes to the Science and Society literature by reconceptualizing digital policing not as a punitive law enforcement function, but as a legitimacy-dependent governance mechanism. The shift toward Online Mediation and Restorative Justice represents a departure from the "coercive law enforcement" focus mentioned in the introduction (Bakir and McStay, 2018; Ferguson, 2017).

By prioritizing persuasion over punishment, Bareskrim is attempting to manage "social trust" as a strategic asset. The study finds that the effectiveness of governing disinformation depends less on the sophistication of the OSINT algorithm and more on whether the "counter-narrative" provided by the state is trusted by the citizenry. If the state is viewed as a biased actor, its educational efforts—no matter how technologically advanced—will be rejected by "echo chambers" (Sunstein, 2017; Tornberg, 2022).

To answer the research question: Preventive digital policing in Indonesia is implemented through a combination of high-tech surveillance (OSINT/AI) and "soft"

interventions (Virtual Alerts/Mediation). However, legitimacy shapes its effectiveness by acting as the "social filter" through which these interventions are received.
The novelty of this study lies in three key areas:

a. The Contextual Shift: Most literature on digital policing focuses on predictive policing in the West (crime hotspots). This study shifts the focus to "Preventive Information Governance" in a Global South democracy.
b. The Preventive Paradigm: It documents a move away from the "arrest-first" culture toward a "notify-first" digital culture, which is a significant institutional evolution for the Indonesian National Police.
c. The Role of Social Trust: It empirically links the success of AI-driven policing to the sociological concept of state legitimacy, arguing that "digital guardianship" is impossible without "societal acceptance."

The interaction between technology and state authority in Indonesia serves as a cautionary tale for other developing nations. If digital policing is perceived as an instrument of surveillance rather than a tool for public safety, it may inadvertently increase affective polarization (Tornberg, 2022). As identified in the results, the "algorithmic noise" of platforms like X or TikTok often rewards the very disinformation the police seek to stop. Therefore, a purely state-centric technological approach is insufficient.

The findings suggest that the governance of disinformation must move toward a collaborative governance model. This involves not only the police and technology, but also civil society and independent fact-checkers (like MAFINDO) to bridge the legitimacy gap. The "technology" of governance must include "social transparency" as its most critical feature.

In conclusion, the digital transformation of Bareskrim Polri is a landmark effort in technology-enabled governance. However, the study confirms that the "preventive" potential of digital policing is currently bottlenecked by institutional capacity and, more importantly, a deficit in digital legitimacy. For Bareskrim to effectively govern the information space, it must move beyond being a "Watcher" to being a "Trusted Mediator." The future of digital policing in Indonesia lies not in more powerful OSINT tools, but in the procedural fairness of its digital alerts and the legal clarity of its interventions.

D. CONCLUSION

This study has examined the implementation of preventive digital policing by Indonesia's Criminal Investigation Agency (Bareskrim Polri) as a technology-enabled governance mechanism for addressing disinformation and hate speech. The findings successfully address the research objective by demonstrating that Bareskrim's digital transformation is not merely a technical upgrade, but a strategic shift toward a "preventive guardianship" model.

The research concludes that the governance of digital harms in Indonesia is currently operationalized through a hybrid architecture of high-tech surveillance utilizing OSINT, multi-platform monitoring, and AI-driven classification, and non-

coercive interventions such as Virtual Alerts and online mediation. This approach effectively moves the policing paradigm away from the "arrest-first" culture toward a "notify-first" digital culture, which aims to de-escalate potential social conflicts before they materialize.

However, the findings reveal that the effectiveness of this technology-enabled strategy is profoundly mediated by institutional legitimacy and social trust. While the technical infrastructure for "digital guardianship" is in place, its governance potential is frequently bottlenecked by regulatory ambiguity, specifically within the ITE Law, and a "legitimacy deficit" stemming from public perceptions of state surveillance. Consequently, the study confirms that the successful governance of disinformation in a Global South context depends less on the sophistication of the algorithms and more on the procedural fairness, transparency, and societal acceptance of state-led interventions.

ACKNOWLEDGMENT

I would like to express my sincere appreciation to the University of Indonesia as my home institution for its support. I also extend my gratitude to the Indonesian National Police for their valuable assistance and support throughout this research.

REFERENCES

1. Afisa, A., Qodir, Z., Habibullah, A., & Sugiharto, U. (2024). Analysis of the ITE Law on digital rights and democratic values in Indonesia. *The Journal of Society and Media*, *8*(2), 424–444. https://doi.org/10.26597/jsm.v8i2.16453
2. APJII. (2024). *Survei Penetrasi Internet Indonesia 2024*. Asosiasi Penyelenggara Jasa Internet Indonesia.
3. Bakir, V., & McStay, A. (2018). Fake news and the economy of emotions: Problems, causes, solutions. *Digital Journalism*, *6*(2), 154–175. https://doi.org/10.1080/21670811.2017.1345645
4. Bradford, B., Jackson, J., & Milani, J. (2017). Police legitimacy. In *Oxford research encyclopedia of criminology*. Oxford University Press.
5. Bradshaw, S., & Howard, P. N. (2019). *The global disinformation order: 2019 inventory of organised social media manipulation*. Oxford Internet Institute.
6. Chan, J., & Bennett Moses, L. (2016). Is Big Data challenging criminology? *Theoretical Criminology*, *20*(1), 21–39. https://doi.org/10.1177/1362480615586614
7. Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, *10*(1), 1–10. https://doi.org/10.1038/s41598-020-73510-5
8. Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, *44*(4), 588–608. https://doi.org/10.2307/2094589

9. Crawford, A., & Hutchinson, S. (2016). Mapping the contours of "everyday security": Time, space and emotion. *British Journal of Criminology*, *56*(6), 1184–1202. https://doi.org/10.1093/bjc/azv121
10. Evans, S. K., Pearce, K. E., Vitak, J., & Treem, J. W. (2017). Explicating affordances: A conceptual framework for understanding affordances in communication research. *Journal of Computer-Mediated Communication*, *22*(1), 35–52. https://doi.org/10.1111/jcc4.12180
11. Ferguson, A. G. (2017). *The rise of big data policing: Surveillance, race, and the future of law enforcement*. New York University Press.
12. Fortuna, P., & Nunes, S. (2020). A survey on automatic detection of hate speech in text. *ACM Transactions on Internet Technology*, *20*(2), 1–30. https://doi.org/10.1145/3372695
13. Gibson, J. J. (2015). *The ecological approach to visual perception*. Psychology Press.
14. Goldsmith, A. J. (2010). Policing's new visibility. *British Journal of Criminology*, *50*(5), 914–934. https://doi.org/10.1093/bjc/azq033
15. Goldsmith, A., & Brewer, R. (2015). Digital drift and the criminal interaction order. *Theoretical Criminology*, *19*(1), 112–130. https://doi.org/10.1177/1362480614550117
16. Hill, M., & Hupe, P. (2001). *Implementing public policy: An introduction to the study of operational governance*. SAGE Publications.
17. Holt, T. J., Bossler, A. M., & Seigfried-Spellar, K. C. (2015). *Cybercrime and digital forensics: An introduction*. Taylor & Francis.
18. Law of the Republic of Indonesia Number 11 of 2008 concerning Electronic Information and Transactions (*Undang-Undang Republik Indonesia Nomor 11 Tahun 2008 tentang Informasi dan Transaksi Elektronik*).
19. Leukfeldt, E. R., & Yar, M. (2016). Applying routine activity theory to cybercrime: A theoretical and empirical analysis. *Deviant Behavior*, *37*(3), 263–280. https://doi.org/10.1080/01639625.2015.1012409
20. MAFINDO. (2024). *Lanskap hoaks 2024: Pemetaan disinformasi di Indonesia*. Masyarakat Anti Fitnah Indonesia.
21. McGuire, M. R., & Holt, T. J. (2017). *The Routledge handbook of technology, crime and justice*. Routledge.
22. Montasari, R., Carpenter, V., & Masys, A. J. (2023). *Digital transformation in policing: The promise, perils and solutions*. Springer Nature Switzerland.
23. Newman, G., Clarke, R. V., & Shoham, S. G. (2021). *Rational choice and situational crime prevention*. Routledge.
24. Rahmadhany, A., Safitri, A. A., & Irwansyah. (2021). Fenomena penyebaran hoax dan hate speech pada media sosial. *Jurnal Teknologi dan Informasi Bisnis*, *3*(1), 1–200. https://doi.org/10.47233/jteksis.v3i1.181
25. Stewart, A. J., McCarty, N., & Bryson, J. J. (2020). Polarization under rising inequality and economic decline. *Science Advances*, *6*(50), Article abd4201. https://doi.org/10.1126/sciadv.abd4201
26. Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media*. Princeton University Press.

27. Tornberg, P. (2022). How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences (PNAS)*, *119*(42), e2207159119. https://doi.org/10.1073/pnas.2207159119
28. Tyler, T. R. (2006). *Why people obey the law*. Princeton University Press.
29. Tyler, T. R., & Huo, Y. J. (2002). *Trust in the law: Encouraging public cooperation with the police and courts*. Russell Sage Foundation.
30. Van Meter, D. S., & Van Horn, C. E. (1975). The policy implementation process: A conceptual framework. *Administration & Society*, *6*(4), 445–488. https://doi.org/10.1177/009539977500600404
31. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559
32. Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. *Proceedings of the 1st Workshop on Abusive Language Online*, 78–84. https://doi.org/10.18653/v1/W17-3012
33. Yin, R. K. (2017). *Case Study Research and Applications: Design and Methods*. SAGE Publications.