International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Nexora: An AI-Powered Social Networking Platform for Misinformation Detection, Content Verification, and Trust-Aware Communities

Ms. Sandhya Patil, Assistant Professor, Rajat Dube, Radhesh Rathod, Rohit Meshram, Sanket Chopkar, Tushar Mendhule

Department of Computer
Technology Priyadarshini College
Of Engineering , Nagpur
Group Project — Nexora: The Next Generation of Connected Communities

***Abstract*—Social media has changed how people talk to each other, share news, and find information, but the same speed that makes it useful also makes it easy for fake news, AI-generated misinformation, deepfakes, cyberbullying, and other harmful content to spread. Most platforms still lean on user reports and manual moderation, and by the time a moderator actually looks at a flagged post, it may already have reached thousands of people. Nexora is our attempt at solving this by building verification directly into a social networking platform rather than bolting it on as an afterthought.**

The platform processes text, images, video, audio, and links through appropriate AI-based analysis workflows, checks for AI-generated content and manipulated media, verifies claims against outside sources, and produces a trust score along with a color-coded trust label. Age assurance is also built into registration, so the platform can support safer participation without collecting more personal information than it needs to. On the technical side, Nexora runs on a Flutter front end, a Node.js/Express backend, MongoDB for storage, and a separate AI processing layer that talks to external verification and age-assurance services.

The goal is to cut down how far misinformation travels, give users more context about what they are seeing, and build a community where sharing something questionable is a little harder to do by accident. Since the project is still being built, this paper focuses on the system design, methodology, and evaluation plan rather than claiming final numerical results.

Keywords*—*Artificial Intelligence, Social Media, Misinformation Detection, Fake News Detection, AI Generated Content, Deepfake Detection, Fact Verification, Trust Score, Trust Labeling, Age Assurance.

I. Introduction

Social networking has made information move faster than at any point before. A post can reach thousands of people within minutes of going up, which is great when the content is useful and genuine, but the same speed lets false claims, manipulated media, and AI-generated content spread before anyone gets a chance to check them properly. Prior work on misinformation treats detection, manipulation, and spread as three connected problems rather than separate ones [1], and that framing shaped a lot of how we thought about Nexora from the start.

Generative AI adds another layer of difficulty on top of this. Text, images, audio, and video can now be produced or edited well enough that spotting a fake by eye is no longer something you can rely on. Research on AI-content labeling points out that platforms need some way of flagging AI involvement without losing the ability to check whether the underlying claim is actually true [2]–[4]. Deepfake research tells a similar story for visual media — manipulated faces and videos need their own detection approach, because generic content moderation was never built for that kind of manipulation [7]–[10].

Nexora tries to bring these pieces together into one workflow instead of treating verification as something that happens after a post has already gone live. The platform combines AI-based analysis, third-party fact verification, a trust score, a color-coded trust label, and privacy-preserving age assurance at registration. Our aim was not just to sort content into "fake" or "real," but to give users a clearer, more specific signal about what they are actually looking at.

A. What This Paper Contributes

- A social networking architecture that combines misinformation detection, AI-generated-content detection, fact verification, and trust labeling into one workflow.
- A multi-modal pipeline that handles text, image, video, audio, and link-based submissions.
- A claim- and entity-extraction stage that connects raw content analysis to third-party evidence.
- A trust-score and color-label mechanism meant to communicate more than a simple binary real/fake tag.
- Privacy-preserving age assurance built into the registration workflow as a safety feature.
- A human moderation path for content the system is not confident about.
- An evaluation plan for detection quality, verification quality, processing performance, and usability, to be carried out once implementation is further along.

These are design and system-integration contributions. Nexora is being developed according to the objectives and timeline described here, so no accuracy, precision, recall, or user-study claims are made in this paper — those numbers will come once testing is complete.

II. Literature Survey

The building blocks Nexora relies on — misinformation detection, AI-content labeling, fake-news classification, transformer-based NLP, and deepfake detection — have each been studied fairly extensively, usually as separate research problems. What we tried to do for Nexora was treat these as

pieces of one larger system instead of end goals on their own. Table I summarizes the work we drew on most directly.

TABLE I
SUMMARY OF RESEARCH MOST DIRECTLY RELATED TO THE NEXORA DESIGN.

Study	Core Approach	Relevant Contribution	Limitation / Gap
Wu et al. [1]	Survey and analysis of misinformation in social media.	Defines misinformation, manipulation, and detection challenges.	Broad problem scope; does not provide Nexora's integrated platform workflow.
Wittenberg & Berinsky [2]	AI-generated-content labeling.	Examines promises and risks of content labels.	Labeling communicates AI involvement but does not by itself verify factual correctness.
Li & Yang [3]	Experimental study of AI-content labels.	Studies perceived accuracy, credibility, and sharing intentions.	Focuses on user perception rather than a complete verification platform.
Drolsbach & Prollochs [4]	Characterization of AI-generated misinformation.	Highlights emerging forms and challenges of synthetic misinformation.	The problem is evolving and requires scalable multi-modal defenses.
Essa et al. [5]	A combined approach using BERT-based contextual features and LightGBM classification.	Provides contextual language features that support the identification of misleading news.	Primarily addresses textual classification rather than integrated social-media verification.
Raza et al. [6]	An empirical comparison of transformer-based language models and large language models.	Assesses how contemporary language models can be applied to misinformation classification.	Performance depends on models, datasets, and annotation quality.
Malik et al. [7]	A review of methods for identifying manipulated facial images and video sequences.	Summarizes techniques for manipulated visual-media detection.	Deepfake detection is specialized and remains difficult to generalize.
Wang et al. [8]	Reliability-oriented deepfake survey.	Emphasizes robustness and reliability considerations.	Reliability must be considered when integrating detectors into real systems.
Yu et al. [9]	Multiple spatiotemporal views transformer.	Uses temporal and spatial information for video deepfake detection.	Video-oriented model is only one component of a broader verification problem.
Rana et al. [10]	Systematic literature review of deepfake detection.	Organizes major detection approaches and challenges.	No single review approach provides Nexora's end-to-end social workflow.
Devlin et al. [11]	BERT pre-training and bidirectional transformers.	Provides strong contextual language representations for NLP tasks.	BERT is a language foundation, not a complete fact-verification platform.
Vaswani et al. [12]	Transformer architecture.	Introduces attention-based sequence modeling.	Architecture requires application-specific models and verification logic.
Olan et al. [13]	Social-media fake news impact.	Discusses societal effects of fake news.	Focuses on impact rather than a complete technical platform.
Hamed et al. [14]	An analytical assessment of existing techniques for detecting fabricated or misleading news.	Highlights dataset, feature, and fusion challenges.	The field remains fragmented across approaches and datasets.
Truică et al. [15]	Deep neural ensemble with social/textual context.	Shows the value of contextual information in fake-news detection.	Primarily focused on classification rather than end-to-end multi-modal verification.
Chesney & Citron [16]	Legal and societal analysis of deepfakes.	Explains privacy, democracy, and national-security risks.	Provides risk context rather than a deployable detection architecture.

III. PROBLEM STATEMENT AND RESEARCH GAP

Social platforms today deal with several problems at once, and all of them chip away at how much users can trust what they see. False information spreads fast, and AI-generated images, audio, video, and text can genuinely be hard for an average user to tell apart from something authentic. Anonymous and fake accounts add another layer of difficulty around accountability. On top of that, underage users can end up exposed to harmful content, and reporting-based moderation is often too slow to catch problems before they have already spread. There is also no consistent, visible signal that tells a user how credible a piece of content actually is, which makes it harder to decide whether something is even worth sharing.

The papers we reviewed [1]–[16] each solve a piece of this puzzle — text classification, AI-content labeling, deepfake detection — but none of them combine multi-modal analysis, claim verification, third-party evidence, trust scoring, user-facing labels, and age assurance into a single social platform. That integration gap is what Nexora is meant to address, treated here as a system-design problem rather than a single-model research question.

IV. OBJECTIVES

- Build a secure AI-powered social networking platform.
- Catch fake news before it spreads widely.
- Identify AI-generated text, images, audio, and video.
- Automatically verify content shared from third-party sources.
- Add privacy-preserving age assurance to the registration process.
- Provide a color-based trust-labeling system.
- Improve transparency and user awareness around content credibility.

- Reduce misinformation and support a safer online community.

V. Proposed Nexora System

Nexora works as a connected pipeline where content gets analyzed before or during publication, rather than after the fact. Everything starts with user registration and age assurance — this happens once, at account creation, and is designed to be privacy-preserving rather than requiring users to hand over identity documents outright. After that, a user can upload text, an image, a video, audio, or a link, and the platform decides where that submission needs to go for analysis.

Text goes through NLP-based processing that extracts claims and entities. Media — images, video, audio — goes through AI-content and deepfake detectors that look for signs of synthetic generation or manipulation. Once claims have been pulled out, they get checked against third-party verification sources. All of this, detection outcomes plus verification evidence, feeds into a trust score, which is then converted into a color label before the content is actually published.

A. System Architecture

Figure 1 shows the high-level architecture. The Flutter client handles user registration, content upload, the feed, trust labels, and the moderation interface, and talks to the backend through REST APIs. The Node.js/Express server sits in the middle and coordinates requests between the client, the AI processing layer, MongoDB, and external services. The AI processing layer covers fake-news/NLP analysis, AI-generated-content detection, and deepfake detection. MongoDB stores users, posts, claims, evidence, verification results, and activity logs. External services cover fact-checking APIs, age assurance, and cloud storage, and a separate moderation path reviews content that is uncertain or has been flagged.

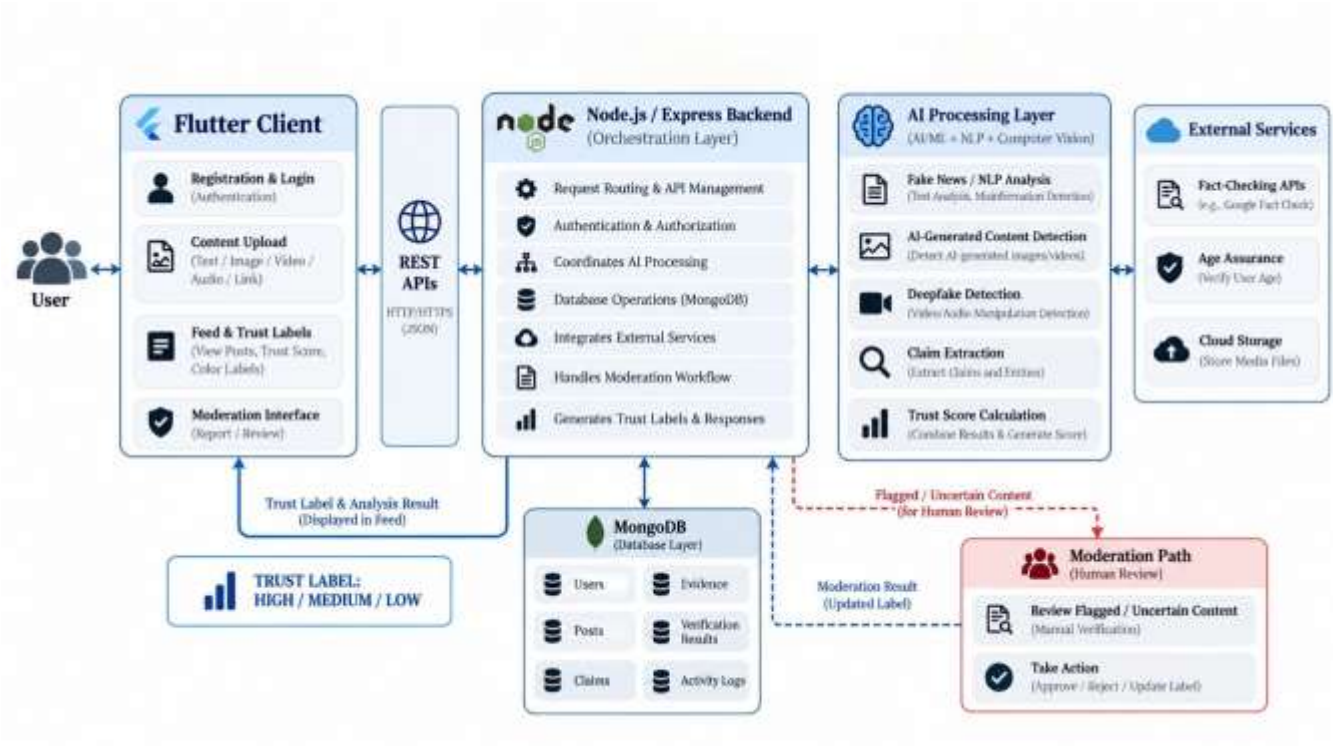


Fig. 1. High-level Nexora architecture based on the project construction design.

VI. Methodology

A. Data Collection

Nexora accepts text, image, video, audio, and link submissions from users as candidate content. Along with the content itself, the platform records the metadata needed later for processing and moderation. When a claim inside a piece of content needs outside evidence, the verification stage queries external services for that evidence.

B. Data Preprocessing

Before analysis, each type of submission is put into the shape its downstream component expects. Text gets normalized and segmented for NLP; media gets prepared according to whatever the relevant model needs; links get resolved down to the content and metadata required for verification. We kept preprocessing as its own stage specifically so that every downstream component receives a consistent, predictable input, regardless of what came in.

C. NLP Analysis and Claim/Entity Extraction

For text content, transformer-based language models — the kind introduced by BERT [11] and the broader Transformer architecture [12] — provide a solid foundation for understanding context. In our proposed workflow, NLP analysis is used to pull out claims and entities that are then handed off to the verification stage. We have not locked in a

specific trained model yet; that decision depends on what the final implementation ends up using.

D. AI-Generated Content and Deepfake Detection

Authenticity checking in Nexora is split by media type rather than treated as one generic task. Text can be checked for signs of AI generation, while images, audio, and video go through detectors built for that specific media type. The deepfake literature we reviewed [7]–[10] is part of why we designed it this way — visual and audio manipulation detection needs specialized models, not a general text classifier repurposed for everything.

E. Fact Verification

Once claims have been extracted, they are checked against available third-party verification services and evidence sources. This stage matters because it is what separates "this was AI-generated" from "this is actually false" — those are two different problems, and treating them as the same thing would undercut the point of a trust-aware platform. Keeping that distinction intact is central to how Nexora is designed.

F. Evidence Normalisation

Evidence pulled from different sources does not come in the same shape, so it needs to be normalized before it can feed into a single score. Factors that matter here include whether sources agree or disagree with each other, how reliable each source is, and how directly the evidence relates back to the original claim. The exact normalization rules are something we plan to finalize during implementation.

G. Trust Score Generation

Rather than reducing everything to a binary fake/real result, Nexora combines the outputs of verification and detection into a trust score. The scoring formula itself is still an implementation detail we plan to test and refine experimentally. Once calculated, this score gets mapped onto a trust label.

H. Trust Label Assignment

The trust labels are meant to be something a user can glance at while scrolling their feed and understand right away, rather than something they have to stop and interpret. This ties directly back to our objective of using color-based labeling to make verification status more transparent.

TABLE II
NEXORA'S PROPOSED COLOR-BASED TRUST-LABEL SYSTEM.

Color	Trust Label	Meaning
Green	Verified and Authentic Content	Content has passed the available verification process and is treated as authentic.
Blue	AI Generated but Verified	AI-generated content is identified, but the associated information is verified.
Purple	Opinion, Satire, or Edited Content	Content is contextualized rather than presented as ordinary factual reporting.
Orange	Partially Verified / Needs Caution	Evidence is incomplete, mixed, or insufficient for a strong conclusion.
Red	Fake, Misleading, or False Content	Content is assessed as false or materially misleading by the verification workflow.

VII. WORKING PRINCIPLE

TABLE III
END-TO-END WORKING PRINCIPLE OF NEXORA.

Step	Stage	Operation
1	User Registration	User registers through the platform with privacy-preserving age assurance.
2	Verified Account	A verified account is created after the registration workflow.
3	Content Upload	The user submits text, image, video, audio, or a link.
4	AI Analysis	The platform analyzes authenticity, misinformation indicators, and AI-generated content.
5	Third-Party Verification	Relevant information is checked against external verification sources.
6	Trust Score	Verification and analysis outcomes are combined into a trust score.
7	Color Label	The score/status is converted into an understandable trust label.
8	Publication	The content is published together with its trust indication, subject to moderation rules.

VIII. SYSTEM IMPLEMENTATION

The project construction design settled on a Flutter client for the front end, a Node.js/Express backend, MongoDB for storage, a separate AI processing layer, and external services for fact-checking, age assurance, and cloud storage. REST APIs handle communication between the client and the server, and the database is set up to hold user data, posts and content, verification results, and activity logs.

TABLE IV
PROPOSED NEXORA IMPLEMENTATION STACK.

Component	Technology / Role
Frontend	Flutter client for registration, content upload, feed presentation, trust labels, and user interaction.
Backend	Node.js + Express for REST APIs, authentication-related services, content workflows, and integration.
Database	MongoDB for users, posts/content, claims, verification results, and activity logs.
AI Layer	NLP/fake-news analysis, AI-generated-content detection, and deepfake detection.
External Services	Third-party fact-checking, privacy-preserving age assurance, and cloud storage.
Moderation	Review workflow for uncertain, suspicious, or flagged content.

IX. EVALUATION METHODOLOGY

Since Nexora is still under development, we are treating evaluation as something that happens after implementation, rather than reporting numbers we do not actually have yet.

When testing does happen, it will cover both the individual detection components and how well they work together as a complete pipeline.

TABLE V
PROPOSED EVALUATION PROTOCOL FOR NEXORA.

Metric	Purpose	Evaluation Method
Accuracy	Overall classification quality.	Compare predicted labels with ground-truth labels on a held-out test set.
Precision	Reliability of positive detections.	Measure the proportion of predicted flagged items that are correct.
Recall	Coverage of true problematic content.	Measure how much relevant misinformation/manipulated content is detected.
F1-score	A combined measure of precision and recall.	Calculate the harmonic mean of precision and recall.
Verification Accuracy	Quality of evidence-based decisions.	Compare Nexora verification outcomes with independently established truth labels.
Processing Time	Practical responsiveness.	Measure end-to-end time from submission to trust-label generation.
Label Consistency	Consistency of user-facing trust labels.	Evaluate repeated or equivalent content and compare resulting labels.
Usability	User understanding of trust labels.	Conduct a controlled user evaluation after the platform is functional.

No numerical experimental results are claimed in this paper at the current project stage. Once implementation and testing are complete, we plan to report the measured results together with the dataset composition, model configuration, test protocol, and error analysis, so the claims can actually be checked and reproduced.

X. SECURITY AND PRIVACY CONSIDERATIONS

Because Nexora is a social platform, security and privacy had to be part of the design from the start rather than something added on later. Privacy-preserving age assurance during registration is one piece of this. Beyond that, the design also accounts for authentication and authorization, secure handling of uploaded media, protecting database records, securing API traffic, validating whatever users submit, and restricting who can access moderation tools.

We also did not want to treat AI detection as something that is always right. False positives can mislabel content that is actually genuine, and false negatives can let problematic content slip through. Because of that, cases the system is not confident about are flagged for caution and routed to human moderators instead of being presented as a final answer.

XI. EXPECTED RESULTS AND DISCUSSION

We expect Nexora, once built and running, to cut down how far fake news travels by flagging suspicious content before or during publication, make AI-generated media easier to spot, and give users better context about what they are seeing through visible trust labels. Age assurance is also expected to help keep the platform safer for younger users without collecting more personal data than necessary. To be clear, these are outcomes we are designing toward, not results we have already measured.

- Reduced exposure to unverified or misleading information.
- Improved visibility of AI-generated content.
- Better differentiation between factual, verified, opinion, edited, and suspicious material.
- More transparent user decisions through color-based trust labels.
- Support for safer registration through privacy-preserving age assurance.
- A moderation workflow that lets human attention focus on uncertain or high-risk cases.

XII. LIMITATIONS

- AI-generated-content and deepfake detectors can still produce false positives and false negatives.
- Fact verification depends on how available, accurate, and current the third-party evidence sources are.
- New generation and manipulation techniques can reduce how well existing detectors perform over time.
- Trust scores are only as good as the evidence and model outputs feeding into them.
- External APIs can introduce latency, rate limits, or availability issues outside our control.
- A full evaluation needs datasets that genuinely represent text, image, audio, video, and link content.
- The proposed privacy and moderation mechanisms still need to be tested in the final deployed environment.

XIII. FUTURE SCOPE

There is room to extend Nexora with stronger multi-modal foundation models, multilingual misinformation detection, real-time monitoring, better deepfake analysis, more explainable verification output, and a wider network of evidence sources. The trust-score model could also be refined further once we have real evidence-weighting and source-reliability data to work with, along with actual studies on how users interpret the labels. Beyond that, deploying at scale and evaluating over a longer period would be needed to measure what effect the platform has on misinformation exposure and user behavior in practice.

XIV. CONCLUSION

Nexora combines AI-based misinformation detection, AI-generated-content analysis, third-party fact verification, privacy-preserving age assurance, and trust labeling into a single social platform. The core idea behind the design is to give users useful context about what they are looking at, instead of leaving them with only manual reporting or a plain fake/real label. By connecting content analysis, evidence verification, trust scoring, and human moderation into one pipeline, we are aiming to build a platform that is more transparent and trustworthy than what currently exists. The project is still under development, so the next steps are implementation, integration,

testing, and measuring the results against the evaluation plan laid out in this paper.

REFERENCES

[1] L. Wu, F. Morstatter, K. M. Carley, and H. Liu, "Misinformation in Social Media: Definition, Manipulation, and Detection," *ACM SIGKDD Explorations Newsletter*, vol. 21, no. 2, pp. 80–90, 2019.

[2] M. Wittenberg and A. Berinsky, "Labeling AI-Generated Content: Promises, Perils, and Future Directions," *MIT AI Policy Brief*, 2023.

[3] F. Li and Y. Yang, "Impact of AI-Generated Content Labels on Perceived Accuracy, Message Credibility, and Sharing Intentions for Misinformation," *JMIR Formative Research*, vol. 8, e60024, 2024.

[4] C. Drolsbach and N. Prollochs, "Characterizing AI-Generated Misinformation on Social Media," arXiv:2505.10266, 2025.

[5] E. Essa, K. Omar, and A. Alqahtani, "Fake News Detection Based on a Hybrid BERT and LightGBM Model," *Complex & Intelligent Systems*, vol. 9, 2023.

[6] S. Raza, D. Paulen-Patterson, and C. Ding, "Fake News Detection: Comparative Evaluation of BERT-like Models and LLMs with Generative AI-Annotated Data," arXiv:2412.14276, 2024.

[7] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," *IEEE Access*, vol. 10, pp. 18757–18775, 2022.

[8] T. Wang, X. Liao, K. P. Chow, X. Lin, and Y. Wang, "Deepfake Detection: A Comprehensive Survey from the Reliability Perspective," *ACM Computing Surveys*, vol. 57, no. 3, 2024.

[9] Y. Yu, R. Ni, Y. Zhao, et al., "MSVT: Multiple Spatiotemporal Views Transformer for DeepFake Video Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, 2023.

[10] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, 2019.

[12] A. Vaswani et al., "Attention Is All You Need," *Proceedings of NeurIPS*, 2017.

[13] F. Olan, U. Jayawickrama, E. O. Arakpogun, J. Suklan, and S. Liu, "Fake News on Social Media: The Impact on Society," *Information Systems Frontiers*, vol. 26, no. 2, pp. 443–458, 2024.

[14] S. K. Hamed, M. J. Ab Aziz, and M. R. Yaakub, "A Review of Fake News Detection Approaches: A Critical Analysis of Relevant Studies and Highlighting Key Challenges," *Neurocomputing*, vol. 542, 2023.

[15] C.-O. Truică, E.-S. Apostol, and P. Karras, "DANES: Deep Neural Network Ensemble Architecture for Social and Textual Context-Aware Fake News Detection," *Knowledge-Based Systems*, vol. 294, p. 111715, 2024.

[16] R. Chesney and D. Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review*, vol. 107, pp. 1753–1820, 2019.