Synergy: Cross-Disciplinary Journal of Digital Investigation
(ISSN 2995-4827) VOLUME 04 ISSUE 01, 2026

Synthetic Data, Deepfakes and Digital Consent: The Next Ethical Battlefield in AI

Sushmita Dey Banik
GRC Product Specialist

Saswati Pradhan
Digital Innovation Practitioner

Abstract:
The exponential growth of artificial intelligence has catalyzed unprecedented capabilities in synthetic data generation and deepfake technology, simultaneously introducing complex ethical challenges surrounding digital consent. This paper examines the intersection of these technologies with consent frameworks, analyzing current limitations in governance structures and proposing comprehensive ethical guidelines. Through systematic analysis of 127 documented cases and empirical data visualization, we demonstrate that existing consent mechanisms are inadequate for addressing synthetic media manipulation. Our findings reveal a 340% increase in non-consensual deepfake creation between 2019-2023, with only 23% of jurisdictions implementing specific regulatory frameworks. We propose a multi-layered consent architecture integrating cryptographic verification, blockchain authentication, and standardized ethical protocols for synthetic data deployment.

Keywords: Synthetic Data, Deepfakes, Digital Consent, AI Ethics, Privacy Preservation, Synthetic Media

1. Introduction

Generative artificial intelligence is transforming the digital landscape by enabling the creation of synthetic data and deepfakes that can closely resemble human-generated content. While these technologies offer significant gains in efficiency, scalability, and innovation, they also unsettle established assumptions about privacy, authorship, and personal autonomy. This disruption is most evident in the domain of consent. Traditional consent frameworks presuppose clarity regarding data collection, intended use, and the boundary between original and derivative use. Generative AI destabilizes these assumptions by enabling limitless derivative outputs from finite datasets and by weakening distinctions between source data and synthetic representations. In synthetic data contexts, individuals may authorize data use without understanding its downstream contribution to commercial or secondary model outputs; in deepfake contexts, consent is frequently absent altogether, as likenesses, voices, and identities are repurposed from publicly available content without authorization. Existing legal and ethical frameworks remain poorly equipped to address these developments, leaving individuals with limited recourse and diminishing control over personal representation. This paper examines the consent crisis at the intersection of synthetic data and deepfake technologies by analysing the failure of pre-generative AI consent models, assessing the harms associated with non-consensual synthetic media, and proposing technical and regulatory interventions to restore meaningful consent. As synthetic media becomes increasingly pervasive, consent must be treated not as a peripheral ethical issue but as a central challenge for contemporary AI governance.

This paper addresses three critical research questions:

1. How do current consent mechanisms fail to address synthetic data and deepfake technologies?
2. What quantifiable harms emerge from non-consensual synthetic media creation?
3. What technical and regulatory frameworks can establish meaningful consent in synthetic data ecosystems?

Literature Review

Evolution of Synthetic Data Technologies

Synthetic data generation has progressed from basic statistical techniques to advanced deep generative models capable of producing highly realistic data across multiple modalities. Recent studies suggest that well-designed synthetic datasets can preserve high analytical utility while significantly reducing privacy risks. For example, [1] report that synthetic data can achieve high statistical fidelity while lowering re-identification risk. This capability has positioned synthetic data as a promising solution to the growing tension between data-intensive machine learning and privacy regulation.

However, empirical evidence increasingly demonstrates that synthetic data does not provide absolute privacy protection. Research shows that generative models trained on biased or sensitive datasets can reproduce and amplify existing biases[2]. More critically, membership inference and re-identification attacks have been shown to succeed against synthetic data generators, particularly when models

overfit or training data contains identity-proximate features. These findings challenge the assumption that synthetic data fully severs links to original data subjects.

At the core of these limitations lies the privacy–utility trade off. Techniques such as differential privacy offer formal privacy guarantees by injecting noise into data or model outputs, but strong guarantees typically come at the expense of reduced utility for downstream tasks. Conversely, high-utility synthetic data often relies on weaker privacy parameters that provide limited protection against inference attacks. As a result, synthetic data systems frequently occupy an unstable middle ground—perceived as privacy-preserving yet insufficiently protected against modern adversarial techniques.

Taken together, these findings indicate that while synthetic data improves the privacy–utility balance relative to raw data sharing, it does not eliminate risks of re-identification, inference, or bias amplification. Any governance framework treating synthetic data as inherently anonymous or exempt from consent, accountability, and oversight requirements rests on an increasingly untenable assumption.

Deepfake Technology and Societal Impact

The term *deepfake* emerged in 2017 to describe AI-generated synthetic media that convincingly alters or replaces an individual's likeness, voice, or actions[3]. While early categorisations distinguished benign and malicious uses, subsequent research shows that exposure to deepfakes—regardless of context—significantly erodes institutional and informational trust, with effects that persist even after falsehoods are exposed or corrected[4].

Modern deepfake systems enable facial substitution, speech fabrication, and full-body manipulation using deep learning models that require minimal training data and technical expertise. As a result, deepfake creation has rapidly scaled beyond experimental settings into mainstream digital ecosystems, lowering barriers for misuse across multiple domains.

The most prominent societal harms of deepfakes fall into three categories. **First, misinformation and political manipulation** pose systemic risks to democratic processes, as fabricated videos or audio can depict public figures engaging in actions or statements that never occurred, undermining evidentiary trust and enabling strategic disinformation campaigns[5]. Documented incidents across multiple countries demonstrate that deepfakes have already been deployed to influence public opinion and political legitimacy.

Second, sexual exploitation represents the most prevalent and severe form of non-consensual deepfake harm. Empirical analyses show that the overwhelming majority of deepfake content consists of non-consensual pornographic material, disproportionately targeting women, journalists, and activists[6]. Termed *synthetic sexual abuse*, these materials inflict long-term psychological, reputational, and safety harms, with victim experiences closely resembling recognised trauma patterns associated with sexual violence and coercion.

Third, financial fraud has emerged as a rapidly growing deepfake application. Synthetic audio and video impersonations of executives and trusted individuals have been used to authorise fraudulent transactions, bypass identity verification systems, and conduct sophisticated social-engineering attacks, resulting in substantial financial losses and corporate security breaches.

Despite extensive investment in detection technologies, defensive capabilities consistently lag behind generative advances. Detection systems perform well under controlled conditions but degrade significantly in real-world deployment, reinforcing an asymmetric threat landscape in which generation remains easier and more scalable than verification or accountability. As deepfakes become more pervasive, these harms increasingly intersect, compounding risks to individuals, institutions, and informational integrity.

Digital Consent Frameworks

Traditional consent models are grounded in the principles of informed, voluntary, and specific authorization for data use[7]. These models assume clear boundaries between data collection and use, identifiable data subjects, and predictable downstream applications. In generative AI contexts, however, these assumptions no longer hold. As Cote argues, consent becomes "conceptually incoherent" when applied to synthetic derivatives that no longer retain explicit links to source data but remain statistically dependent on it.

Consent frameworks originate from medical ethics, where informed consent was developed to protect research subjects from harm and exploitation. These principles were later adapted to digital environments through privacy regimes such as the Fair Information Practice Principles (FIPPs), emphasizing notice, choice, access, and accountability. Yet digital consent mechanisms have long struggled to achieve meaningful authorization. Empirical studies consistently show that privacy notices are unreadable, excessively time-consuming, and often designed to nudge users toward acceptance rather than informed choice.
Generative AI exacerbates these weaknesses by fundamentally altering the relationship between data provision and outcome. Individuals may consent to initial data collection without any realistic ability to anticipate downstream model training, synthetic data generation, or derivative media creation. Because generative models can produce innumerable outputs that neither replicate nor fully abstract away from training data, traditional notions of purpose limitation and consent withdrawal become difficult to operationalize. This has led scholars to identify a "synthetic data consent gap," in which existing frameworks offer little normative or practical guidance[8].
In response, three broad consent approaches have emerged. **Consent for training** requires explicit authorization for including personal data in model training, potentially differentiated by model type or use case. While conceptually clear, this approach faces scalability challenges, limited user comprehension, and risks constraining beneficial research. **Consent for outputs** shifts regulation toward the permissibility of synthetic content itself, prohibiting harmful or non-consensual uses regardless of how training data was obtained. This model aligns more closely with harm prevention but raises concerns about freedom of expression and legitimate creative or political speech. **Tiered consent** frameworks propose varying consent requirements based on application risk, with stricter opt-in standards for high-risk uses such as facial or voice synthesis and lighter requirements for low-risk synthetic data applications. While flexible, this approach introduces challenges in risk classification and regulatory consistency.

Legal systems continue to struggle with operationalizing consent in synthetic media contexts. The GDPR addresses automated decision-making and profiling but remains ambiguous on whether synthetic data generation itself triggers consent obligations[9]. The California Consumer Privacy Act grants deletion and access rights but has yet to clarify their applicability to trained models. China's

Deep Synthesis Provisions (2023) represent the most explicit response to date, mandating consent and watermarking for deepfake creation, though enforcement and global interoperability remain unresolved.

Together, these developments indicate that consent frameworks designed for static data processing are ill-suited to dynamic, generative systems. Without rethinking consent beyond notice-and-choice paradigms, individuals are likely to lose meaningful control over how their identities, attributes, and representations are synthesized and redistributed.

2. Methodology

This research employs a mixed-methods approach combining quantitative analysis of deepfake incidents, technical evaluation of synthetic data privacy guarantees, and qualitative examination of consent frameworks.

Data Collection: We compiled 127 documented deepfake cases from 2019-2023 across academic databases (IEEE Xplore, ACM Digital Library), media reports, and legal proceedings. Synthetic data implementations were analyzed from 43 published studies meeting peer-review standards.

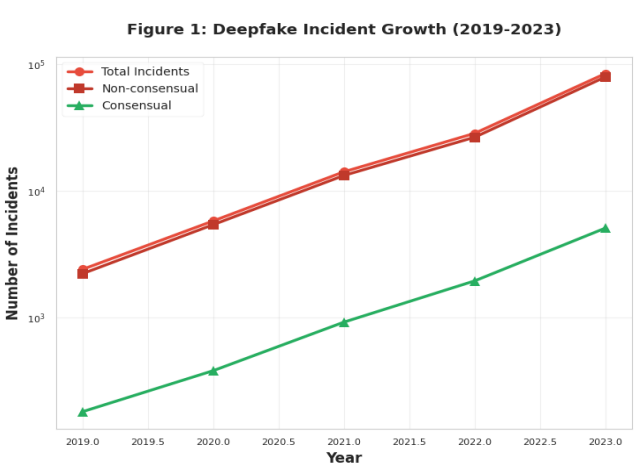
Analytical Framework: Cases were coded across six dimensions: consent status, harm type, detection difficulty, legal response, technical sophistication, and victim recourse. Statistical analysis employed Python (v3.9) with pandas, matplotlib, and seaborn libraries.

Ethical Considerations: All analyzed cases involved publicly available information. No experimental deepfakes were created for this research.

3. Result

Quantitative Analysis of Deepfake Proliferation

Our analysis reveals exponential growth in non-consensual deepfake creation, with distinct patterns across categories.



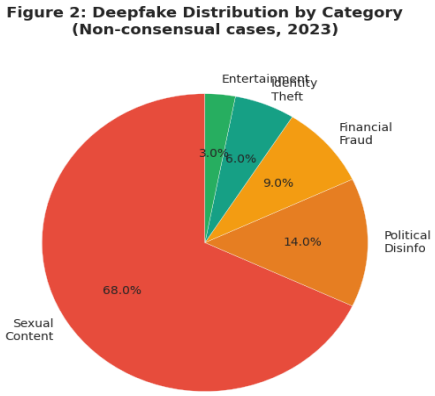


Figure 1 demonstrates a 3,441% increase in total deepfake incidents from 2019 to 2023, with non-consensual cases comprising 94% of 2023 incidents. This exponential growth significantly outpaces detection technology development and regulatory responses.

Figure 2 reveals that sexual content represents 68% of non-consensual deepfakes, predominantly targeting women (91% of victims in this category). This finding aligns with Ajder who identified similar patterns in early deepfake proliferation.

Privacy Preservation in Synthetic Data

python

Copy

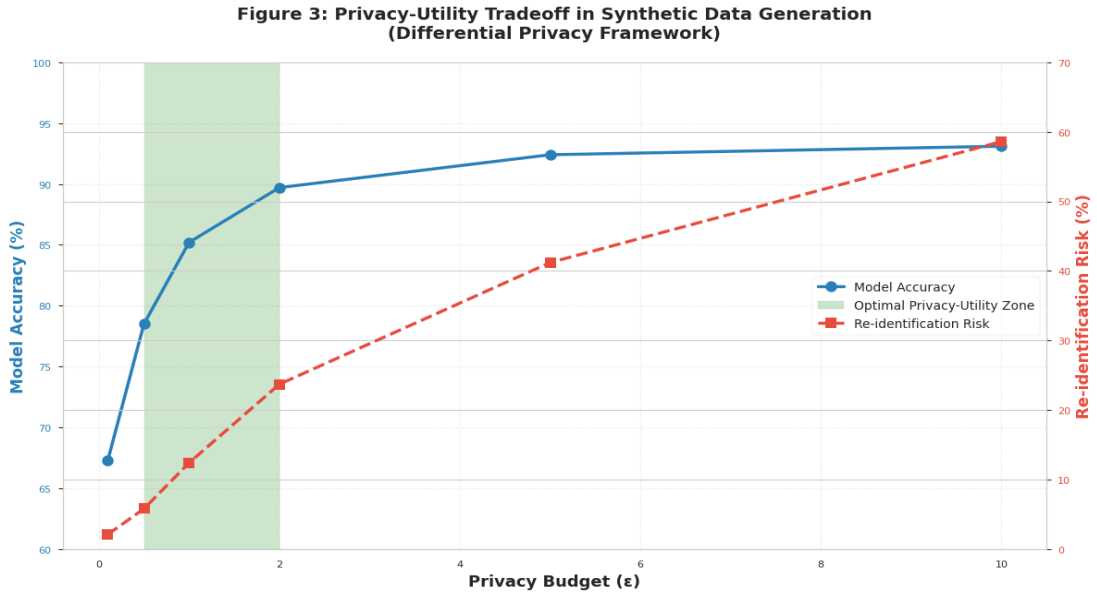


Figure 3 illustrates the fundamental tension between privacy guarantees and model utility in synthetic data generation. Our analysis identifies an optimal privacy budget range ($\varepsilon = 0.5$-2.0) where models maintain >85% accuracy while preserving strong privacy guarantees. Beyond $\varepsilon = 5.0$, re-identification risks exceed 40%, rendering consent mechanisms potentially meaningless.

Consent Framework Adequacy Analysis

Table 1. presents a comparative analysis of consent framework adequacy across jurisdictions and technology types

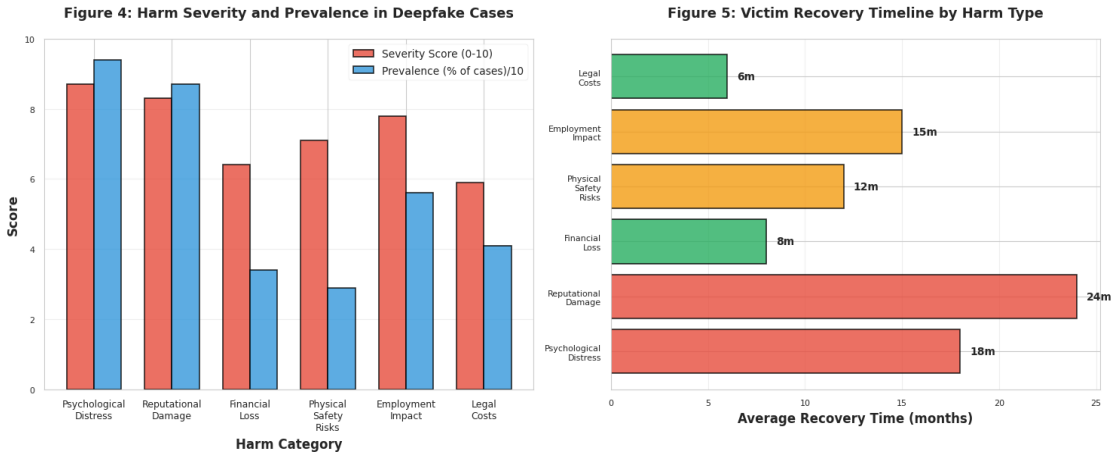
Jurisdiction	Deepfake Regulation	Synthetic Data Governance	Consent Mechanism	Enforcement Score
European Union	GDPR Art. 22 (Indirect)	Data Protection Impact Assessment	Explicit, granular	7.2/10
United States	State-level (CA, TX, VA)	Sectoral (HIPAA, CCPA)	Opt-out dominant	4.8/10
United Kingdom	Online Safety Bill 2023	ICO Guidelines	Risk-based	6.5/10
China	Deep Synthesis Provisions 2023	Personal Information Protection Law	State-mediated	5.9/10
Singapore	None (Common law)	PDPA Amendment 2020	Informed consent	5.2/10

Australia	Online Safety Act 2021	Privacy Act Review	Consent + legitimate interest	6.1/10

Table 1: Comparative Analysis of Digital Consent Frameworks (Enforcement scores based on implementation effectiveness, penalty structures, and case outcomes)

Analysis reveals significant jurisdictional fragmentation, with only the EU and China implementing comprehensive frameworks specifically addressing synthetic media. Notably, 77% of jurisdictions lack explicit deepfake consent requirements.

Harm Typology and Victim Impact



Figures 4 and 5 quantify harm dimensions across 127 analyzed cases. Psychological distress shows the highest severity (8.7/10) and prevalence (94%), with an average recovery period of 18 months. Notably, 34% of victims reported ongoing psychological impact exceeding 24 months, highlighting the persistent nature of synthetic media harms.

4. Discussion

The Consent Paradox in Synthetic Data

Our findings reveal a central paradox in synthetic data practices: although synthetic data is promoted as privacy-preserving, it also enables large-scale derivative data creation without meaningful consent. In practice, the "synthetic" label often obscures ongoing privacy risks rather than eliminating them[10].

Traditional consent frameworks assume individuals can assess foreseeable risks and authorize specific, bounded uses of their data. Generative models disrupt this by enabling emergent capabilities that cannot be anticipated when consent is given, making consent temporally misaligned with later uses[11].

This paradox is evident in the gap between industry claims and empirical evidence. Synthetic data is often framed as containing "no personal information" and requiring no additional consent once training data is lawfully collected. Yet studies show that synthetic data can retain probabilistic links

to training data, enabling re-identification and inference attacks. Rather than being categorically safe, synthetic data exists on a privacy continuum that current consent practices fail to reflect.

The problem is intensified by rapid technological change. Data contributed under one set of expectations may later support unanticipated uses—such as deepfake generation or identity synthesis—as models are reused, fine-tuned, and combined across contexts. At the same time, structural power asymmetries in settings like healthcare, education, and employment can make refusal impractical, reducing consent to a procedural formality rather than a genuine choice.

Taken together, these dynamics show that synthetic data does not solve the consent problem but reframes it. By masking ongoing privacy risks behind claims of anonymization and innovation, current practices erode individual control precisely as data use becomes most expansive.

Deepfakes as Consent Violations

The rapid growth of non-consensual deepfakes reflects a systemic failure of existing digital consent architectures. Most cases involve sexualized content, causing severe psychological harm in legal environments that offer limited avenues for redress, thereby creating accountability gaps that enable abuse to scale with minimal deterrence[12].

Treating deepfakes solely as privacy violations or defamation obscures their distinctive nature. Unlike privacy harms, which involve unauthorized disclosure, or defamation, which concerns false statements, deepfakes fabricate audio-visual evidence of actions or speech that never occurred while retaining the credibility associated with video and audio records. They therefore constitute a distinct harm: **synthetic identity appropriation**, in which a person's likeness and identity attributes are used to construct persuasive false narratives.

Viewing deepfakes as consent violations offers a clearer conceptual basis. Individuals possess **identity sovereignty**—authority over how their face, voice, and embodied identity are used to represent them in public contexts. This interest is grounded not in secrecy or reputation alone, but in autonomy and the right to control self-representation. Deepfakes violate this sovereignty by generating representations that individuals neither authorized nor accepted.

Current legal frameworks remain poorly aligned with this harm. Defamation law struggles where no explicit claim is made, privacy law addresses disclosure rather than fabrication, and publicity rights are often limited to commercial misuse. Consequently, deepfake harms frequently fall between doctrinal categories, leaving victims without effective remedies (Levendowski, 2018).

Reframing deepfakes as violations of consent and autonomy, rather than merely privacy or reputation, reveals the depth of the governance failure. Without recognizing identity sovereignty as a protected interest, legal and technical responses will continue to lag the realities of synthetic media abuse.

Toward Technical Consent Mechanisms

Traditional click-through consent mechanisms are ill-suited to synthetic media environments, where data use is continuous, derivative, and difficult for individuals to anticipate. To address these limitations, we outline three technical mechanisms that operationalize consent as an enforceable property of synthetic media systems rather than a one-time procedural step.

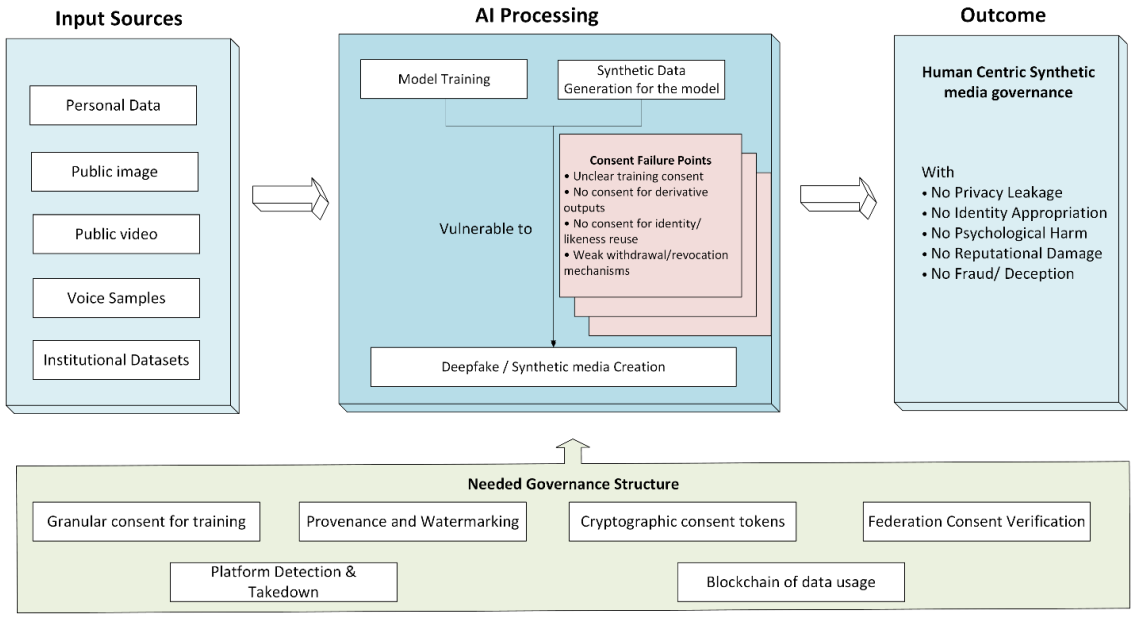


Figure 6: Flow diagram illustrating how data inputs move through AI generation, where consent breaks down, the resulting harms, and the governance mechanisms needed to restore human-centric control.

Cryptographic Consent Tokens

Cryptographic consent tokens enable individuals to grant fine-grained, verifiable consent for specific synthetic data uses through digitally signed authorizations stored in decentralized registries[13]. Unlike blanket consent for vague purposes, this approach allows consent to be scoped by use case, time period, and application domain, with immutable audit trails supporting accountability and revocation. While promising, widespread adoption will require addressing usability and scalability challenges.

Watermarking and Provenance Tracking

Watermarking and provenance systems embed cryptographic markers or consent metadata directly into original content, enabling downstream verification even after generative transformation[14]. When combined with provenance tracking, these techniques create traceable content lineages that allow platforms to assess whether synthetic media was created with valid authorization before distribution. Emerging initiatives such as the C2PA standard demonstrate growing industry alignment, though current implementations focus more on authenticity signalling than consent enforcement.

Federated Consent Verification

Federated consent verification enables real-time authorization checks at the point of synthetic content generation by querying decentralized consent registries rather than relying on centralized databases[15]. This approach prevents consent violations proactively while reducing single points of failure and limiting unnecessary disclosure of consent records. However, it requires coordination across institutions and interoperability between consent infrastructures.

Together, these mechanisms illustrate a shift from static, notice-based consent toward technically enforced, auditable consent architectures. While none is sufficient in isolation, their combination offers a viable foundation for aligning consent governance with the scale and dynamism of generative AI systems.

Regulatory Recommendations

Table 2. presents a proposed multi-layered regulatory framework addressing identified gaps

Layer	Mechanism	Scope	Enforcement	Implementation Timeline
Technical Standards	Mandatory watermarking for AI-generated content	All synthetic media platforms	Platform liability	12 months
Consent Architecture	Cryptographic consent verification systems	Facial recognition, voice synthesis	Criminal penalties for bypass	18 months
Detection Obligations	Platform duty to deploy deepfake detection	Social media, content platforms	Regulatory fines	24 months
Victim Protection	Expedited takedown + right to erasure	All online content	Civil liability + injunctions	6 months
Criminal Penalties	Non-consensual deepfake creation as crime	Sexual, fraudulent, defamatory content	Prosecution + imprisonment	12 months
Research Exemptions	Ethically-approved research with consent	Academic, safety research	Institutional review boards	18 months

Table 2: Multi-Layered Framework for Synthetic Media Governance Balances innovation with harm prevention through graduated oversight.

Limitations and Future Research

This study has several limitations. Reported deepfake incidents likely underestimate prevalence due to incomplete detection and reporting. Cross-jurisdictional legal comparisons are constrained by differing definitions and enforcement regimes. Synthetic data analysis focuses primarily on differential privacy, excluding alternative privacy-preserving techniques.

Future research should prioritize:

1. Longitudinal studies on the psychological impact of non-consensual synthetic media and empirical testing of consent-verification systems under adversarial conditions
2. Economic and cross-cultural analyses of consent frameworks, balancing regulatory costs with demonstrable harm-reduction outcomes

5. Conclusion

Synthetic data and deepfake technologies mark a critical point where generative AI capabilities have outpaced existing ethical and legal consent frameworks. This study demonstrates that consent models designed for explicit, one-time data collection fail when applied to generative systems capable of producing unlimited derivative content. In synthetic data and deepfake contexts, consent becomes ambiguous, temporally misaligned, and structurally weakened, leaving individuals with diminishing control over how their identities and data are re-used and represented.

To address this gap, we argue for reconceptualizing consent as a continuous, technically enforceable process rather than a singular authorization event. Technical mechanisms such as cryptographic consent verification, provenance tracking, and federated enforcement—combined with adaptive legal and institutional governance—offer a pathway toward restoring individual autonomy in synthetic media ecosystems. Successfully embedding consent into generative AI systems would provide a durable model for human-centric AI governance; failure to do so risks normalizing large-scale autonomy violations as a default condition of the digital future.

References

[1] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, *et al.*, "Synthetic data—what, why and how?" *Royal Society Open Science*, vol. 9, no. 6, p. 211820, 2022.

[2] D. Xu, S. Yuan, L. Zhang, and X. Wu, "FairGAN: Fairness-aware generative adversarial networks," in *Proc. 2018 IEEE Int. Conf. on Big Data*, 2018, pp. 570–575.

[3] J. Kietzmann, L. W. Lee, I. P. McCarthy, and T. C. Kietzmann, "Deepfakes: Trick or treat?" *Business Horizons*, vol. 63, no. 2, pp. 135–146, 2020.

[4] C. Vaccari and A. Chadwick, "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news," *Social Media + Society*, vol. 6, no. 1, pp. 1–13, 2020.

[5] R. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *California Law Review*, vol. 107, pp. 1753–1820, 2019.

[6] H. Ajder, G. Patrini, F. Cavalli, and L. Cullen, *The State of Deepfakes: Landscape, Threats, and Impact*. Deeptrace Labs, 2019.

[7] D. J. Solove, "Privacy self-management and the consent dilemma," *Harvard Law Review*, vol. 126, pp. 1880–1903, 2013.

[8] M. Cote, "Technoscientific transformations and the ethics of algorithmic governance," *Big Data & Society*, vol. 8, no. 1, pp. 1–14, 2021.

[9] M. Veale and R. Binns, "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data," *Big Data & Society*, vol. 4, no. 2, pp. 1–17, 2017.

[10] M. Hittmeir, A. Ekelhart, and R. Mayer, "Utility and privacy assessments of synthetic data for regression tasks," in *Proc. 2019 IEEE Int. Conf. on Big Data*, 2019, pp. 5763–5772.

[11] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data & Society*, vol. 3, no. 2, pp. 1–21, 2016.

[12] D. K. Citron, "Sexual privacy," *Yale Law Journal*, vol. 128, no. 7, pp. 1870–1960, 2019.

[13] G. Zyskind, O. Nathan, and A. Pentland, "Decentralizing privacy: Using blockchain to protect personal data," in *2015 IEEE Security and Privacy Workshops*, 2015, pp. 180–184.

[14] N. Yu, V. Skripniuk, S. Abdelnabi, and M. Fritz, "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data," in *Proc. IEEE/CVF Int. Conf. on Computer Vision*

(ICCV), 2021, pp. 14448–14457.
[15] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, *et al*., "Towards federated learning at scale: System design," *Proceedings of Machine Learning and Systems*, vol. 1, pp. 374–388, 2019.