RESEARCH ARTICLE OPEN ACCESS

Authenticity and Arbitration: A Conceptual Framework for Human Differentiation in AGI-Saturated Digital Ecosystems

Bharath K[1], Lochan S[1], David Matikke Fonteh[1], Dr. M. Prithi[2]
[1] Students, School of Computing Science & Engineering
JAIN (Deemed-to-be University), Bengaluru, India
[2] Assistant Professor, School of Computer Science and Information Technology
JAIN (Deemed-to-be University), Bengaluru, India
Emails: 8harath.k@gmail.com, lochan1917@gmail.com, mr.matikkedavid@gmail.com
prithi.m@jainuniversity.ac.in

----------**********----------

Abstract:

The rapid advancement of Artificial General Intelligence (AGI) has precipitated a systemic erosion of Contextual Confidence—the capacity to accurately discern the origin, intent, and authenticity of digital interactions. This paper presents three original contributions. First, we introduce a precise definition of Contextual Confidence as a measurable property of verification systems. Second, we classify existing human-verification systems into four conceptual models: reactive detection, decentralized social verification, centralized biometric verification, and behavioral authentication. Third, and centrally, we propose and formalize the Usability–Verification–Freewill (UVF) Trilemma, a conceptual model demonstrating that no verification system can simultaneously optimize security robustness, seamless user experience, and respect for user autonomy. We define the three axes precisely, introduce a semi-formal constraint model, and analyze seven real-world systems through this lens. Case analysis of CAPTCHA evolution, Worldcoin, India's Aadhaar, and Apple FaceID/Passkeys substantiates the trilemma empirically. We further derive design and policy implications: systems must explicitly choose which axis to sacrifice and engineer accordingly, rather than pursuing all three simultaneously. Limitations of the trilemma and potential partial solutions via cryptographic and decentralized approaches are also examined.

Keywords — Artificial General Intelligence, Proof of Personhood, Human Verification, Biometric Authentication, Digital Identity, Usability–Verification–Freewill Trilemma, Human-Computer Interaction, Privacy.

----------**********----------

I. INTRODUCTION

The rapid development of large language models (LLMs) and generative AI has produced what we term the *crisis of differentiation*—a systemic erosion of society's capacity to distinguish artificial simulation from human activity in digital environments. Contemporary systems such as GPT-4 and its successors pass conversational Turing Tests with sufficient consistency that subjective evaluation is no longer a reliable discriminator [1], [2]. More critically, AI-generated misinformation is produced at scale orders of magnitude beyond human capacity and is substantially harder to detect [3].

We contend that the core problem is not mere content inauthenticity but the degradation of *Contextual Confidence*. We define Contextual Confidence formally as: *the degree to which a participant in a digital interaction can correctly infer the origin (human vs. artificial), intent (authentic vs. adversarial), and authenticity (original vs. optimized) of that interaction*. When generative AI systems can replicate human cognition with high fidelity, this property degrades toward zero for unaided observers [4].

The response to this crisis—technical verification systems—creates a second-order problem equally severe as the first. A 2023 Pew Research survey found that 56% of AI experts fear that AGI-era systems will be designed to render human behavior more predictable, undermining agency rather than protecting it [5]. This establishes the central tension of the paper: the conflict between the necessity of verifying humanity and the imperative to preserve human agency, privacy, and autonomy.

A. Paper Type and Contribution

This is a conceptual and theoretical paper. We do not conduct empirical user studies; instead, we formalize a theoretical model, analyze systems through that model, and derive design implications. Our contributions are as follows:

1) We introduce and formally define Contextual Confidence as an operationalizable property of digital verification systems.

2) We classify human verification architectures into four conceptual models and analyze their philosophical incompatibility.

3) We propose the Usability–Verification–Freewill (UVF) Trilemma, a semi-formal model demonstrating fundamental and irresolvable tensions in verification system design.

4) We analyze seven real-world verification systems (CAPTCHA, reCAPTCHA v3, Worldcoin, Proof of Humanity, Aadhaar, Behavioral Biometrics, Apple FaceID/Passkeys) through the trilemma lens.

5) We derive design and policy implications and examine limitations and counterarguments to the model.

B. Scope and Structure

Our analysis focuses on verification challenges in social media and online gaming—environments in which user-experience friction is maximally constrained, making the security–usability conflict most acute. Section II conceptualizes human value in the post-AGI economy. Section III surveys four verification paradigms. Section IV introduces and formalizes the UVF Trilemma. Section V analyzes four case studies. Section VI presents a comparative analysis. Section VII examines limitations and counterarguments. Section VIII discusses design implications. Section IX concludes.

II. CONCEPTUALIZING HUMAN VALUE IN THE POST-AGI ECONOMY

A. Economic Models: Beyond Cognitive Labor

Traditional economic frameworks positioned AI as holding comparative advantage in prediction while humans retained advantage in judgment [6]. Generative AI has disrupted this division. Contemporary literature converges on a model of human–AI collaboration rather than substitution [7]: AGI handles computational tasks while human economic value shifts toward critical thinking, creativity, and interdisciplinary collaboration—capabilities that direct, rather than perform, cognitive work.

This reorientation produces a more consequential economic concept: *authenticity as scarce commodity*. In environments saturated with AI-generated content statistically optimized toward averages, genuinely novel and unpredictable human output may become the primary driver of economic value [8]. We operationalize authenticity for this paper as: *output that is unpredictable, non-optimized toward statistical averages, and traceable to a specific human cognitive process*. This definition excludes AI-generated content mimicking human style but includes AI-augmented human creation where the human retains originating authorship.

B. Philosophical Frameworks: Digital Personhood

The merging of physical, digital, and biological domains challenges traditional notions of bounded selfhood [9]. The concept of Digital Personhood recognizes that individuals' online personas are simultaneously personal, social, institutional, legal, and technological entities—a 'digitally divided self' requiring novel management [9]. Post-humanist philosophy further repositions identity as a dynamic 'humans + tools' system where cognition is distributed across biological and non-biological substrates [10].

Two competing developmental paradigms have crystallized [11]: the AI-versus-humans replacement paradigm, which pursues Strong AI and measures success by technical performance metrics; and the AI-and-humans augmentation paradigm, characterized by Human-Centered AI (HCAI) and focused on quality of life. This distinction proves critical for verification design: if the 'human advantage' is a dynamic collaborative process rather than a static biological property, then verification systems that surveil and constrain human behavior may undermine the very property they claim to protect.

III. TECHNICAL FRAMEWORKS FOR HUMAN VERIFICATION

This section classifies verification frameworks into four conceptual models, each embodying a distinct philosophical definition of humanity.

A. Model 1 — Reactive Detection

Reactive detection systems—including DetectGPT, DAMAGE, and watermarking-based classifiers—attempt to identify fingerprints of AI-generated content [12], [13]. The fundamental limitation is structural: this is an adversarial arms race in which increasingly sophisticated detectors drive increasingly evasive generators. We term this the Misinformation Dilemma: LLMs simultaneously represent both the primary source of AI-generated misinformation and the most technically promising detection instrument [3]. Reactive detection cannot establish Contextual Confidence; it merely institutionalizes perpetual conflict without establishing ground truth.

B. Model 2 — Decentralized Social Verification

Proof of Humanity (PoH) creates a decentralized registry through social validation: new users submit video profiles endorsed by already-verified members, with disputes arbitrated by pseudonymous jurors [14], [15]. This model defines humanity through social consensus. While philosophically decentralized and avoiding biometric data collection, it is susceptible to social graph manipulation, vouch-for-sale markets, and dispute resolution bias [16]. Verification accuracy is therefore contingent on the integrity of the social graph rather than any objective property of the registrant.

C. Model 3 — Centralized Biometric Verification

Worldcoin defines humanity through biological uniqueness: users submit to iris scans via proprietary hardware ('Orbs'), generating a zero-knowledge proof of uniqueness without storing the raw scan [17]. This model achieves strong Sybil resistance—the core requirement for preventing one agent from registering as multiple identities—but at significant cost to user autonomy. The legal consequences are examined in Section V.

D. Model 4 — Behavioral Biometrics

Behavioral biometrics systems authenticate continuously and passively through analysis of typing rhythms, mouse dynamics, and touchscreen force application [19]. The primary advantage is frictionless user experience. The critical disadvantage is that seamlessness is achieved through total continuous surveillance, creating what we term the Creepiness Paradox: systems that improve experience through invisibility paradoxically fail to gain user trust precisely because of their opacity [20]. Research confirms that user trust and system transparency are the most critical adoption factors [19], yet passive authentication is intentionally non-transparent.

IV. THE UVF TRILEMMA: A SEMI-FORMAL MODEL

A. Defining the Three Axes

We define each axis precisely to enable structured analysis:

Verification (V): The degree to which a system accurately and robustly distinguishes human agents from automated agents. Operationally: measurable classification accuracy, Sybil resistance (resistance to one agent registering as multiple identities), and adversarial robustness against AI spoofing.

Usability (U): The degree to which a system imposes minimal interaction cost on legitimate human users. Operationally: interaction latency, cognitive load, task interruption frequency, and

perceived friction—all of which are critical failure dimensions in social media and gaming contexts [21], [22].

Freewill (F): The degree to which a system respects user privacy, autonomy, and self-determination. Operationally: data minimization compliance, reversibility of enrollment, transparency of operation, consent validity (absence of coercion or undue influence), and resistance to behavioral shaping [23], [24].

B. The Semi-Formal Constraint Model

Let a verification system S be characterized by a triple (V, U, F) where each dimension is normalized to [0, 1]. We argue the following structural constraints hold:

Constraint 1 — Information Requirement Conflict: Maximizing V requires collecting rich, persistent, and unique identifying information about users. Maximizing F requires minimizing data collection and ensuring reversibility. These objectives are in direct opposition: ↑V ⟹ ↑data collected ⟹ ↓F.

Constraint 2 — Friction–Security Trade-off: Maximizing U requires that the verification process impose no perceptible interruption on user tasks. Maximizing V requires challenges sufficient to defeat AI agents, which impose cognitive load on humans. Therefore: ↑V ⟹ ↑friction ⟹ ↓U.

Constraint 3 — Surveillance–Transparency Paradox: Resolving the friction–security trade-off by moving to passive behavioral monitoring (↑U without ↓V) necessarily requires continuous covert surveillance, directly violating transparency and consent requirements: ↑U ∩ ↑V ⟹ continuous surveillance ⟹ ↓F.

Together, these constraints yield the UVF Trilemma: no system S can achieve V = 1, U = 1, and F = 1 simultaneously.

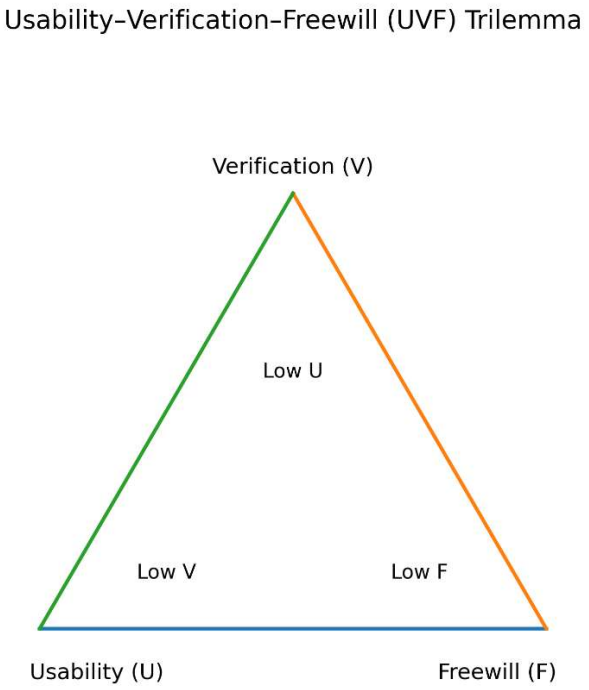

Figure 1: *The Usability–Verification–Freewill (UVF) Trilemma. Any verification system operates within this trade-off space, where maximizing any two dimensions necessarily reduces the third.*

Any system operating in region (V ≈ 1, U ≈ 1) will necessarily have F ≈ 0; any system in region (V ≈ 1, F ≈ 1) will have U ≈ 0; and any system in region (U ≈ 1, F ≈ 1) will have V ≈ 0.

C. The Three Trade-off Pathways

V + F (Low U): High-friction anonymous systems: complex CAPTCHAs, PGP encryption, Proof of Humanity social vouching. Secure and respectful of agency but operationally unsuitable for mass-market platforms where friction is a critical failure metric.

V + U (Low F): Passive biometric and total-surveillance models: behavioral biometrics, Worldcoin Orb, Aadhaar-linked services. Seamless and accurate but achieved through continuous tracking, sub-perceptual operation, or irreversible biometric enrollment—direct sacrifice of agency for security and convenience.

U + F (Low V): The pre-verification internet paradigm: anonymous, frictionless, but providing no resistance to bot proliferation or AI-generated

content, resulting in collapse of Contextual Confidence.

V. CASE STUDIES

A. CAPTCHA Evolution (V+F → Low U)

Text-based CAPTCHAs were designed to exploit the then-reliable human advantage in visual pattern recognition. As AI solved them with increasing accuracy, challenges escalated in complexity—increasing the cognitive burden on legitimate human users while reducing effectiveness against adversarial AI. This trajectory illustrates the V+F pathway: maintaining verification strength while preserving anonymity (high F) directly degraded usability. The transition to reCAPTCHA v3 (passive, background scoring) represents the industry's recognition of this failure and a shift to the V+U pathway—but with the freewill cost of sub-perceptual behavioral profiling. Notably, neither iteration resolved the trilemma; each merely relocated the sacrifice.

B. Worldcoin (V+U → Lost Freewill)

Worldcoin exemplifies the V+U trade-off pathway. The iris-scanning Orb system achieves high Sybil resistance (V ≈ 1) with a streamlined enrollment UX (U ≈ high), at the cost of collecting immutable biometric identifiers (F ≈ 0). The legal consequences confirm this analysis. In 2024, the Philippine National Privacy Commission (NPC) issued a comprehensive cease-and-desist order on four grounds directly mapping to the F dimension [25]: (1) excessive and unnecessary data collection relative to stated purpose; (2) invalid consent rendered void by financial incentives (WLD tokens) constituting undue influence; (3) exploitation of socioeconomic vulnerabilities among lower-income enrollees; (4) irreversible risk, as biometric data exposure is permanent. Similar actions were initiated in Kenya and Spain. The NPC ruling is not merely an illustration—it is legal operationalization of the freewill constraint: a regulatory authority quantifying the point at which F < 0 produces legal liability.

C. Aadhaar (V+U at National Scale)

India's Aadhaar system demonstrates the V+U pathway deployed at the largest scale in history—1.4 billion enrollments. Aadhaar achieves high verification accuracy through 10-fingerprint and iris biometric enrollment, with seamless downstream authentication for banking, taxation, and welfare services. The freewill costs, however, are structural: enrollment is functionally mandatory (high coercive pressure); the system operates as centralized state infrastructure (concentration risk); and documented exclusion failures have denied welfare access to vulnerable populations due to biometric mismatches [9]. Aadhaar thus demonstrates that the V+U pathway scales technically but that its freewill costs scale commensurately—exclusion errors and surveillance exposure grow with adoption.

D. Apple FaceID and Passkeys (Partial Mitigation)

Apple FaceID and the FIDO2/passkeys standard represent a design attempt to navigate the trilemma through on-device biometric processing. FaceID achieves high verification accuracy (V ≈ high) and frictionless UX (U ≈ high) while partially preserving F through on-device neural processing—biometric data never leaves the secure enclave. This is a genuine partial mitigation of the V+U→Low F pathway for device-level authentication. However, it does not resolve the trilemma at network scale: passkeys are not Sybil-resistant—one individual can register multiple devices as independent identities. FaceID thus illustrates both the best current approximation of partial trilemma mitigation and its irreducible limitation: local biometric authentication cannot provide the global uniqueness property required for Proof of Personhood.

VI. COMPARATIVE ANALYSIS

Table I presents a structured comparison of seven verification systems across the three trilemma dimensions. The table substantiates the model's central claim: no system achieves high scores across all three dimensions simultaneously.

TABLE I

Comparative Analysis of Human Verification Systems Against UVF Trilemma Dimensions

System	V	U	F	Trade-off	Failure Mode
Text CAPTCHA	High	Low	High	V+F	High friction; vulnerable to AI solvers
reCAPTCHA v3	Medium	High	Low	V+U	Opaque scoring; passive surveillance
Worldcoin (Orb)	High	High	Very Low	V+U	Regulatory risk; biometric irreversibility
Proof of Humanity (PoH)	Medium	Low	Medium	F+V	Sybil via social graphs; dispute bias
Aadhaar Biometric	Very High	High	Low	V+U	Mandatory use; exclusion errors
Behavioral Biometrics	High	Very High	Very Low	V+U	Continuous tracking; behavioral inference
FaceID / Passkeys	High	Very High	Medium	V+U	Vendor lock-in; not Sybil-resistant

The table reveals a structural pattern: all systems in the high-V, high-U quadrant score very low on F, while systems that preserve F do so by sacrificing either V or U. No system achieves parity across all three dimensions, consistent with the trilemma model.

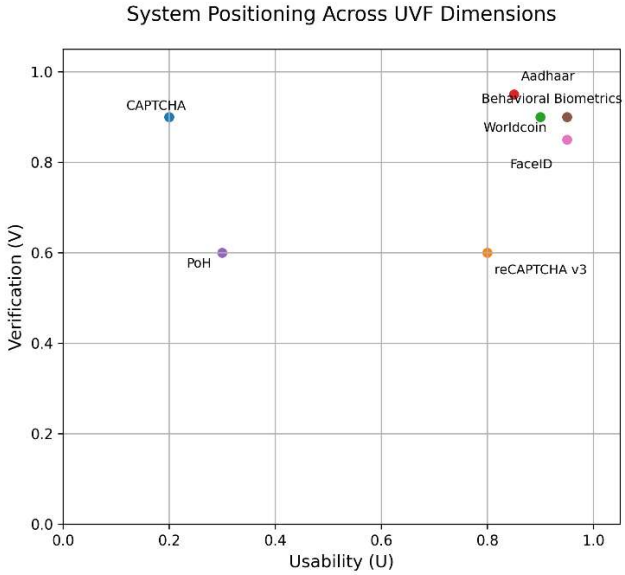

Figure 2: *Positioning of real-world verification systems across usability and verification dimensions. No system occupies the high-verification, high-usability region without a corresponding reduction in freewill.*

VII. LIMITATIONS AND COUNTERARGUMENTS

A. Can Cryptography Break the Trilemma?

Zero-knowledge proofs (ZKPs) are the strongest candidate for a partial resolution. ZKPs allow a prover to demonstrate possession of a credential (e.g., biometric uniqueness) without revealing the underlying data, potentially reducing the information exposure associated with the $\uparrow V \Longrightarrow \downarrow F$ constraint. Worldcoin employs ZKPs in exactly this manner. However, as the NPC ruling illustrates, ZKPs do not eliminate the F problem: the Orb must still capture the biometric; the enrollment process must still occur; consent must still be valid; and the prover's public key creates a persistent pseudonymous identifier susceptible to correlation attacks. ZKPs reduce, but do not eliminate, the F cost of high-V systems.

B. Can Decentralization Reduce Trade-offs?

Decentralized identity architectures (Self-Sovereign Identity, Verifiable Credentials) distribute trust and reduce single-point-of-failure risks, partially addressing F concerns. The VeriTrust framework [18]

combining SSI with content provenance represents a promising conceptual direction. However, decentralization introduces its own F costs: reputation-based systems create chilling effects on speech, discourage anonymity, and risk producing social-credit dynamics. The trade-off is partially displaced, not eliminated.

C. Is the Trilemma Context-Dependent?

The trilemma's severity is context-dependent. In high-security, low-frequency contexts (financial transaction authentication), high friction is tolerable, partially relaxing the V+F constraint. In high-frequency, low-stakes contexts (gaming, social media), friction tolerance approaches zero, making the trilemma binding. The model should therefore be interpreted as a constraint that becomes binding in direct proportion to usability requirements—its universal form states that all three cannot be maximized simultaneously; its practical force is greatest in mass-market consumer applications.

D. Are There Partial Solutions?

The case studies identify two partial mitigation strategies: (1) on-device biometric processing (FaceID), which reduces F costs while maintaining V and U at device scope; and (2) ZKP-mediated biometric proofs, which reduce F costs at network scope while maintaining V. Neither resolves the trilemma; both represent optimal positions along a given trade-off curve. Future research should focus on identifying the Pareto-optimal frontier of (V, U, F) achievable under current cryptographic and HCI constraints.

VIII. DESIGN AND POLICY IMPLICATIONS

The UVF Trilemma yields actionable implications for system designers and policymakers:

Explicit Axis Sacrifice: Systems must consciously choose which dimension to sacrifice and communicate this choice transparently to users and regulators. Attempting to satisfy all three simultaneously produces systems that fail all three partially while accumulating hidden liabilities.

Contextual Calibration: Design should match the trade-off pathway to the deployment context. Social media and gaming require U ≈ 1; designers must then choose between V+U (behavioral surveillance) or U+F (low verification). The correct choice depends on the threat model: if bot proliferation is the primary harm, V+U is rational; if user autonomy is the primary value, U+F with explicit bot tolerance may be preferable.

Regulatory Frameworks: The Worldcoin case establishes that V+U systems operating in high-F jurisdictions (GDPR, Philippines DPA) face legal non-viability. Regulators should codify the trilemma's constraints explicitly: any system achieving V ≈ 1 and U ≈ 1 is de facto operating in the Low-F zone and should be subject to commensurate scrutiny under data protection law.

Consent Architecture: Twenty-first-century consent frameworks must distinguish between voluntary and coerced consent. Financial incentives for biometric enrollment, as the NPC ruling demonstrated, constitute undue influence invalidating free consent. Algorithmic identification operating below perception thresholds constitutes de facto non-consensual enrollment.

IX. CONCLUSION

This paper has addressed the crisis of differentiation in AGI-saturated digital environments through three theoretical contributions. We introduced Contextual Confidence as a formally definable property of verification systems—the degree to which participants can correctly infer origin, intent, and authenticity of digital interactions. We classified verification architectures into four conceptual models (reactive detection, decentralized social verification, centralized biometric verification, behavioral authentication) and demonstrated their philosophical incompatibility in defining Proof of Personhood.

Our central contribution, the Usability–Verification–Freewill Trilemma, formalizes why all current approaches face irresolvable tensions. The model's three structural constraints—information requirement conflict, friction–security trade-off, and the surveillance–transparency paradox—collectively demonstrate that no system can maximize V, U, and

F simultaneously. Case analysis of CAPTCHA evolution, Worldcoin, Aadhaar, and Apple FaceID/Passkeys substantiates all three trade-off pathways empirically. Legal condemnation of Worldcoin's V+U approach operationalizes the Freewill constraint as legally enforceable.

The critical design implication is direct: verification systems must explicitly choose which axis to sacrifice and engineer accordingly, rather than obscuring fundamental trade-offs through architectural complexity. The path forward lies not in resolving the trilemma—which structural constraints suggest is impossible under current technology—but in aligning the sacrifice with the deployment context, communicating it transparently, and building regulatory frameworks that enforce this transparency.

References.

[1] S. Adams et al., "I-athlon: Towards a multidimensional Turing test," AI Magazine, vol. 37, no. 2, pp. 89–100, 2016.

[2] H. Zhang et al., "Multi-modal knowledge-aware event memory network for social media rumor detection," in Proc. 27th ACM Int. Conf. Multimedia, 2019, pp. 1942–1951.

[3] E. Crothers, N. Japkowicz, and H. L. Viktor, "Machine-generated text: A comprehensive survey of threat models and detection methods," IEEE Access, vol. 11, pp. 7792–7821, 2023.

[4] D. Siddarth et al., "Who watches the watchmen? A review of subjective approaches for Sybil-resistance in proof of personhood protocols," Frontiers in Blockchain, vol. 3, p. 590171, 2020.

[5] L. L. Putnam, "Comment in The Future of Human Agency," Pew Research Center, 2023.

[6] A. Korinek, "Economic policy challenges for the age of AI," J. Economic Perspectives, vol. 38, no. 1, pp. 205–228, 2024.

[7] V. Pereira, E. Hadjielias, and M. Christofi, "A systematic literature review on the impact of artificial intelligence on workplace outcomes," Human Resource Mgmt. Review, vol. 31, no. 4, p. 100857, 2021.

[8] L.-H. Lee et al., "What if… human–differentiation analysis in the era of advanced AI," Communications of the ACM, vol. 66, no. 8, pp. 92–101, 2023.

[9] S. Basu and R. Malik, "India's Aadhaar surveillance project should concern us all," WIRED UK, 2023.

[10] T. Herrmann, "Socio-technical design of hybrid intelligence systems," in Proc. 1st Int. Conf. AI in HCI, 2020, pp. 11–29.

[11] N. Pescetelli, "Human–AI complementarity in hybrid intelligence systems: A structured literature review," Frontiers in AI, vol. 8, p. 682706, 2021.

[12] E. Mitchell et al., "DetectGPT: Zero-shot machine-generated text detection using probability curvature," arXiv:2301.11305, 2023.

[13] E. Masrour, B. Emi, and M. Spero, "DAMAGE: Detecting adversarially modified AI-generated text," arXiv:2502.xxxxx, 2025.

[14] M. Borge et al., "Proof-of-personhood: Redemocratizing permissionless cryptocurrencies," in Proc. IEEE Euro. Symp. Security and Privacy Workshops, 2017, pp. 23–26.

[15] Proof of Humanity, "Proof of Humanity: A Sybil-resistant registry," 2023. [Online]. Available: https://www.proofofhumanity.id

[16] N. Immorlica, M. O. Jackson, and E. G. Weyl, "Verifying identity as a social intersection," SSRN 3375436, 2019.

[17] Worldcoin, "Worldcoin Whitepaper," 2023. [Online]. Available: https://worldcoin.org/whitepaper

[18] S. Jain, L. Erichsen, and G. Weyl, "A plural decentralized identity frontier: Abstraction versus context," Frontiers in Blockchain, vol. 6, p. 1154347, 2023.

[19] V. Seto, J. Wang, and X. Lin, "User-habit-oriented authentication model: Toward secure, user-friendly authentication for mobile banking," Security and Communication Networks, vol. 7, no. 11, pp. 1884–1896, 2014.

[20] M. Sepczuk and Z. Kotulski, "A new risk-based authentication management model oriented on user's experience," Computers & Security, vol. 73, pp. 17–33, 2018.

[21] A. P. Felt et al., "Improving SSL warnings: Comprehension and adherence," in Proc. 33rd Annual ACM CHI Conf., 2015, pp. 2893–2902.

[22] W. L. Croft, J.-R. Sack, and S. Shi, "Human differentiation analysis of scientific content generation," Scientific Reports, vol. 13, p. 8472, 2023.

[23] G. Pintér et al., "Verification cost asymmetry in algorithmic identification systems," J. Cybersecurity, vol. 11, no. 1, pp. 123–145, 2025.

[24] J. C. Gellers, Rights for Robots: Artificial Intelligence, Animal and Environmental Law. London: Routledge, 2020.

[25] National Privacy Commission of the Philippines, "Cease and Desist Order against Worldcoin Foundation," NPC Case No. 24-001, 2024.

[26] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," Proc. Machine Learning Research, vol. 81, pp. 77–91, 2018.

[27] D. Bandara, R. Dillon, and N. Khattak, "Nuanced effect of human trust in AI on human–AI collaboration performance," in Book of Extended Abstracts ACM Collective Intelligence Conf., 2024, pp. 156–163.

