Private and Decentralized Age Verification Architecture

Sofía Celi
Brave Software & University of Bristol

Kyle den Hartog
Brave Software

Hamed Haddadi
Brave Software & Imperial College London

Christian Knabenhans
EPFL

Elizabeth Margolin
University of Pennsylvania

Abstract

Today, it's widely acknowledged that we face serious challenges in controlling what content is accessible to children online, and more importantly we currently lack effective tools to address this in a privacy-preserving and effective way. The core difficulty lies in the tradeoffs involved. But if we approach the problem thoughtfully, we can strike a balance: preserving the Web's openness and user agency, minimizing unnecessary data collection and privacy harms, and empowering guardians (whether parents, teachers, or school IT administrators) to better manage what children are exposed to. In doing so, we might even unlock broader benefits, such as combating misinformation, and curbing manipulation and fraud by bots or foreign actors. These are ambitious goals: but how do we make them a reality?

1 Introduction

Efforts to protect children online frequently converge on the concept of *age verification*. However, the prevailing framing of this challenge (verifying the user's age to gate access to content) conflates multiple underlying issues into a single, overly rigid mechanism. As a result, current proposals often compromise user privacy, enforce centralized authority, and fail to scale effectively across the Web's decentralized infrastructure.

To build a system that upholds core Web values, such as user agency, privacy, and openness, we must reframe the problem. What is commonly referred to as "age verification" is in fact a fusion of two distinct but interconnected challenges:

- **The Content Moderation Problem [21, 23, 24, 40]:** How can content be restricted in a technically enforceable, scalable, and verifiable privacy-preserving manner? How does it match user-expectations, interpretations of online harm that across cultures [25], communities [42], and individuals; and interface design?
- **The Guardianship Problem:** Who decides what content should be restricted, and by what mechanism is that decision enforced?

In the sections that follow, we explore these two problem domains. We first examine the limitations and opportunities surrounding content moderation on the Web. We then address the guardianship challenge, highlighting how enforcement mechanisms can be decentralized and tailored to reflect local values and preferences. Finally, we show how this modular approach leads to a more robust, equitable, and scalable framework for protecting children online, and how it may also generalize to broader challenges, such as curbing misinformation and automated fraud.

2 The Content Moderation Problem

The content moderation problem can be broadly defined as the technical and policy challenge of enabling users (or services acting on their behalf) to filter, block, or otherwise regulate access to specific content on the Web due to its perceived harmful nature. Although the conceptual goal is straightforward, current approaches are limited both in scope and in scalability. Most notably, the ability to filter content remains inconsistent across platforms, lacks standardization, and often depends on brittle or non-generalizable heuristics. At a practical level, comprehensive and fine-grained content moderation remains computationally infeasible for all but the largest content providers. As a result, most solutions rely on coarse-grained techniques, such as domain-level filtering, or employ simplistic rule-based systems.

One example is network-level content filtering [22, 26], most commonly implemented at the DNS layer. Internet Service Providers (ISPs) often use this method to block access to domains associated with malicious or illegal activity. End users may configure similar protections using tools like PiHole or through DNS-based features embedded in consumer VPNs. These systems are primarily used to block malware, trackers, or advertising domains, and operate at a level of abstraction that lacks semantic understanding of the actual content.

Site-level moderation tools [18] are also available but tend to be platform-specific and highly fragmented. Common features include *block*, *mute*, *filter*, and *report* functionalities. For instance, a user on social media may choose to block another user, mute specific keywords (e.g., "Elon Musk"), or report posts labeled as Not Safe For Work (NSFW). These mechanisms, while effective in limited contexts, vary widely across platforms in terms of availability, usability, and enforcement consistency. Some modern platforms adopt a heuristic-driven model that prioritizes and filters content based on domain-level reputational signals. These systems are typically underpinned by manually curated lists, originally intended for adblocking, which are maintained by a small number of contributors. While effective for certain categories of content (e.g., advertisements or malware), these approaches do not scale well to the full semantic range of content requiring moderation.

Fundamentally, content moderation poses a semantic and normative challenge that resists centralized enforcement. Centralized service providers, by acting as arbiters of classification, introduce bottlenecks and introduce both interpretive and logistical limitations. Two central questions illustrate the depth of this problem:

(1) **Who determines whether content is classified correctly?** Manual classification introduces subjectivity and variability across moderators, while algorithmic approaches (such as those based on machine learning classifiers) are limited by biases in training data and the inherent opaqueness of model decisions.
(2) **How can moderation mechanisms apply across the long-tail of the Web, including small-scale or decentralized platforms and real-time media such as livestreams?** In these contexts, jurisdictional complexity and technical limitations render centralized enforcement impractical or ineffective.

Historical analogs, such as the discretionary censorship reveal a persistent dynamic: enforcement, even when technologically mediated, is ultimately subject to the discretion and interpretation of the censor. This creates a structural vulnerability to the so-called *chilling effect*, in which individuals self-censor in anticipation of punitive action. Consequently, centralized architectures for content moderation not only struggle with scale and accuracy, but may also exacerbate harms to user autonomy and freedom of expression. These limitations underscore the need for alternative, decentralized approaches that better align with the principles of privacy, scalability, and moral pluralism.

A critical but often overlooked dimension of content moderation debates is the rights of children themselves [39]. While much of the discourse on online safety is framed around protecting minors from explicit or harmful content, typically invoking examples such as pornography or violent media, the same regulatory tools and rhetorical framings are frequently weaponized to restrict access to identity-affirming or informational resources. In several jurisdictions, laws justified under the banner of "child protection" have been used to block or remove content related to LGBTQ+ identities [5, 37], gender-affirming healthcare, and sexual education. This raises profound ethical and political concerns about who defines what is "child-appropriate" and on whose behalf those decisions are made. The appeal of a decentralized, user- or guardian-configurable content moderation framework is that it resists the imposition of a singular moral standard dictated by centralized actors: whether governments or platform owners.

3 The Guardianship Problem

Complementary to the content moderation problem is what we refer to as the *guardianship problem*. This problem centers on a fundamental normative and architectural question: *who* determines which content should be accessible to which users, and *how* that decision is enforced in practice.

Contemporary approaches to age verification (as codified in numerous legislative efforts across jurisdictions) typically assume that governments should serve as the primary authority in defining and enforcing access control to online content. This is operationalized by imposing legal obligations on centralized service providers (often referred to as *relying parties*) to verify the age of users and regulate access accordingly. These providers, in turn, are incentivized to comply under the threat of regulatory penalties.

At a technical level, most current systems adopt a model based on third-party attestation. Under this approach, a trusted third party issues a credential or assertion regarding a user's age or eligibility. This credential is then presented by the user to the service provider to gain access to content gated by age or other regulatory criteria. However, these systems are typically implemented in a centralized, server-side fashion. Credentials are issued by a limited set of recognized authorities (henceforth, the "issuer"), and verification occurs over the plaintext data (with access to the full information on the credential and identity of the issuer) on the server hosting the gated content. This architecture carries significant implications: it effectively shifts the locus of guardianship away from individuals (e.g., parents, teachers, or local institutions) and toward centralized issuers and content providers, who must enforce content policies uniformly across all users. In doing so, these actors are forced to apply a single set of moral and regulatory judgments across a globally heterogeneous user base.

Such an arrangement introduces further several concerns. First, it risks eroding user agency by denying guardians closer to the user the ability to enforce content policies that reflect local values or individual preferences. Second, it introduces privacy risks, as users must disclose sensitive personal attributes (e.g., age and any fields that the credential has) and the identity of the issuer (which introduces risks in terms of deanonymizing the user via its citizenship, for instance) to access content. Finally, centralized guardianship mechanisms are structurally misaligned with the decentralized and user-centric ethos of the Web, limiting their effectiveness and scalability in heterogeneous environments.

This motivates the need for alternative models of guardianship that decentralize enforcement authority, preserve individual agency, and reduce reliance on centralized verification infrastructures.

3.1 Limitations of Digital Credential Frameworks

Recent regulatory proposals, such as the European Union's Digital Identity Framework [1] (eIDAS 2.0) [15] and similar efforts in the UK [38], US [27, 34, 35], Canada [8], Australia[3], and elsewhere, emphasize digitally verifiable credentials, particularly those augmented with privacy-preserving primitives such as zero-knowledge proofs (ZKPs) [2], as a foundation for online age verification and content access control. These frameworks envision users receiving cryptographically signed

[1] See a response of cryptographers to it [30].

attestations (e.g., of age, residency, or citizenship) from authorized entities, which they can then present to online services to demonstrate eligibility. Similar cryptographic mechanisms are being explored in academic and industry contexts, particularly in blockchain systems:

- **zk-TLS:** Systems such as [1, 13, 29, 43, 45, 46] allow a user to commit to a TLS session transcript and later prove facts about their interaction with an unmodified web service (e.g., account balances), enabling the construction of web-based oracles for smart contracts.
- **zk-Authorization:** Tools like zkLogin [9] and zkCreds [36] allow users to prove properties about a JSON Web Token (JWT) from an OAuth or OpenID provider, without revealing the token itself.
- **zk-Compilation:** Proofs that an executable or bytecode (e.g., ELF, WASM) corresponds to a known source and compiler toolchain [16], supporting provenance and auditability.
- **zk-Optimization:** Systems such as Otti [6] allow proving that a private optimization process (e.g., university admissions) was computed according to committed policies and inputs, without revealing the inputs.
- **zk-Middleboxes:** Given a commitment to network protocol streams (e.g., DNS), a sender can prove that traffic satisfies policies without revealing its content [20, 33, 44].

Despite these advances, credential-based systems face multiple limitations in both theory and deployment.

Security fragility. The use of ZKPs in these systems remains fragile. Many cryptographic protocols described as 'zero-knowledge" in academic literature do not satisfy rigorous formal definitions (are not zero-knowledge—which preserve privacy– or are not sound–which prevents forgeability–) or composability guarantees [12, 14, 17, 28, 32]. Moreover, their security often relies on idealized models (e.g., the random oracle model), and negative results exist for general composability in more realistic models [10, 11, 19]. Implementation is also non-trivial: many ZKP systems, even well-audited ones, have suffered from subtle vulnerabilities [31].

Insufficient privacy guarantees. ZKPs do not guarantee meaningful privacy unless applied carefully and correctly [41]. For example, a proof that a user's age lies in the range 20–21 may inadvertently disclose that the user is indeed 21. In practice, even semantically "private" range proofs may leak significant information based on how they are constructed or interpreted. Moreover, privacy loss can compound over time when multiple proofs are issued across a temporal sequence. For instance, a user who first proves their age is between 20 and 21, and then two months later proves they are over 21, has effectively disclosed a narrow interval for their exact date of birth. This temporal leakage highlights the need for systems to manage *privacy through change*, not just in isolated proofs, but across repeated interactions. Providing users with transparency about the cumulative privacy loss from sequences of attribute-based disclosures, and establishing baseline guarantees of unlinkability across sessions, will be essential to the safe deployment of such systems.

Furthermore, if not applied with appropriate scoping constraints, ZKP systems can devolve into a form of client-side scanning [4], where arbitrary attestations are made about the contents of committed data such as TLS transcripts, JWT tokens, or digital credentials. Such attestations may inadvertently break the security or privacy guarantees of the underlying protocol or data structure. It must therefore be verifiable, and externally auditable, that the statement being proved does not enable indirect deanonymization or policy circumvention.

Centralization and inclusion risks. These systems often rely on a small set of recognized credential issuers (e.g., governments, telecom providers, account providers). Only credentials from these issuers are accepted by relying parties, leading to exclusion of individuals without formal IDs, residence, or institutional affiliation. Moreover, the issuer's identity is usually disclosed to the verifier, which itself may leak sensitive information, such as citizenship or jurisdiction.

Parsing and semantic mismatch. A central, and often underappreciated, limitation in current ZKP-based systems lies in the semantic gap between low-level commitments (e.g., raw byte streams) and the structured data representations over which the zero-knowledge proof is intended to operate. Many existing systems implicitly assume that the input is already well-formed: for example, that a JSON object adheres to the appropriate grammar [9, 46], or that a digital credential conforms to a standardized syntax[2]. However, in the absence of formal guarantees about parsing correctness, malformed or adversarially crafted byte streams can undermine soundness. For instance, `zkLogin` [9] assumes that JSON keys do not contain escape characters, violating this assumption enables attacks that can break the system's security.

To mitigate these issues, some systems reveal select portions of the data to the verifier for direct inspection, but this undermines the core privacy properties that ZKPs are meant to preserve. Moreover, the challenge is not only in verifying the presence of a particular value (e.g., an `age` number value), but also in verifying its *position* and *context* within the document structure. For example, a valid proof must ensure that the `age` value appears as the value associated with a top-level "age" key, rather than being nested under an unrelated field or fabricated through structural ambiguity. Without formally verified parsing, the proof system cannot soundly claim that the input satisfies the intended property.

Additional ecosystem limitations. Several further challenges:

- **Lack of protocol standardization:** No widely adopted ZKP protocol stack exists across jurisdictions or vendors, hampering interoperability.

[2]A recent work [7] tackles this problem in an efficient and correct manner.

- **Interoperability concerns:** Even compatible schemes vary in encoding, supported predicates, and proof formats, fragmenting the ecosystem.
- **Revocation challenges:** Privacy-preserving revocation mechanisms are underdeveloped and hard to integrate with unlinkability guarantees.
- **Trust ambiguity:** Even with ZKPs, the verifier must know whether to trust the issuer, reintroducing central trust dependencies.
- **Poor user experience:** Asking users to configure selective disclosure or predicate proofs creates cognitive and UX burdens.

The above concerns are not meant to dismiss the value of zero-knowledge proofs, which remain a powerful and essential tool in privacy-preserving system design. Rather, they highlight the risks of deploying ZKPs "out-of-the-box," without rigorous formal verification, comprehensive security analysis, user-centered interface design, and thoughtful integration into the broader system architecture.

Empowering Users in the Proof Generation Process. Beyond the limitations outlined above, we believe there is significant promise in interaction models that give users greater visibility and agency in the proof-generation process itself. One promising design pattern is to allow the user to inspect and possibly modify the statement to be proved before it is compiled into a zkCircuit [3]. In this model, the user is not simply asked to passively authorize a proof of possession of some attribute (e.g., age, location, or citizenship), but is empowered to participate in constructing the proof in a transparent way.

More advanced variants can allow the user to not only inspect the statement but also modify the *data* being passed into it. In such cases, the application mediating the proof can additionally construct a zero-knowledge proof of *equivalence* between the original (private) and modified (public) values. This proof ensures that the semantic meaning of the original predicate is preserved, even under redacted or transformed inputs. The transformed data and the equivalence proof are then jointly passed to the authentication or relying-party system for verification.

This interaction pattern supports both privacy and user autonomy: users gain meaningful insight into what is being proved, and can validate or sanitize the inputs before the proof is constructed. Trust in the compilation and transformation process can be distributed across trusted third-party tooling, formal verification pipelines, or endorsement schemes, adding flexibility while maintaining integrity.

4 An Alternative Approach

A more private, decentralized, and principled solution to the challenges of filtering via age verification becomes possible when we disentangle the *content moderation* and *guardianship* problems and address them independently.

To address the content moderation problem, we begin with the assumption that content can be classified into granular categories (e.g., `safe`, `unsafe`, `adult`, `violent`, etc.). This assumption holds whether moderation occurs through centralized mechanisms (e.g., server-side filtering based on age-verification credentials) or decentralized ones, such as client-side filtering informed by curated content lists. In current practice, privacy-preserving mechanisms such as SafeBrowsing and adblocking systems (e.g., EasyList, uBlock Origin, Brave Shields) rely on domain-level heuristics to filter requests in the browser. While limited in semantic depth, these heuristics have proven effective in many contexts. More recent developments, such as SafeBrowsing v5 in Google Chrome, introduce on-device real-time classification to detect malicious or harmful content. This paradigm could be extended to encompass other content types by, for example, introducing semantic classification tags in HTML served by websites. A client-side browser engine could then consult user-defined or guardian-enforced policies to determine whether content should be rendered, blurred, blocked, or require additional approval.

Such a system would support a more generalizable, user-centered form of content filtering. Individual users could define personal filters (e.g., blocking all mentions of a specific public figure), or subscribe to third-party curated lists tailored to particular moral, cultural, or informational goals. Importantly, this approach satisfies the principle of agency by default: the system would be opt-out rather than mandatory, allowing users to configure their experience or defer it to their trusted entities.

However, in the context of child protection, the challenge lies in ensuring that such settings cannot be easily circumvented. Here, the guardianship problem becomes relevant. Existing enforcement mechanisms in educational contexts (such as device management and network-level filtering by IT administrators) can be extended to support system-level configuration of content moderation policies. If integrated into the operating system, such policies could propagate to applications (e.g., browsers, games) in a uniform and enforceable manner. In this model, the browser would default to blocking content that cannot be confidently classified as appropriate. Guardians (e.g., parents, teachers, or school IT administrators) could authorize temporary access through explicit override mechanisms, or configure the system to log access attempts for later review. For bring-your-own-device (BYOD) scenarios, guardians could configure devices through mobile device management (MDM) solutions or consent to remote provisioning policies from trusted institutions. This architecture decentralizes enforcement, allowing guardianship to be delegated to a broad and diverse ecosystem of actors. Rather than relying on centralized institutions to issue and verify credentials, enforcement can be grounded in local norms and preferences. For example, some parents may choose to allow access to sensitive educational material, while others may restrict access to certain topics. This flexibility

[3] A *zkCircuit* is a low-level, fixed-function representation of a computation compiled into arithmetic constraints suitable for zero-knowledge proving systems. It encodes a specific function or statement as an arithmetic circuit, usually defined over a finite field, where proving involves showing that a secret input (the *witness*) satisfies the circuit constraints.

avoids the imposition of monolithic content standards and mitigates the identity politicization that often accompanies centralized regulatory schemes.

Privacy-preserving digital credentials still play a role in this architecture, but the trust model shifts. Instead of requiring issuance by a globally recognized authority (e.g., a government agency), credentials can be issued by authorized guardians identified via decentralized identifiers (DIDs). For instance, a teacher could issue a temporary access credential from a managed device to permit a student to view a specific website. The browser (or application) acts as the verifier of this credential, and the device's operating system serves as the holder. Because the number of trusted issuers is bounded by the configuration of the device, there is no need to establish global interoperability between all credential authorities.

From the user's perspective, the experience remains seamless. If a child attempts to access gated content, and no valid credential is available, the content is blocked (e.g., via a browser interstitial or page blurring). A credential request can then be issued to the guardian's device, which prompts for approval with contextual information (e.g., target site, duration of access, logging preferences). This mechanism enables fine-grained, consent-based access control, with transparency and auditability configurable by the guardian.

This architecture satisfies key privacy goals by avoiding server-side access to user credentials and personal information. Sites are not required to learn users' ages, identities, or guardian relationships. Their role is limited to tagging content, either manually or via automated tools, with appropriate classification metadata. Regulators may still issue guidelines for classification schemes, but ultimate enforcement lies with the user or guardian, not with the server.

Note however that while decentralization offers important benefits in resisting centralized censorship and enabling local control, it also introduces risks when enforcement authority is delegated to entities that may themselves act oppressively or discriminatorily. For instance, entrusting schools or parents with full guardianship over content access can result in restrictions on identity-affirming information, particularly for LGBTQ+ youth. In certain jurisdictions, school administrations have used filtering tools to block access to resources on gender identity, sexual health, or queer advocacy, under the guise of protecting children. Similarly, parental control technologies can be weaponized to isolate, or suppress a child's access to affirming or educational content: a phenomenon that some scholars and advocates have described as a form of *intimate digital violence*.

These scenarios illustrate a key tension in decentralization: while it dismantles centralized, one-size-fits-all content regimes, it may also re-inscribe hierarchical power structures within the family or local institutions. The challenge, then, is to design systems that uphold children's rights to information [39] while still allowing for age-appropriate protections. This might require new forms of accountability, transparency, and recourse—such as allowing children to appeal filtering decisions, or integrating multiple guardian roles (e.g., involving educators, counselors, or rights advocates) to mitigate the risks of unilateral control.

In sum, by reassigning roles in the content access trust architecture, we enable a more decentralized, privacy-preserving, and morally pluralistic system. Users can choose whether to self-moderate, delegate moderation to trusted third parties, or defer decisions to guardians. Credential issuance becomes local and ephemeral, rather than centralized and persistent. This approach not only strengthens protections for children online, but also provides a blueprint for addressing other forms of content-related harms (such as misinformation) through decentralized and agency-preserving mechanisms.

Acknowledgments

We thank François Dupressoir and Carmela Troncoso for valuable discussions and feedback that shaped this work.

References

[1] TLS Notary, 2023. https://tlsnotary.org/.

[2] G: Zero Knowledge Proof. https://eu-digital-identity-wallet.github.io/eudi-doc-architecture-and-reference-framework/latest/discussion-topics/g-zero-knowledge-proof/, 2024. Accessed July 2025.

[3] 1News. Debate rages as australia set to ban children from social media. https://web.archive.org/web/20240910083505/https://www.1news.co.nz/2024/09/10/debate-rages-as-australia-set-to-ban-children-from-social-media/, September 2024. Archived via Wayback Machine; Accessed July 2025.

[4] Harold Abelson, Ross Anderson, Steven M Bellovin, Josh Benaloh, Matt Blaze, Jon Callas, Whitfield Diffie, Susan Landau, Peter G Neumann, Ronald L Rivest, Jeffrey I Schiller, Bruce Schneier, Vanessa Teague, and Carmela Troncoso. Bugs in our pockets: the risks of client-side scanning. *Journal of Cybersecurity*, 10(1), January 2024.

[5] Avram Anderson and A. L. Roth. Queer erasure: Internet browsing can be biased against lgbtq people, new exclusive research shows. *Index on Censorship*, 49(1):75–77, 2020.

[6] Sebastian Angel, Andrew J. Blumberg, Eleftherios Ioannidis, and Jess Woods. Efficient representation of numerical optimization problems for SNARKs. 2022.

[7] Sebastian Angel, Sofía Celi, Elizabeth Margolin, Pratyush Mishra, Martin Sander, and Jess Woods. Coral: Fast succinct non-interactive zero-knowledge CFG proofs. Cryptology ePrint Archive, Paper 2025/1420, 2025.

[8] ARPA Canada. Age verification bill reintroduced in the senate. https://arpacanada.ca/articles/age-verification-bill-reintroduced-in-the-senate/, 2025. Accessed July 2025.

[9] Foteini Baldimtsi, Konstantinos Kryptos Chalkias, Yan Ji, Jonas Lindstrøm, Deepak Maram, Ben Riva, Arnab Roy, Mahdi Sedaghat, and Joy Wang. zklogin: Privacy-preserving blockchain authentication with existing credentials. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2024.

[10] B. Barak. How to go beyond the black-box simulation barrier. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 106–115, 2001.

[11] James Bartusek, Liron Bronfman, Justin Holmgren, Fermi Ma, and Ron Rothblum. On the (in)security of kilian-based SNARGs. Cryptology ePrint Archive, Report 2019/997, 2019.

[12] Sofía Celi, Shai Levin, and Joe Rowell. CDLS: Proving knowledge of committed discrete logarithms with soundness. In Serge Vaudenay and Christophe Petit, editors, *AFRICACRYPT 24*, volume 14861 of *LNCS*, pages 69–93. Springer, Cham, July 2024.

[13] Sofía Celi, Alex Davidson, Hamed Haddadi, Gonçalo Pestana, and Joe Rowell. DiStefano: Decentralized infrastructure for sharing trusted encrypted facts and nothing more. Cryptology ePrint Archive, Paper 2023/1063, 2023.

[14] Quang Dao, Jim Miller, Opal Wright, and Paul Grubbs. Weak fiat-shamir attacks on modern proof systems. In *2023 IEEE Symposium on Security and Privacy*, pages 199–216. IEEE Computer Society Press, May 2023.

[15] European Commission. Proposal for a Regulation of the European Parliament and of the Council amending Regulation (EU) No 910/2014 as regards establishing a framework for a European Digital Identity. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0281, June 2021. COM/2021/281 final.

[16] Zhiyong Fang, David Darais, Joseph P Near, and Yupeng Zhang. Zero knowledge static program analysis. 2021.

[17] Ariel Gabizon. On the security of the BCTV pinocchio zk-SNARK variant. Cryptology ePrint Archive, Report 2019/119, 2019.

[18] Sarah A. Gilbert. "i run the world's largest historical outreach project and it's on a cesspool of a website." moderating a public scholarship site on reddit: A case study of r/askhistorians. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), May 2020.

[19] Shafi Goldwasser and Yael Tauman. On the (in)security of the Fiat-Shamir paradigm. Cryptology ePrint Archive, Report 2003/034, 2003.

[20] Paul Grubbs, Arasu Arun, Ye Zhang, Joseph Bonneau, and Michael Walfish. Zero-knowledge middleboxes. 2022.

[21] Anaobi Ishaku Hassan, Aravindh Raman, Ignacio Castro, Haris Bin Zia, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. Exploring content moderation in the decentralised web: the pleroma case. In *Proceedings of the 17th International Conference on Emerging Networking EXperiments and Technologies*, CoNEXT '21, page 328–335, New York, NY, USA, 2021. Association for Computing Machinery.

[22] Nguyen Phong Hoang, Arian Akhavan Niaki, Jakub Dalek, Jeffrey Knockel, Pellaeon Lin, Bill Marczak, Masashi Crete-Nishihata, Phillipa Gill, and Michalis Polychronakis. How great is the great firewall? measuring china's DNS censorship. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 3381–3398. USENIX Association, August 2021.

[23] Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2), October 2023.

[24] Jialun Aaron Jiang, Peipei Nie, Jed R. Brubaker, and Casey Fiesler. A trade-off-centered framework of content moderation. *ACM Trans. Comput.-Hum. Interact.*, 30(1), March 2023.

[25] Jingying A Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R Brubaker. Understanding international perceptions of the severity of harmful content online. *PLOS ONE*, 16(8):e0256762, 2021.

[26] Lin Jin, Shuai Hao, Haining Wang, and Chase Cotton. Understanding the impact of encrypted dns on internet censorship. In *Proceedings of the Web Conference 2021*, WWW '21, page 484–495, New York, NY, USA, 2021. Association for Computing Machinery.

[27] Ash Johnson. The path to digital identity in the united states. Report, Information Technology and Innovation Foundation, September 2024. Accessed July 2025.

[28] Dmitry Khovratovich, Ron D. Rothblum, and Lev Soukhanov. How to prove false statements: Practical attacks on fiat-shamir. Cryptology ePrint Archive, Report 2025/118, 2025.

[29] Jan Lauinger, Jens Ernstberger, Andreas Finkenzeller, and Sebastian Steinhorst. Janus: Fast privacy-preserving data provenance for TLS. 2025.

[30] Anja Lehmann et al. Cryptographers' feedback on the eu digital identity's arf. https://www.hpi.de/oldsite/fileadmin/user_upload/fachgebiete/lehmann/cryptographers-feedback.pdf, 2024. Accessed July 2025.

[31] Wouter Lueks, Bogdan Kulynych, Jules Fasquelle, Simon Le Bail-Collet, and Carmela Troncoso. zksk: A library for composable zero-knowledge proofs. In *Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society*, WPES'19, page 50–54, New York, NY, USA, 2019. Association for Computing Machinery.

[32] Jim Miller. Coordinated disclosure of vulnerabilities affecting girault, bulletproofs, and plonk. *Trail of Bits Blog*, April 2022. Accessed July 2025.

[33] David Naylor, Richard Li, Christos Gkantsidis, Thomas Karagiannis, and Peter Steenkiste. And then there were more: Secure communication for more than two parties. 2017.

[34] AP News. Utah lawsuit over online porn age-verification dismissed. *Associated Press*, February 2025. Accessed July 2025.

[35] Melissa Quinn. Supreme court upholds texas law on age verification for porn sites. *CBS News*, June 2025. Accessed July 2025.

[36] Michael Rosenberg, Jacob White, Christina Garman, and Ian Miers. zk-creds: Flexible anonymous credentials from zkSNARKs and existing identity infrastructure. Cryptology ePrint Archive, Paper 2022/878, 2022.

[37] Gabriel Shaw and Xian Zhang. Cyberspace and gay rights in a digital china: Queer documentary filmmaking under state censorship. *China Information*, 32(2):270–292, 2017.

[38] UK Government. Digital identity. https://www.gov.uk/guidance/digital-identity, 2025. Accessed July 2025.

[39] United Nations General Assembly. Convention on the rights of the child. https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-child, 1989. Adopted by General Assembly resolution 44/25 of 20 November 1989. Accessed July 2025.

[40] Aleksandra Urman, Aniko Hannak, and Mykola Makhortykh. User attitudes to content moderation in web search. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1), April 2024.

[41] Zhipeng Wang, Stefanos Chaliasos, Kaihua Qin, Liyi Zhou, Lifeng Gao, Pascal Berrang, Benjamin Livshits, and Arthur Gervais. On how zero-knowledge proof blockchain mixers improve, and worsen user privacy. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2022–2032, New York, NY, USA, 2023. Association for Computing Machinery.

[42] Galen Weld, Amy X. Zhang, and Tim Althoff. What makes online communities 'better'? measuring values, consensus, and conflict across thousands of subreddits. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1121–1132, May 2022.

[43] Xiang Xie, Kang Yang, Xiao Wang, and Yu Yu. Lightweight authentication of web data via garble-then-prove. 2024.

[44] Collin Zhang, Zachary DeStefano, Arasu Arun, Joseph Bonneau, Paul Grubbs, and Michael Walfish. Zombie: Middleboxes that don't snoop. 2024.

[45] Fan Zhang, Ethan Cecchetti, Kyle Croman, Ari Juels, and Elaine Shi. Town crier: An authenticated data feed for smart contracts. 2016.

[46] Fan Zhang, Deepak Maram, Harjasleen Malvai, Steven Goldfeder, and Ari Juels. DECO: Liberating web data using decentralized oracles for TLS. 2020.