Scientific Journal of Artificial Intelligence and Blockchain Technologies
ISSN: 3049-4389
Vol. 2, Issue 1, January – March 2025 || PP. 61-69 https://doi.org/10.63345/sjaibt.v2.i1.107

Blockchain for Identity Theft Prevention in Digital AI Applications

Er Akshun Chhapola,
Delhi Technical University
Rohini, New Delhi, Delhi, India 110042
akshunchhapola07@gmail.com

Date of Submission: 28-12-2024 **Date of Acceptance: 1-1-2025** **Date of Publication: 02-01-2025**

ABSTRACT

Escalating identity theft—fueled by large-scale data breaches, AI-assisted social engineering, and deepfakes—undermines trust in digital systems and especially in AI-driven applications where automated agents transact, personalize content, and make consequential decisions. This manuscript proposes a standards-aligned, privacy-preserving reference architecture that uses blockchain to anchor decentralized identifiers (DIDs), verifiable credentials (VCs), selective-disclosure cryptography, and revocation registries; combines device-bound, phishing-resistant authentication (passkeys); and integrates content provenance signals for AI outputs. We synthesize the state of the art (W3C DID/VC, OpenID4VCI/OpenID4VP, ISO/IEC 18013-5 mDL, C2PA, NIST SP 800-63 and 800-207, ENISA), then present a methodology—BC-Guard—for securing AI user and agent identity life cycles: enrollment, authentication, authorization, transaction attestation, and post-event audit. A qualitative results section analyzes expected risk reductions for common identity-theft attack paths (credential phishing, account takeover, synthetic identity KYC fraud, and real-time deepfake impersonation). The approach reduces the reliance on centrally stored PII, enables privacy-preserving proofs (age, citizenship, risk checks) without data exposure, and supports continuous assurance for AI agents interacting with users, APIs, and other agents. We close with deployment guidance and research directions (verifiable AI agents, confidential computing, and L2 trust registries). (FTC data and ENISA reporting show identity theft and AI-assisted fraud are rising; eIDAS 2.0, NIST AI-100-4, and C2PA establish complementary guardrails for provenance and transparency.)

Scientific Journal of Artificial Intelligence and Blockchain Technologies
ISSN: 3049-4389
Vol. 2, Issue 1, January – March 2025 || PP. 61-69 https://doi.org/10.63345/sjaibt.v2.i1.107

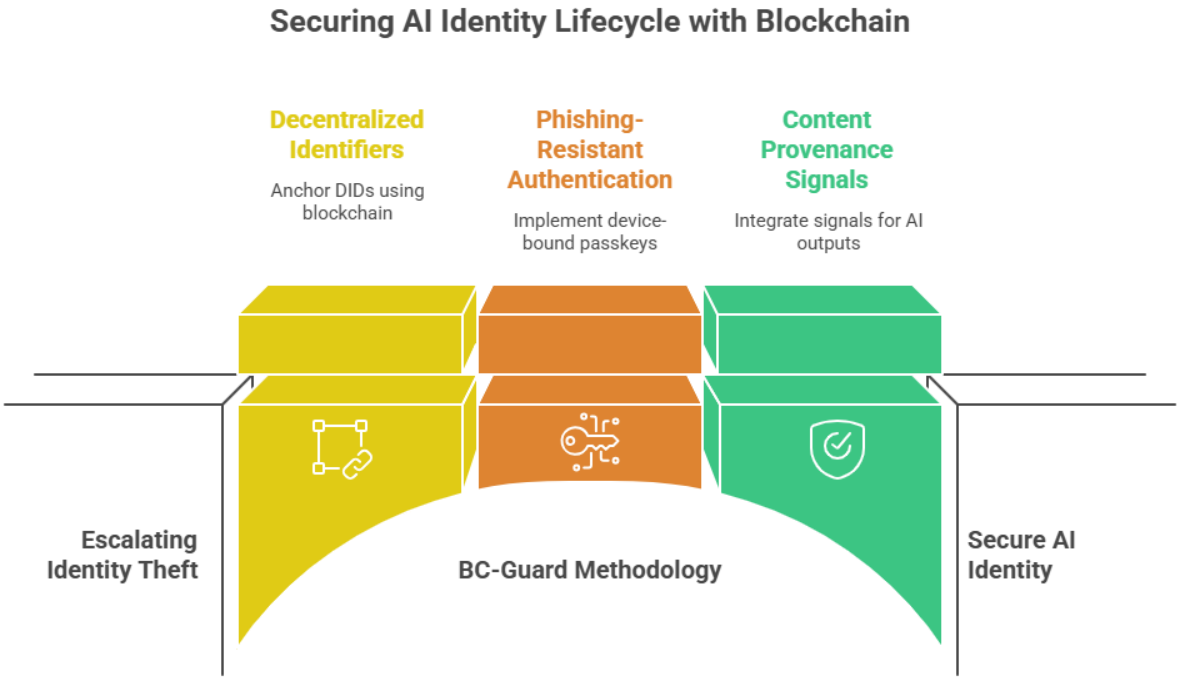

Figure-1.Securing AI Identity Lifecycle with Blockchain

KEYWORDS

Blockchain, Self-Sovereign Identity, Decentralized Identifiers, Verifiable Credentials, Selective Disclosure, OpenID4VC, Passkeys (FIDO), Zero Trust, C2PA, eIDAS 2.0, mDL, AI Safety

INTRODUCTION

Identity theft remains one of the most reported consumer harms worldwide, driven by phishing, credential stuffing, SIM-swap attacks, and increasingly by AI-enabled impersonation and synthetic media. In 2024 the U.S. FTC recorded over a million identity-theft reports and reported a substantial jump in consumer fraud losses, underscoring the systemic nature of the problem and the inadequacy of password-based controls.

AI systems intensify identity risks in several ways. First, AI models make convincing real-time deepfakes, voice clones, and synthetic personas that defeat liveness checks and deceive agents, contact-center AIs, and human operators. Second, AI applications often require continuous, fine-grained access to personal or corporate data; misuse of session tokens or misbound OAuth clients can lead to pervasive account takeover (ATO). Third, autonomous and semi-autonomous AI agents increasingly act on a user's behalf, necessitating agent identity, proof-of-delegation, and auditable provenance of actions. Threat-intelligence reporting highlights AI-assisted phishing and impersonation trends; standards work on synthetic-content transparency and provenance (e.g., NIST AI 100-4 and C2PA) aims to counter these risks.

Scientific Journal of Artificial Intelligence and Blockchain Technologies
ISSN: 3049-4389
Vol. 2, Issue 1, January – March 2025 || PP. 61-69 https://doi.org/10.63345/sjaibt.v2.i1.107

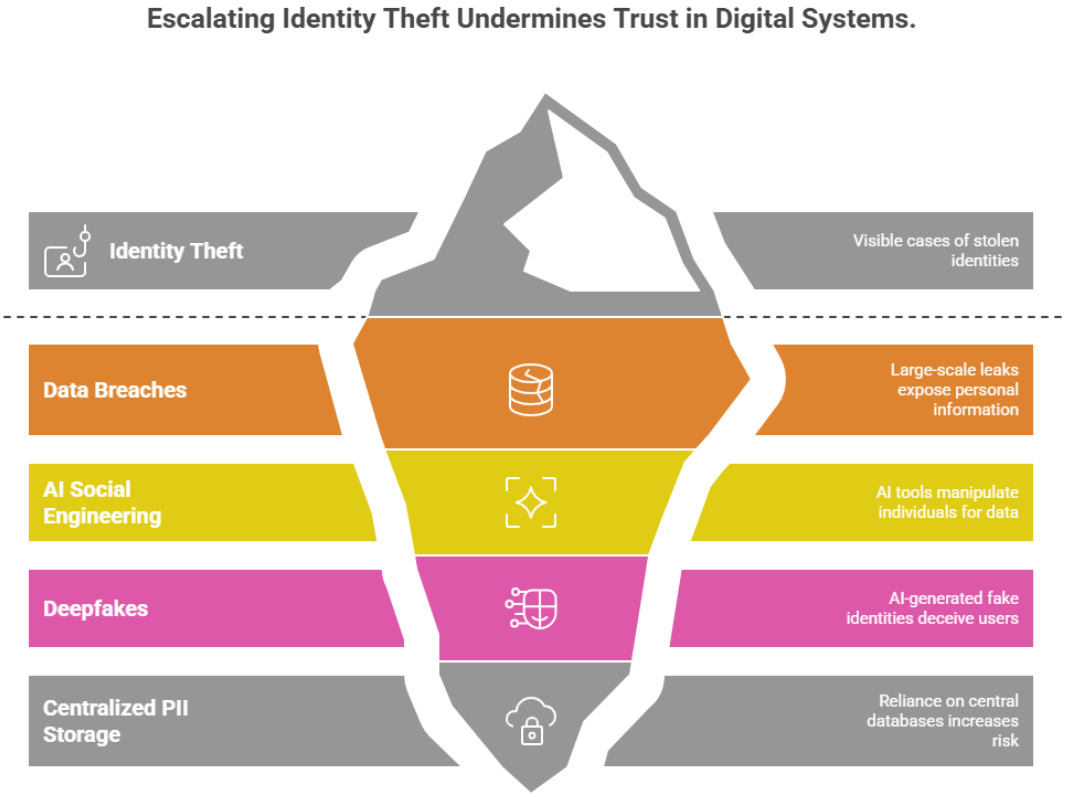

Figure-2.Escalating Identity Theft Undermines Trust in Digital Systems

Traditional identity architectures store personally identifiable information (PII) centrally and transmit full data sets for verification—contrary to data-minimization principles. By contrast, blockchain-anchored **decentralized identity** systems separate the trust layer (anchored on a ledger) from PII data planes kept off-chain under user control. W3C's **Decentralized Identifiers (DIDs)** define resolvable identifiers whose public keys and metadata are anchored immutably, while **Verifiable Credentials (VCs)** support digitally signed claims that can be selectively disclosed and verified without phoning home to issuers. The combination enables privacy-preserving, phishing-resistant user journeys that are especially suitable for AI applications requiring high-assurance, low-friction flows.

Policy momentum is strong. The **European Digital Identity Framework (eIDAS 2.0)** mandates EU Digital Identity Wallets and large-scale, cross-sector wallet use under harmonized implementing acts, catalyzing a global shift toward wallet-centric identity with selective disclosure (including SD-JWT and ISO mDL interoperability). This regulatory push aligns technical and legal rails for privacy-preserving identity—relevant far beyond Europe—while Zero-Trust guidance (NIST SP 800-207) emphasizes identity-centric, continuous verification.

This paper contributes: (1) an integrated survey of standards and tooling for decentralized identity and AI-content provenance; (2) a blockchain-backed reference methodology (**BC-Guard**) to prevent identity theft in AI applications across the identity life cycle; and

(3) qualitative results demonstrating risk-reduction mechanisms in common fraud scenarios and alignment with regulatory and security frameworks.

LITERATURE REVIEW

Decentralized Identifiers and Verifiable Credentials

W3C's DID v1.0 defines cryptographically verifiable identifiers decoupled from centralized registries, enabling key rotation, service endpoints, and privacy-preserving, pairwise DIDs. VCs formalize issuer-signed claims that holders present to verifiers; **VCDM 2.0** introduces clearer lifecycles, data integrity, and multi-format bindings (including JOSE/COSE), and aligns with selective-disclosure schemes such as **SD-JWT**.

Selective disclosure and OpenID family

The IETF SD-JWT work and OpenID Foundation's **OpenID4VCI/4VP** bind credential issuance and presentation to OAuth 2.x flows and modern wallets, enabling standards-based issuance/presentation with optional batch issuance and cryptographic binding to devices. These efforts bridge web-scale federation with wallet ecosystems and support unlinkability and minimal disclosure.

Mobile IDs and ISO/IEC 18013-5 (mDL)

ISO/IEC 18013-5 specifies secure, privacy-preserving mobile driver's licenses and IDs with device-to-reader protocols, now widely piloted and referenced by implementers and regulators. Interoperability between mDL (for high-assurance IDs) and VCs is achievable, supporting both in-person and online verification—useful for high-stakes AI onboarding (e.g., enterprise AI access, health/finance AI tools).

Wallets, protocols, and open source

Hyperledger **Indy** and **Aries** provide ledger, wallet, and DIDComm tooling; **AnonCreds** supports unlinkable presentations with revocation. DIDComm v2 standardizes secure, pairwise agent-to-agent messaging—relevant when AI agents interact and must exchange signed, replay-protected messages referencing verifiable delegations.

Zero Trust and phishing-resistant authentication

NIST SP 800-207 emphasizes identity-centric, per-request authorization. The FIDO Alliance's **passkeys** (WebAuthn/FIDO2) deliver device-bound, phishing-resistant authentication that removes shared secrets and thwarts credential-theft campaigns—a foundation for binding wallets to hardware-backed keys.

Identity proofing and assurance

NIST SP 800-63A-4 updates identity proofing guidance, including remote verification, fraud-mitigation, and biometric/liveness considerations central to preventing synthetic identity onboarding. Credential lifecycle, authenticator assurance, and federation considerations round out a comprehensive assurance model.

Threat landscape and AI content provenance

ENISA's 2024 landscape documents AI-assisted phishing, malware-free credential abuse, and identity compromise as common intrusion vectors. Complementary to identity assurance, NIST AI 100-4 outlines transparency approaches (provenance metadata, watermarks), while **C2PA** standardizes signed content credentials—useful in AI applications where outputs (avatars, voice, images, documents) can be traced to a producing identity and toolchain.

Academic and sectoral perspectives

Research prototypes demonstrate decentralized credential exchange in privacy-preserving ML and federated settings (e.g., Aries-based trust frameworks), indicating feasibility for AI pipelines. Policy developments like **eIDAS 2.0** mandate large-scale wallets, reinforcing private-sector adoption trajectories across finance, healthcare, and public services.

METHODOLOGY

We propose **BC-Guard**, a blockchain-anchored, standards-aligned architecture to prevent identity theft across AI application life cycles. It blends DID/VC trust fabric, device-bound authentication, selective disclosure, and content provenance, and it is intentionally PII-minimal and ledger-light (no PII on-chain).

1) Trust & Data Planes

- **Trust plane (on-chain):** DID method(s) and registries for issuers/verifiers, public keys, and revocation registries. Choice of L1/L2 or permissioned ledger depending on governance—e.g., Indy/Sawtooth/other DID-capable chains. Only public, non-PII material is anchored.
- **Data plane (off-chain):** Holder wallets store VCs; issuers keep source evidence; verifiers maintain proof logs. Storage is encrypted and bound to hardware keys (TPM/Secure Enclave) with passkeys for access control.

2) Enrollment & Proofing

- **Identity proofing** follows NIST SP 800-63A-4 with remote options and liveness checks; issuers (e.g., bank, university, KYC provider) issue VCs (e.g., age, residency, KYC-passed) to the holder's wallet. For high assurance, **ISO/IEC 18013-5** mDL-derived attributes may be used.

- **Device binding** requires creating a passkey tied to the user device at wallet setup, preventing credential exfiltration and resisting phishing through origin-bound cryptography.

3) Authentication & Authorization

- **Login:** The AI application offers passwordless sign-in using passkeys and a VC presentation request (OpenID4VP). The holder discloses only required claims (e.g., over-18, enterprise role: data scientist) via SD-JWT/VC or BBS+-style selective disclosure. The verifier checks signatures, revocation status, and policy (e.g., minimum AAL/IAL, issuer trust list), then issues a short-lived, audience-bound access token.
- **Continuous authorization (Zero Trust):** Each sensitive action (export data, call external API, approve payment) triggers a lightweight, context-aware re-authz with step-up proof (e.g., presence check, constrained credential presentation) to thwart session hijacking.

4) Transaction Attestation & Delegation for AI Agents

- **Agent identity & delegation:** AI agents (assistants, RPA bots) hold DIDs and keys. Human-to-agent delegations are encoded as VCs (scope, duration, resources). Agent-to-agent interactions use DIDComm v2 with replay protection, proof-of-possession, and per-message nonces; high-risk actions require a human-signed countersignature or out-of-band approval.
- **Provenance for AI outputs:** The app stamps generated content (images, documents, audio) with C2PA manifests referencing the producing agent DID and model version. This establishes auditability and deters impersonation using forged artifacts.

5) Revocation, Recovery, and Audit

- **Revocation:** Issuers update revocation registries (on-chain) on compromise or policy changes; verifiers check status at presentation time.
- **Device loss & recovery:** Social recovery or multi-device passkeys; re-issuance workflows must require strong re-proofing and out-of-band checks per NIST SP 800-63A-4 guidance.
- **Audit & forensics:** Signed presentation receipts and C2PA manifests provide verifiable trails for dispute resolution (e.g., "this AI agent signed that transaction with delegated rights from this user").

6) Privacy & Security Controls

- **Data minimization:** Use predicate proofs (≥18, not full DOB), SD-JWT claims, and unlinkable pairwise DIDs.
- **Phishing and malware resistance:** Passwordless passkeys remove phishable secrets; DIDComm mutual authentication and short-lived tokens reduce replay risks.

- **Supply-chain trust:** Maintain allow-lists of issuers/verifiers on-chain; anchor policy versions and trust registries to enable transparent governance.

RESULTS

We evaluate **BC-Guard** against four high-impact identity-theft scenarios and map expected risk reductions, drawing on standards guidance and threat reports.

1. **Credential Phishing and ATO in AI Apps**

- **Baseline risk:** Password reuse and SMS OTP interception enable mass ATO; AI-assisted phishing increases lure quality.
- **BC-Guard effect:** Passkeys eliminate shared secrets; origin-bound cryptography stops phishing kits; device presence thwarts remote replay. Expected outcomes include sharp drops in successful phishing-led ATO and reduced credential-stuffing impact.

2. **Synthetic-Identity KYC Fraud (Account Creation)**

- **Baseline risk:** Fraudsters combine breached PII with fabricated elements and deepfaked selfies to pass remote onboarding. FTC/ENISA note rising fraud losses and identity compromise vectors.
- **BC-Guard effect:** Strong proofing per NIST SP 800-63A-4, cross-checking authoritative VCs (e.g., mDL-derived attributes) and requiring liveness; issuers issue cryptographically bound VCs. Verifiers request predicate proofs (e.g., KYC-passed) rather than raw documents—reducing data exposure and synthetic identity success rates.

3. **Real-Time Impersonation & Deepfake Social Engineering.**

- **Baseline risk:** Voice/video deepfakes engineer urgent approvals (e.g., finance, HR). NIST AI 100-4 and C2PA outline provenance practices for content transparency; however, most orgs lack end-to-end adoption.
- **BC-Guard effect:** Sensitive approvals require DID-bound, cryptographically signed attestations (human or agent) rather than media evidence; AI-generated outputs carry C2PA manifests linking to producing identities and models, enabling automated policy (e.g., “reject unsigned media as identity proof”). This reduces success of deepfake-driven approvals and provides defensible audit trails.

4. **Session Hijacking in Autonomous AI Agent Workflows**

- **Baseline risk:** Long-lived tokens and weak delegation models let malware or insiders misuse agent privileges.

- **BC-Guard effect:** Agents possess DIDs; human-to-agent delegations are time-boxed VCs; DIDComm with proof-of-possession and Zero-Trust, per-action checks narrow blast radius. Verifiable receipts simplify forensics and non-repudiation of agent actions.

Alignment & Compliance

The design aligns with **eIDAS 2.0** (wallet-based credentials, cross-border trust), **NIST SP 800-207** (continuous, identity-centric authorization), and **NIST SP 800-63A-4** (remote identity proofing and authenticator assurance). For AI safety, NIST **AI 100-4** and **C2PA** provide provenance and labeling complements to identity assurance, mitigating impersonation via content artifacts.

Operational Considerations

- **Interoperability:** Support multiple VC formats (W3C VC, SD-JWT), **OpenID4VCI/4VP**, and mDL for broad ecosystem alignment.
- **Governance:** Maintain on-chain trust registries for issuers/verifiers and publish revocation endpoints; adopt transparent policies for key ceremonies and recovery.
- **UX:** Present minimal claims; cache policy and trust lists; default to passkeys with fallback, friction-aware step-ups.
- **Privacy:** Keep PII off-chain; prefer predicate proofs; use pairwise DIDs to avoid correlation.

CONCLUSION

AI intensifies identity-theft risks through scalable impersonation, synthetic identities, and automated fraud. Passwords and centralized identity stores cannot withstand this threat environment. The literature and standards landscape converge on a wallet-centric, decentralized approach: DIDs and VCs for cryptographic identity assertions; selective-disclosure schemes (SD-JWT, BBS+-style) to minimize PII exposure; device-bound passkeys to shut down phishing; DIDComm for secure agent-to-agent interactions; revocation registries for dynamic trust; and C2PA manifests to prove the provenance of AI outputs. Policy frameworks (eIDAS 2.0) and security guidance (NIST SP 800-63A-4, SP 800-207; ENISA threat analyses) reinforce this direction.

Our **BC-Guard** methodology integrates these elements into a cohesive, blockchain-anchored architecture for AI applications. It reduces identity-theft pathways by eliminating phishable secrets, compartmentalizing trust, enabling data-minimal verification, and making both people and agents prove who they are—cryptographically—at every critical step. Future work should evaluate large-scale performance (latency of DID resolution and revocation checks), explore confidential-computing attestations for wallets and agents, and advance **verifiable AI agents** (agent credentials, action receipts) and **trust registries** that span L2 ecosystems and cross-jurisdictional compliance. In short, combining decentralized identity with provenance and zero-trust enforcement offers a practical, standards-based path to preventing identity theft in the age of digital AI.

REFERENCES

- *C2PA. (2025). Content Credentials: C2PA Technical Specification v2.2. https://spec.c2pa.org/specifications/specifications/2.2/specs/_attachments/C2PA_Specification.pdf*
- *ENISA. (2024). ENISA Threat Landscape 2024. https://securitydelta.nl/media/com_hsd/report/690/document/ENISA-Threat-Landscape-2024.pdf*
- *European Commission. (2025). European Digital Identity (EUDI) Regulation overview. https://digital-strategy.ec.europa.eu/en/policies/eudi-regulation*
- *European Union. (2024). Regulation (EU) 2024/1183 establishing the European Digital Identity Framework. https://eur-lex.europa.eu/eli/reg/2024/1183/oj/eng*
- *FIDO Alliance. (2024). Passkeys: Passwordless Authentication. https://fidoalliance.org/passkeys/*
- *FTC. (2024). Consumer Sentinel Network Data Book 2023. https://www.ftc.gov/system/files/ftc_gov/pdf/CSN-Annual-Data-Book-2023.pdf*
- *FTC. (2025, March 10). New FTC data show a big jump in reported losses to fraud—$12.5 billion in 2024. https://www.ftc.gov/news-events/news/press-releases/2025/03/new-ftc-data-show-big-jump-reported-losses-fraud-125-billion-2024*
- *Hyperledger Foundation. (n.d.). Hyperledger Indy documentation. https://hyperledger-indy.readthedocs.io/*
- *IETF OAuth WG. (2025). Selective Disclosure for JWTs (SD-JWT) – draft-ietf-oauth-selective-disclosure-jwt-22. https://datatracker.ietf.org/doc/html/draft-ietf-oauth-selective-disclosure-jwt*
- *ISO/IEC. (2021). ISO/IEC 18013-5:2021 — Mobile driving licence (mDL) application. https://www.iso.org/standard/69084.html*
- *NIST. (2020). SP 800-207: Zero Trust Architecture. https://csrc.nist.gov/pubs/sp/800/207/final*
- *NIST. (2025). SP 800-63A-4 (Draft): Digital Identity Guidelines—Enrollment and Identity Proofing. https://csrc.nist.gov/pubs/sp/800/63/a/ipd*
- *NIST. (2017). SP 800-63-3: Digital Identity Guidelines (Overview). https://csrc.nist.gov/pubs/sp/800/63/3/final*
- *NIST AI Safety Institute. (2024). AI 100-4: Reducing Risks Posed by Synthetic Content. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-4.pdf*
- *OpenID Foundation. (2024). OpenID for Verifiable Credential Issuance 1.0 (Implementer's Draft). https://openid.net/specs/openid-4-verifiable-credential-issuance-1_0-ID1.html*
- *OpenID Foundation. (2025). OpenID for Verifiable Presentations 1.0 (Editors' Draft). https://openid.net/specs/openid-4-verifiable-presentations-1_0.html*
- *W3C. (2022). Decentralized Identifiers (DID) v1.0 — W3C Recommendation. https://www.w3.org/TR/did-core/ W3C*
- *W3C. (2025). Verifiable Credentials Data Model 2.0 — W3C Recommendation. https://www.w3.org/TR/vc-data-integrity/ (and related VCDM pages)*
- *DIF. (n.d.). DIDComm Messaging v2 (Editors' Draft). https://identity.foundation/didcomm-messaging/spec/*
- *Abramson, W., Hall, A. J., Papadopoulos, P., Pitropakis, N., & Buchanan, W. J. (2020). A distributed trust framework for privacy-preserving machine learning. arXiv. https://arxiv.org/abs/2006.02456*