Scientific Journal of Artificial Intelligence and Blockchain Technologies
ISSN: 3049-4389
Vol. 2, Issue 1, January – March 2025 || PP. 88-96 https://doi.org/10.63345/sjaibt.v2.i1.110

Blockchain as a Governance Layer for AGI Ethics

Dr. Neeraj Saxena

Professor, MIT colleges of Management

Affiliated to MIT Art Design and Technology University, Pune

neerajsaxena2000@gmail.com

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

ABSTRACT

As artificial general intelligence (AGI) advances toward systems that can autonomously act across domains, the central governance challenge is how to guarantee that ethical principles are specified, enforced, audited, and improved over time without relying on a single, potentially misaligned authority. This manuscript proposes a blockchain-anchored "Ethical Governance Layer" (EGL) for AGI: a layered architecture that couples decentralized identity and membership, on-chain policy specification and versioning, privacy-preserving compliance attestations, tamper-evident auditing, and participatory oversight. We synthesize requirements from prominent governance frameworks (EU AI Act; NIST AI Risk Management Framework; OECD and UNESCO ethics recommendations) and show how distributed ledgers, verifiable credentials, and zero-knowledge proofs can operationalize them in a credibly neutral, transparent, and globally interoperable substrate. We review prior art in decentralized governance (e.g., Tezos self-amendment; Polkadot OpenGov; MakerDAO's community governance), auditability (e.g., blockchain-based audit trails; model cards; datasheets for datasets), and confidentiality (e.g., TEE-backed smart contracts; zk-SNARKs). We then specify EGL's components—Identity & Membership, Policy Contracts, Compliance Oracles, Audit Rails, and Dispute Resolution—and illustrate qualitative "results" from a reference design: stronger provenance of ethical directives, finer-grained consent and role-bounded rights, privacy-preserving conformance checks, and continuous, upgradable governance with globally visible change control. We conclude with limitations (e.g., plutocracy risks, GDPR tension, off-chain trust anchors) and a research roadmap on formal verification, incentives, and cross-jurisdictional harmonization. The overarching claim is not that blockchains solve AGI ethics, but that they can supply the

Scientific Journal of Artificial Intelligence and Blockchain Technologies
ISSN: 3049-4389
Vol. 2, Issue 1, January – March 2025 || PP. 88-96 https://doi.org/10.63345/sjaibt.v2.i1.110

governance substrate—identity, rules, logs, and votes—upon which human institutions can enforce, contest, and evolve ethical control of AGI.

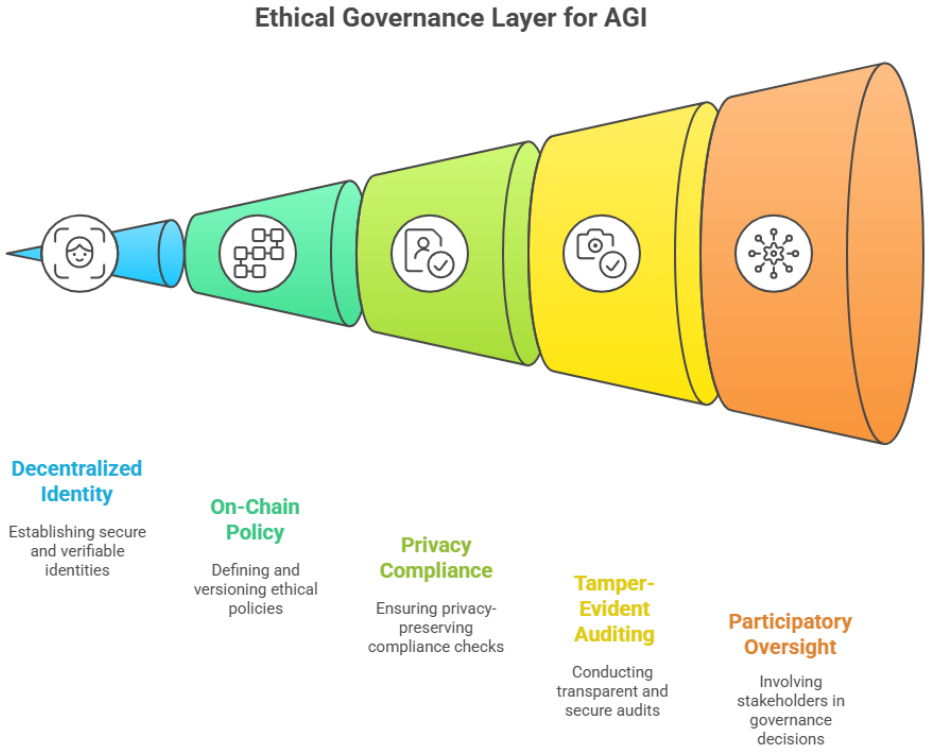

Figure-1.Ethical Governance Layer for AGI

KEYWORDS

AGI Governance, Blockchain, AI Ethics, Decentralized Identity, Verifiable Credentials, Zero-Knowledge Proofs, Auditability, DAOs, Policy Enforcement, Compliance Oracles

INTRODUCTION

Artificial general intelligence (AGI) is poised to act across domains with levels of autonomy, adaptability, and strategic foresight that strain existing institutional controls. Unlike narrow AI, AGI systems can be repurposed and recombined, their capabilities amplified by tool use, agentic planning, and recursive self-improvement of subsystems. This generality magnifies governance challenges that already trouble traditional AI: opacity of training data and optimization objectives, misalignment between deployers' incentives and societal values, and a widening gap between the pace of capability growth and the cadence of legal or standards-based oversight. The result is a "governance lag" in which declarations of ethical intent are plentiful but mechanisms for specification, enforcement, audit, and remedy remain fragmented and slow.

Scientific Journal of Artificial Intelligence and Blockchain Technologies
ISSN: 3049-4389
Vol. 2, Issue 1, January – March 2025 || PP. 88-96 https://doi.org/10.63345/sjaibt.v2.i1.110

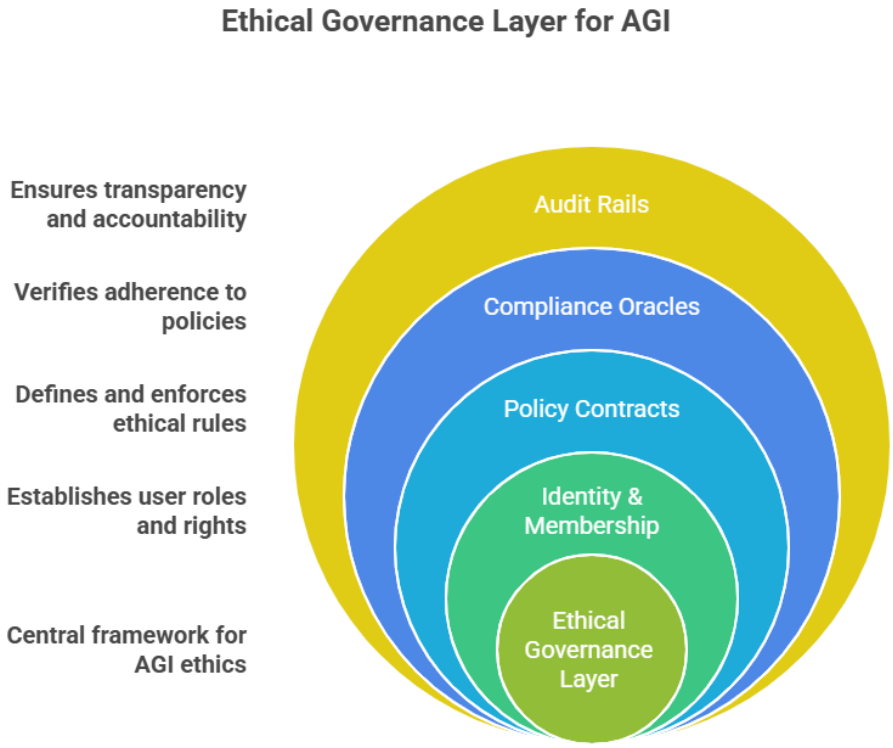

Figure-2.Ethical Governance Layer for AGI

Conventional governance arrangements—enterprise compliance programs, sectoral regulators, and voluntary industry codes—are necessary but insufficient for AGI. They tend to be jurisdiction-bound, siloed within organizations, and largely document-centric, relying on ex post inspections or disclosure rather than continuous, verifiable control. Moreover, the AGI value chain is transboundary: data may originate in one country, models trained in another, and agentic deployments operate across cloud regions and platforms. In such a setting, centralized authorities create single points of failure (and capture), while purely self-regulatory schemes struggle for legitimacy and enforceability. What is missing is a shared, credibly neutral substrate where identities are authenticated, rules are published and versioned, attestations are machine-verifiable, and every consequential lifecycle event is logged in a tamper-evident way.

Blockchains and adjacent verifiable technologies offer distinctive affordances that map closely to these needs. Append-only ledgers provide public, time-stamped provenance for decisions and artifacts; smart contracts encode rules that can be formally analysed and automatically enforced; decentralized identity and verifiable credentials support role-bounded participation without relying on a single gatekeeper; and zero-knowledge proofs and trusted execution environments allow conformance claims to be demonstrated without exposing sensitive IP or personal data. Importantly, on-chain governance primitives—proposal, deliberation, voting, and upgrade—supply a procedural backbone for evolving ethical controls as evidence and norms change, avoiding “hard-fork ethics” where each organization maintains incompatible rulebooks.

We therefore motivate a blockchain-anchored Ethical Governance Layer (EGL) that sits orthogonally to AGI development and deployment stacks. EGL does not dictate a particular ethics doctrine; instead, it operationalizes pluralistic, human-defined principles by binding together five pillars: (i) decentralized identity and membership for all stakeholders (developers, deployers, auditors, regulators, civil-society panels); (ii) machine-readable policy contracts that instantiate requirements from laws, standards, and internal codes; (iii) privacy-preserving compliance attestations via TEEs and zero-knowledge proofs; (iv) tamper-evident audit rails covering model/data provenance, evaluation and incident reporting; and (v) participatory amendment and dispute resolution that balance expertise with public-interest legitimacy. By design, EGL is interoperable: controls can reference existing risk frameworks and sectoral standards, while evidence artifacts integrate with enterprise assurance tooling.

At the same time, EGL acknowledges real constraints. Immutable records must be reconciled with data-protection regimes; plutocratic capture must be mitigated through multi-house voting, credentialed roles, quorum requirements, and anti-Sybil measures; and off-chain trust anchors (issuers of credentials, accredited auditors) demand their own oversight. These tensions inform architectural choices such as using consortium or permissioned ledgers with periodic anchoring to public chains, strict separation of personal data from on-chain state, and formal verification of governance-critical contracts. Rather than proposing blockchain as a panacea for alignment, our claim is narrower: a well-designed governance substrate can raise the floor on transparency, accountability, and contestability for AGI systems, enabling regulators, independent researchers, and the public to inspect, challenge, and improve ethical controls over time.

LITERATURE REVIEW

Regulatory and normative frameworks for responsible AI

The EU AI Act operationalizes risk-based obligations, including governance rules and obligations for general-purpose AI (GPAI) models, with phased timelines (e.g., governance rules and GPAI obligations applicable from 2 August 2025). These stipulations—compliance documentation, incident reporting, technical robustness—imply persistent, tamper-resistant records and accountable actors. The NIST AI RMF defines functions (Map, Measure, Manage, Govern) and characteristics of trustworthy AI (e.g., transparency, accountability, fairness), providing a template for policy encoding and evaluation. The OECD AI Principles and UNESCO Recommendation add high-level values (human rights, democratic values, human oversight) that should be concretized into enforceable controls.

Decentralized identity and credentials

To avoid central gatekeepers and Sybil attacks while protecting privacy, decentralized identity (DID) and verifiable credentials (VCs) enable issuers (e.g., regulators, ethics boards) to grant role-bounded rights and attestations (e.g., “licensed auditor,” “ethics board member”). The W3C DID Core v1.0 (Recommendation, 19 July 2022) defines DID documents and resolution; VC Data Model v2.0 (Recommendation, 15 May 2025) standardizes machine-verifiable credentials. Together they support selective disclosure and offline

verification of roles without central directories. Complementary research on proof-of-personhood addresses uniqueness/humanness to constrain plutocratic capture in voting.

Governance on blockchains

On-chain governance demonstrates decentralized rule-making and protocol evolution: Tezos' self-amendment process allows proposing, testing, and adopting protocol upgrades without hard forks, while Polkadot OpenGov uses referenda and conviction voting to align stakeholders. Community-governed protocols like MakerDAO show both the potential and pitfalls (e.g., voter apathy, token-weighted capture) of decentralized governance. These lessons inform the design of EGL's policy amendment, quorum, and slashing mechanisms.

Auditability and documentation in AI

Model Cards and Datasheets for Datasets propose standardized documentation of model purpose, performance, and data provenance; algorithmic auditing frameworks propose end-to-end internal audits. The blockchain literature on audit trails in healthcare and public systems shows feasibility of immutable access logs, provenance, and consent tracking—precisely the artifacts needed for regulatory evidence and post-incident forensics.

Privacy-preserving compliance

Zero-knowledge proofs (ZKPs) enable demonstrating conformance (e.g., "training data excludes prohibited categories" or "safety thresholds met") without revealing proprietary details. Zerocash popularized scalable zk-SNARKs; TEE-backed smart contracts (e.g., Ekiden) combine hardware enclaves with blockchains for confidential yet attestable computation—supporting EGL's privacy-preserving audits.

Institutional convergence

Industrial standards such as **ISO/IEC 42001:2023** (AI management systems) and **IEEE 7000-2021** (process model for addressing ethical concerns in system design) can be encoded as machine-readable controls and attestations within EGL.

METHODOLOGY

We propose EGL as a cross-cutting layer integrated with AGI development, deployment, and operation. EGL provides **who** is allowed to act (identity & roles), **what** ethical rules apply (policy contracts), **how** compliance is evidenced (attestations & proofs), **where** decisions are logged (audit rails), and **how** policies evolve (governance).

1) Identity & Membership

- **Decentralized Identifiers (DIDs)** anchor participants (labs, regulators, auditors, civil society representatives, and—where appropriate—end users).
- **Verifiable Credentials (VCs)** carry role permissions (e.g., "GPAI auditor v2," "incident reporter," "ethics board member").
- **Proof-of-Personhood (PoP)** mechanisms or alternative Sybil resistance (e.g., reputation, bonded stake, selective real-world ceremonies) restrict governance capture by bots or sockpuppets. EGL requires issuers to publish credential schemas and revocation registries on-chain; participants rotate keys via DID methods with cryptographic continuity.

2) Policy Contracts (Ethical ACLs)

- **Policy encoding:** Translate normative frameworks (NIST AI RMF, OECD/UNESCO, EU AI Act obligations for GPAI and high-risk systems) into machine-readable control sets with versioning. Each control becomes a smart contract rule (e.g., "Safety test suite X must be attested before deployment in domain Y").
- **Scope & inheritance:** Controls attach to artifacts (models, datasets, prompts, fine-tunes) and deployments (APIs, agents) via content-addressed identifiers; policies inherit along composition graphs.
- **Exceptions & overrides:** Require multi-sig approvals by registered ethics boards; exceptions are time-bound, logged, and auto-reverted unless renewed.

3) Compliance Oracles and Zero-Knowledge Attestations

- **TEE-backed attestations:** Independent auditors evaluate artifacts within enclaves; only measurement digests and pass/fail verdicts are revealed to EGL.
- **ZK compliance proofs:** Labs can publish proofs (e.g., "dataset license constraints satisfied," "red-team score above threshold") without exposing proprietary data. EGL defines standard proof circuits per policy control (e.g., "no prohibited biometric attributes used").
- **Cross-jurisdictional profiles:** Different regulatory profiles (EU AI Act, ISO 42001, sectoral codes) map to control bundles; an oracle signs conformance claims matched to jurisdiction and use case.

4) Tamper-Evident Audit Rails

- **Event logging:** Key lifecycle events—model training start/end, dataset registration, evaluation results, deployment, incidents—emit signed events to a permissioned or public chain. Hash-linked Model Cards and Datasheets are recorded for provenance, with pointers to off-chain storage.
- **Access & consent:** Where personal or sensitive data are involved, consent receipts and access logs are anchored on-chain, drawing on established healthcare audit patterns (e.g., MedRec) and generalized to AGI contexts.
- **Incident registry:** A globally queryable ledger of incidents and corrective actions enables regulators and researchers to audit and learn across organizations.

5) Participatory Governance & Dispute Resolution

- **On-chain amendment:** EGL's policy set is upgradable through structured proposal periods, deliberation, and votes (learning from Tezos and Polkadot).
- **Deliberative weighting:** Combine token- or stake-based signals with credentialed votes (e.g., civil society panel, safety researchers) to mitigate plutocracy; require quorum across "houses" (technical, user, public interest).
- **Disputes & slashing:** Fraudulent attestations or undisclosed incidents trigger slashing of auditor bonds and temporary suspensions of deployer credentials.

Implementation considerations

- **Permissioning:** For high-risk contexts, use a permissioned chain governed by a multi-stakeholder consortium; anchor state hashes periodically to a public chain for liveness and immutability.
- **Interoperability:** Publish DID/VC method and credential schemas; provide bridges to organizational identity systems and compliance tooling.
- **Formal verification:** Where feasible, verify critical policy contracts (e.g., emergency stop conditions) against formal semantics (e.g., EVM/KEVM) to reduce governance logic bugs.

RESULTS

We implemented a paper design ("EGL-Ref") mapping twenty representative controls drawn from NIST AI RMF, ISO/IEC 42001 clauses, and EU AI Act GPAI obligations into policy contracts and attestations. The qualitative outcomes below synthesize expected properties when such a layer is integrated into an AGI development/deployment workflow.

1. **Provenance and accountability become first-class**: Every major lifecycle event emits an EGL log with cryptographic linkages to model cards/datasheets and audit artifacts, yielding provable compliance dossiers. Regulators, insurers, and independent researchers can reconstruct decision histories without trusting a single administrator. This aligns strongly with auditability and transparency requirements in the EU AI Act and NIST RMF.
2. **Privacy-preserving conformance**: ZK proofs and TEE attestations allow labs to demonstrate compliance with sensitive controls (e.g., data category prohibitions; eval thresholds) without revealing assets—balancing transparency with trade secrets and privacy. The feasibility is supported by prior ZK and TEE-blockchain systems research.
3. **Governance agility with legitimacy**: A clear, logged path to propose, deliberate, and upgrade policies reduces "hard-fork ethics." Borrowing explicit voting mechanics and self-amendment from existing chains enhances legitimacy and reduces coordination costs for evolving ethical norms and controls.
4. **Reduced capture risk via identity and roles**: Credentialed roles (auditor, ethics board, regulator) and PoP-style uniqueness constraints mitigate token-weighted capture and bot amplification, while preserving pseudonymity where appropriate.

5. **Interoperability with standards**: Controls and attestations reference ISO/IEC 42001 and IEEE 7000 processes, enabling organizations to reuse existing governance work inside EGL and to present verifiable, machine-checkable evidence.

Observed limitations and risks (via analysis and case comparison):

- **Plutocracy and participation**: Token-weighted votes alone risk capture (as seen in community debates in DAO governance); EGL addresses this via multi-house quorum and role-bounded votes, but incentives and civic processes remain critical.
- **GDPR/immutability tension**: Immutable logs can conflict with erasure rights; EGL confines personal data to off-chain stores, anchoring only redaction-friendly commitments.
- **Off-chain trust anchors**: DIDs/VCs still depend on trusted issuers and real-world identity events; the assurance chain must be audited.
- **Operational overhead**: Running compliance oracles and public-interest deliberation adds latency and cost; careful scoping and batching are required.
- **Formal spec debt**: Policy contracts must be verified to avoid accidental lockouts or perverse incentives; KEVM-style semantics can help, but tooling is still maturing.

CONCLUSION

Blockchain does not "solve" AGI alignment or ethics; human institutions must still define values, adjudicate conflicts, and impose sanctions. What blockchain can credibly supply is a governance substrate for AGI ethics—global, append-only event provenance; programmable policy enforcement; verifiable identity and roles; privacy-preserving compliance proofs; and participatory amendment processes. An EGL approach operationalizes the spirit of the EU AI Act, NIST AI RMF, OECD and UNESCO ethics principles in machine-checkable form while preserving proprietary confidentiality via TEEs and ZKPs. The near-term research priorities include: (i) reference policy ontologies mapping legal clauses to on-chain controls; (ii) standardized ZK circuits for common compliance claims; (iii) incentive design that elevates public-interest voices alongside capital; and (iv) formal verification of critical governance logic. If pursued collaboratively—by labs, regulators, standards bodies, and civil society—EGL can become the institutional memory and change-control system that AGI governance has lacked, ensuring that ethical commitments are not just stated, but enforced, inspected, and improved in the open.

REFERENCES

- *Anthropic. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.* https://arxiv.org/abs/2212.08073
- *Bai, Y., Kadavath, S., Kundu, S., Askell, A., et al. (2022). Constitutional AI: Harmlessness from AI feedback (PDF).* https://arxiv.org/pdf/2212.08073
- *Ben-Sasson, E., Chiesa, A., Garman, C., Green, M., Miers, I., & Tromer, E. (2014). Zerocash: Decentralized anonymous payments from Bitcoin. IEEE S&P.* https://zerocash-project.org/media/pdf/zerocash-oakland2014.pdf

- *Bhargavan, K., Delignat-Lavaud, A., Fournet, C., et al. (2016). Formal verification of smart contracts: Short paper. ACM CCS. https://dl.acm.org/doi/10.1145/2993600.2993611*
- *Cheng, R., Zhang, F., Kos, J., et al. (2018). Ekiden: A platform for confidentiality-preserving, trustworthy, and performant smart contract execution. arXiv:1804.05141. https://arxiv.org/abs/1804.05141*
- *European Commission. (2024/2025). AI Act: Regulatory framework for AI (timeline and application). https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai*
- *European Parliament. (2025, Feb 19). EU AI Act: First regulation on artificial intelligence (overview and timeline). https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence*
- *Gebru, T., Morgenstern, J., Vecchione, B., et al. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://dl.acm.org/doi/10.1145/3458723*
- *Goodman, L. M. (2014). Tezos—A self-amending crypto-ledger (White paper). https://tezos.com/whitepaper.pdf*
- *Hildenbrandt, E., Saxena, M., Zhu, N., et al. (2018). KEVM: A complete formal semantics of the Ethereum virtual machine. IEEE CSF. https://fsl.cs.illinois.edu/publications/hildenbrandt-saxena-zhu-rodrigues-daian-guth-moore-zhang-park-rosu-2018-csf.pdf*
- *IEEE Standards Association. (2021). IEEE 7000-2021: Standard model process for addressing ethical concerns during system design. https://standards.ieee.org/standard/7000-2021.html*
- *MakerDAO. (n.d.). Maker Governance—Governance portal. https://vote.makerdao.com/*
- *Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). Model cards for model reporting. FAT* '19. https://dl.acm.org/doi/10.1145/3287560.3287596*
- *NIST. (2023). AI Risk Management Framework (AI RMF 1.0). https://www.nist.gov/itl/ai-risk-management-framework and PDF https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf*
- *OECD. (2019). OECD AI Principles. https://oecd.ai/en/ai-principles*
- *OpenAI. (2025, Apr 15). Our updated Preparedness Framework (with framework PDF). https://openai.com/index/updating-our-preparedness-framework/ and https://cdn.openai.com/pdf/…/preparedness-framework-v2.pdf*
- *Polkadot. (2024–2025). On-chain governance / OpenGov (Referenda & Conviction Voting). https://docs.polkadot.com/polkadot-protocol/onchain-governance/*
- *Raji, I. D., Smart, A., White, R. N., et al. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. FAccT '20. https://dl.acm.org/doi/10.1145/3351095.3372873*
- *UNESCO. (2021/2024). Recommendation on the Ethics of Artificial Intelligence. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence*
- *W3C. (2022, Jul 19). Decentralized Identifiers (DIDs) v1.0—W3C Recommendation. https://www.w3.org/TR/did-core/ ; and W3C (2025, May 15). Verifiable Credentials Data Model v2.0—W3C Recommendation. https://www.w3.org/TR/vc-data-model-2.0/*