GBBC
Global Blockchain
Business Council

AI CONVERGENCE REPORT

GLOBAL STANDARDS MAPPING INITIATIVE 6.0

TOWARD DECENTRALIZED & OPEN-SOURCE AI: TRANSPARENCY, PRIVACY, SECURITY, AND RELIABILITY

GBBC GSMI 6.0

ACKNOWLEDGEMENTS

Diana Oreto (Barrero Zalles)
Head of GSMI & Research, GBBC

John deVadoss - CHAIR
Co-Founder, Providentia Capital

Thank you to our working group participants and review committee for your inputs.

GLOBAL BLOCKCHAIN BUSINESS COUNCIL

DC Location:
1629 K St. NW, Suite 300
Washington, DC 20006

Geneva Location:
Rue de Lyon 42B
1203 Geneva
Switzerland

EXECUTIVE SUMMARY

AI deployment has moved from pilot to production – across content, decision support, and domain copilots. The upside is material; so are the risks. In the face of this trend, and the opportunities and risks it represents, this paper takes a position that is pragmatic: evaluating the use of blockchain strategically to make AI verifiable. For instance, blockchain technology can facilitate provenance of data and models, event-driven audit, and machine-readable attestations that third parties can test without exposing sensitive content. Building on prior GSMI reports on Blockchain & AI Convergence,[1] which established foundational topics on how blockchain can enhance AI deployments, this report comments specifically on recent developments toward decentralized and open-source AI, distinguishing hype from reality in light of existing challenges toward trustworthy AI, and ultimately offering an approach toward AI transparency, privacy, security, and reliability that is more attainable in the short term.

Better solutions & worse problems? AI can amplify value and, just as easily, scale existing bias and perpetuate system societal challenges. Blockchain technology can advance provenance and auditability, as levers to guide toward better solutions. Putting the AI explosion into perspective, consumer tools have reset expectations, and enterprises are weaving copilots into core workflows. Yet velocity has outpaced assurance. The goal is not to rush into maximal function in ways that can increase compliance issues, governance failures, and other risks; instead, the goal is to ensure responsible functions that can be trusted in the long term — with provable inputs, accountable execution, and verifiable outcomes.

The goals of AI transparency, privacy, security, and reliability are aligned with privacy goals, emerging governance scaffolds including standards and regulatory developments (e.g., ISO/IEC 42001, NIST AI, RMF, GenAI Profile, EU AI Act GPAI/Code), and illustrated through operating models that enterprises can adopt now.

FAST AI ADOPTION, FASTER THAN ASSURANCE

Increasingly in the recent months and years, AI adoption has accelerated across all firm sizes, from startups to major enterprises, and in all economic sectors, with tools ranging from content generation and decision support to domain-specific copilots embedded in enterprise workflows. In the EU, for instance, over 40% of large enterprises used AI in 2024.[2] This growth is not merely experimental; it is tied to revenue ambitions, cost containment, and competitive differentiation. Yet the velocity of deployment has outpaced the maturation of AI assurance, with much to be developed in areas like robust governance, run-time observability, auditable provenance, and defensible accountability. Leaders face three recurring questions:

- **What data went in and how was it processed to reach an outcome?**
- **What exactly happened at inference time?**
- **Who is accountable when outcomes drift or other unintended functions cause harm?**

At the same time, the regulatory and standards landscape is continuing to evolve around aspects like lifecycle governance—specifying what to document, what to prove, and how to verify it. Notably, ISO/IEC 42001 introduces a certifiable AI management system framework[3]; the NIST AI Risk Management Framework[4] and Generative AI Profile[5] translate risk concepts into operational controls; and EU obligations for general-purpose AI (GPAI)[6] begin phasing in, supported by a GPAI Code of Practice[7] with obligations and a Code of Practice.

Ultimately, in order for AI to scale business outcomes responsibly, assurance can no longer an ethics add-on; it becomes an operational requirement. The case for blockchain in AI is therefore not ideological but instrumental. To make claims verifiable—provenance of data and models, tamper-evident audit trails, and machine-readable attestations that external parties can test without exposure of sensitive content. In other words: transparency, privacy, security, and reliability. As many of the most tangible AI & Blockchain Convergence wins still flow through use cases and foundation models, this paper's emphasis is on the assurance fabric that makes those wins sustainable.

THE IDEAL OF OPEN-SOURCE AND DECENTRALIZED AI

In late 2004 and early 2005, DeepSeek took off as an open-weight AI company with a revolutionary approach where anyone can modify and deploy models. Worth noting is that while DeepSeek quickly brought attention to open-source AI, its own model is technically open-weight, not fully open-source. For readability purposes, the rest of this document will mean "open-weights" when referring to "open-source."

Releasing a series of powerful AI models that were significantly cheaper to train and run than existing competitors, DeepSeek sent shockwaves across the AI market, having found a way to optimize algorithms and hardware to enabling high performance at a fraction of the cost of existing models out of Silicon Valley. The draw of DeepSeek's open-source approach goes hand in hand with its decentralized AI capability where anyone can download, modify, and deploy AI models without centralized oversight. This characteristic is unlike that of proprietary, SaaS-based models and opens access to AI tools at a broader level, democratizing access to advanced AI tools for developers and researchers around the world. DeepSeek therefore allows community-driven deployment and usage without the need for a single controlling entity, and often relying on decentralized storage networks to reduce dependence on centralized servers. It is worth noting, however, that only DeepSeek could create newer versions of models, having the source code for pre-training and training, as well as the lineage of data. Other participants could 'distill' DeepSeek (e.g., fine-tune, RAG, etc).

Within policy and technical communities, open-source AI and decentralized AI are often presented as remedies to opacity and concentration concerns, leading toward greater reliability. The ideal is attractive: **models whose code and weights are inspectable; datasets with declared provenance and licenses; contributions recognized via verifiable credentials; and execution spread across diverse nodes to avoid single points of failure**.

In this vision, AI becomes more explainable, reproducible, and inclusive, with community governance mitigating monopolistic control. Moreover, privacy can be preserved with self-hosting solutions and users' control over their data use and processing.

Decentralization extends the ideal by distributing compute, storage, and decision rights, aligning incentives through tokenized or reputation-based systems, and enabling federated and privacy-preserving training that respects data sovereignty. Together, open and decentralized approaches promise to widen participation, increase resilience, and ground model behavior in provable contributions. New business models can arise from decentralized AI marketplaces, open models, and cooperative ownership structures.

Open-source and decentralized AI models can reveal exactly how models were trained and where the data was sourced, providing full visibility and certainty rather than a “black box” of unknown data provenance and processing leading to a result. In “black box” scenarios, it is difficult to understand why models may hallucinate or perpetuate biases, and leaving these concerns unaddressed can risk the compounding of even more problematic behavior with each new version of a model. Without accurate data provenance, it is impossible to predict how much AI can deviate in different directions from the expected outcome if access to data is not sourced in a quality manner.

If open source and decentralized AI can allow concerns to be more effectively identified, there can also be mechanisms of economic incentives (e.g., credits, points) to reward participants globally for contributing to address these concerns, which can lead to greater clarity on who and how sources come from. Finaly, more open-source and decentralized approaches to AI will allow a broader view of lived experiences globally that is beneficial to all – from embracing transparency and accountability for training data, to provenance of data sets.

In short, open-source AI could level the playing field by democratizing access to technology solutions, in ways that lower the entropy of coordination, while decentralized AI spreads rights and responsibilities. A primary benefit of open-source AI may be that it is the optimal ethos, and best cultural and technical fit for decentralized AI. This combination increases traceability, which enterprises need to trust and scale AI.

BENEFITS OF OPEN-SOURCE AI

Open-source models provide an alternative to major concerns around AI being controlled with monopoly power. Open-source AI is attractive for the following characteristics:

TRANSPARENCY & INSPECTABILITY

Public code, model cards, and—where feasible—training declarations allow independent scrutiny, in ways that reduce information asymmetries and increase verifiability. Anyone can inspect a model's code, data, and model architecture. This can reduce risks of hidden biases, unethical practices, or backdoors.

REPRODUCIBILITY & SUPPLY-CHAIN INTEGRITY

When artifacts (e.g., weights, tokenizers, loaders) are signed, hashed, and anchored, participants can independently rebuild and replay pipelines. This can enhance education and research, as an excellent resource for learning and experimentation that encourages academic collaboration (e.g., artifact signing solutions standard and tooling).

CUSTOMIZABILITY & SELF-HOSTING

Organizations can tailor, fine-tune, and deploy locally to preserve confidentiality and enforce policy. Developers can tailor models for specific tasks or environments. Customizability also facilitates experimenting with modifications and fine-tuning.

DEMOCRATIZATION

Democratization: Lower barriers to entry can facilitate free access to individuals, startups, and researchers, especially in the Global South and traditionally underrepresented populations, seeking to build and experiment with AI solutions, without the need for substantial budgets.

COLLABORATION

Open contributions from community-driven innovation foster rapid developments, where bugs or other concerns may be found and fixed more quickly by a global community.

ATTRIBUTION & INCENTIVES

Verifiable credentials and on-chain attestations enable granular credit for data, compute, and evaluation—foundations for sustainable contribution markets.

BENEFITS OF DECENTRALIZING AI

As an ideal, many of the benefits of decentralized AI overlap with the benefits of open-source AI, which yet again supports the fact that these two attributes are a natural fit to be deployed together. The attributes and value added by decentralization in AI can be summarized as the following:

SECURITY & PRIVACY

By avoiding central data lakes, decentralized approaches reduce single-point compromise risk, as there is no central database to hack or manipulate. In addition, rather than centralized platforms owning user data, individuals and organizations can better control their own data in a model that provides greater data sovereignty. With confidential computation, sensitive data may be used for training purposes or inference, without being exposed. This enables privacy-enhancing computation (e.g., TEEs, MPC, ZK proof systems) that performs learning or inference without exposing raw data.

FAIRNESS & INCLUSIVITY

Broader participation—both in data contribution and in governance—can improve representativeness and reduce geographic or cultural bias. Diversity in nodes and datasets expands the model's "lived experience" beyond a handful of players or synthetic data which can become unrealistic over iterations of reprocessing. Diverse data sources, by including data from many nodes and communities, also improves representation and reduces bias. With democratized access, any participant can also contribute to, train, and utilize AI models without the need to rely on a few large tech players. Moreover, models and updates can be managed collectively, which leads to reduced monopoly power.

RESILIENCE & RELIABILITY

Distributed topologies mitigate outages and censorship, while consensus-anchored events provide tamper-evident histories of model changes and policy updates. Decentralization creates an environment of no single point of failure, where even if some nodes may fail or go offline, continued AI services remain. This also results in censorship resistance, where it becomes more difficult for a single entity (e.g., government or corporation) to shut down or censor AI models.

INNOVATION & EFFICIENCY

Shared compute marketplaces and cooperative training enable resource pooling and resource sharing, where idle compute power can be pooled together (e.g., edge devices, GPUs). Fine-tuning can propagate through verifiable release channels that preserve lineage and licensing, in a model of collaborative building involving several parties contributing to building models. In this context, training can become scalable, with federated learning and swarm intelligence training models across nodes distributed globally.

ECONOMIC ALIGNMENT ACROSS THE ECOSYSTEM

Tokenized and credential-based systems reward data providers, validators and nodes, red-teamers, and evaluators, creating a more complete AI supply chain with accountable roles. Shared infrastructure also reduces dependence on otherwise expensive compute (e.g., centralized cloud services).

REALITY CHECK: NEITHER OPEN-SOURCE NOR DECENTRALIZED AI ARE A REALITY AT SCALE

While AI delivered in an open-source and decentralized format has much promise, this has yet to materialize. As both consumer-facing and enterprise-focused AI solutions are expected to alter how human work is performed, the pace of AI adoption today has far exceeded the ability to manage risks and enforce governance. Estimates suggest that 82%[8] of open-source software components are considered risky due to factors like poor maintenance, outdated code, and security flaws. Moreover, many open-source projects are run by small teams or individual volunteers with limited resources, leaving them vulnerable to attacks. Threats are becoming more sophisticated, with the potential for supply chain interference in ways that can have significant ripple effects.[9]

For instance, despite the excitement created around a model like DeepSeek, users expressed concerns around data privacy, leading a portion of them to run the model in their own devices as a way to avoid making their data openly available. In the current geopolitical context, concerns around sending data to China, where DeepSeek originated and where its servers may store data, also raised questions, especially in the event of vulnerabilities that could allow sensitive data to be sent over unencrypted channels. Certain governments (e.g., Australia, Italy, Taiwan) have even blocked or restricted access to DeepSeek on government devices, due to national security and privacy concerns, while regulators have raised concerns over lack of transparency and potential user data exposure. It is worth noting, however, that it is possible to download and run DeepSeek (and other models from China) on one's own infrastructure, which would minimize these concerns.

These concerns are an indication that today's AI ecosystem, which has yet to mature, falls short of both ideals of open-source and decentralized AI. "Open" models are often open-weights with incomplete training transparency. Replication is possible in part, but end-to-end reproducibility is theoretical. Decentralized networks grapple with coordination overhead, uneven quality, poisoning and backdoor risks, and fragmented governance. Meanwhile, enterprises gravitate to centralized API providers because they deliver performance, tooling, and support agreements—features that are still maturing in the open and decentralized stack. It becomes evident that open-source and decentralized AI systems have their associated risks that would need to be addressed in order to scale.

Because we haven't yet realized yet true open-source AI or decentralized AI today, and there are still multiple technical and related challenges to get there, it may be deceptive to utilize these terms in a context where nobody has truly released an AI model that is as open and decentralized as to allow truly replicating results.

LIMITATIONS AND RISKS OF OPEN-SOURCE AI

Open-source ecosystems carry specific trade-offs:

Security & Misuse: Functional models can be repurposed for deception, intrusion, or automated abuse without central chokepoints. Because open models pose no restrictions over who uses them and how, bad actors can propagate deepfakes, misinformation, spam, and even repurpose open models for harmful purposes if there are no controls, oversight, or accountability.

Cybersecurity: Because models depend on vast community-managed supply chains, there is no guarantee of legitimate training data, and attackers can analyze and exploit architecture-level weaknesses. Bad actors can inject malicious code, copy or modify weights and redistribute them with hidden backdoors, or re-release "trojaned" versions of models. Compromised model weights or poisoned training scripts can propagate before being detected. Sensitive data can also be leaked and misused.

Quality Control & Maintenance: Documentation, testing, and security hardening vary widely, such that sustainability may depend on contributor bandwidth. Open-source does not guarantee that all projects would uphold high standards, and some may indeed lack adequate governance. Moreover, as many open-source projects rely on unpaid contributors, long-term maintenance, support, and sustainability may be unpredictable.

Fragmentation & Incompatibility: As forks multiply; governance can diverge and standardization would lag, which would ultimately hinder interoperability. This poses challenges for standardizing and regulating a growing ecosystem.

Reproducibility Gaps: Without full data lineage and environment capture, "open" does not equal "replicable."

LIMITATIONS AND RISKS OF DECENTRALIZED AI

Decentralizing AI is currently an evolving process and can be viewed as a moving target, to be achieved at some point in the future, with several challenges if we are to eventually get there at all. Due to the concerns over AI holding hidden biases and obscurity on sources of data, many projects have sought to brand themselves as decentralized AI. Yet centralized approaches still have distinct advantages that are clearly beneficial today, especially as decentralized AI continues to mature. Decentralized approaches to AI also face specific tradeoffs, with the following considerations:

Technical Coordination: Synchronizing many nodes reduces efficiency and complicates versioning and rollbacks because it requires several resources, and it may also be difficult to ensure all nodes contribute clean, high-quality data or compute. Performance inefficiencies may also arise, with slower training and inference (e.g., limits of distrubuted training and federation). Fragmentation may occur, with multiple and inconsistent model versions may also arise across the network.

Security & Trust: Data poisoning and adversarial updates demand robust verification (e.g., cryptographic proofs, attested execution) that are still costly to operate. Bad actors may also insert backdoors. If not designed carefully, even decentralized, federated and peer-to-peer learning models can leak sensitive patterns.

Governance & Accountability: Disputes over upgrades, licensing, and ethical bounds, stemming from a lack of clear authority, can stall progress, such that it may be unclear who is accountable when harm occurs. Ensuring legitimate model updates and results requires robust cryptographic proofs and consensus mechanisms – pointing to the need for effective governance models. Governance, moreover, can become fragmented if different communities adopt divergent rules or protocols, which may ultimately lead to lack of standardization. Community-driven governance brings risks of decision-making being misaligned with broader societal values, leading to potential ethical drift.

Jurisdiction-specific Compliance: Especially when it comes to specific requirements on issues like privacy, data storage and sovereignty, and right to be forgotten in different jurisdictions (e.g., US, EU, China), requirements can be fragmented across different jurisdictions. This can make it challenging for a decentralized AI network, with nodes operating across multiple jurisdictions, to comply simultaneously with conflicting or incompatible laws. In the EU, for instance, GDPR restricts cross-border data flows and requires a "right to erasure" of data. US laws regarding data use may vary by state. China's PIPL regime enforces state oversight and strict localization. Countries like India, Brazil, Singapore, and Canada also have their own mandates on data and AI. IT can be extremely difficult to ensure that data never leaves a given region, while model updates may leak certain sensitive data. Moreover, compliance with erasure rights can be nearly impossible without strong cryptographic controls and techniques for models to un-learn. Accountability in cases of harm may also be very difficult to determine.

Economics & Adoption: Incentives may over-reward activity over quality, and network effects continue to favor centralized incumbents which have more compute, data, and capital. There are high barriers to entry when new infrastructure is required to participate, making it difficult for decentralized alternatives to compete at scale.

Regulatory & Ethical Uncertainty: Supervisory oversight is made more difficult without a central operator or clear standards or harmonized regulatory requirements globally. Data-sovereignty and sectoral rules (e.g., GDPR, HIPAA, data localization and sectoral guidance, etc.) raise open questions for cross-border operations, which can complicate decentralized training.

A MORE PRAGMATIC AVENUE: TRANSPARENT, PRIVATE, SECURE, AND RELIABLE AI

To sustainably implement the benefits of AI, the role of blockchain is key to provide a governance framework to address current challenges with a focus on greater transparency, privacy, security, and reliability. Rather than focus on open-source and decentralized AI, a more realistic alternative may be to focus first on attributes that pave way toward these ideal outcomes. Because open-source and decentralized AI have yet to become a scalable reality in the space, a more realistic approach may be to treat "open" and "decentralized" as gradients, focusing instead on ways to make AI more transparent, private, secure, and reliable.

This way, even though open-source and decentralized AI are not a reality at scale at this moment, they remain aspirational as they have are aspects we can draw on today that can pave way toward a desired AI future. In the short term, a proposed architecture and infrastructure that prioritizes the attributes of transparency, privacy, security, and reliability will pave the way toward future stages of open source and decentralized. One approach can be to identify specific factors that yield measurable assurance of these three attributes and integrate them into production systems that must meet concrete regulatory and business obligations. To ensure greater transparency, privacy, security, and reliability, blockchain technology is a crucial tool to ensure desired outcomes - both for attributes for training the AI models themselves and attributes at inference time, which have important distinctions. AI solutions should adopt the following :

AUDITABILITY

This refers to the ability to independently verify how an AI system was built, how it operates, and whether its claims are legitimate, using transparent, tamper-resistant records and blockchain technology. In such a framework, third-party forensic analysts can query blockchain-based systems of record to confirm that the human-readable assertions made by an AI operator accurately correspond to verifiable on-chain evidence. This bridges the trust gap between what an AI provider claims and what an outside auditor can independently validate, relying on blockchain proofs, cryptographic guarantees, and other mechanisms for establishing legitimate claims. Auditability extends to examining data provenance, the processing of data, and the evolution of model weights, including the ability to use mechanistic interpretability queries to understand how specific inputs, prompts, or query structures influenced model behavior or tuning.

In this context, auditability ensures the quality of a system, allowing it to be reliably examined through durable records, documentation, and evidence, and ultimately ensuring compliance, accuracy, and accountability. To support this, an event-driven architecture is often required: audit events must be designated, captured, and finally processed on-chain or through equivalent mechanisms that record data provenance, runtime performance characteristics, weight attributions, or regulatory compliance checkpoints. Effective AI auditability requires a governance process that clearly defines what constitutes an audit event, encodes the relevant requirements, and ensures the system can produce human-interpretable outputs. These outputs must be paired with a cryptographic verification process that traces the entire auditing workflow. This can function through a structured checklist that validates each step of the AI system's behavior against secure, immutable evidence.

DATA SOVEREIGNTY AND PRIVACY

AI data sovereignty and privacy revolve around giving end users meaningful control over how their data is used by AI systems, including the ability to choose, both at a granular and contextual level, which parts of their data can be accessed for a specific query, and to understand exactly what portions of their interactions, integrations, or data provided are utilized. Three operating models exist with specific user interactions with AI and data control considerations.

Scenario 1	Involves consumers or organizations using an API-driven model operated by a third party. For example, an employer using an external AI service under a licensing or hosting framework, or running a corporate AI model locally where all data and operations remain inside the enterprise environment.
Scenario 2	Represents true self-sovereign AI model, in which individuals use fully local, isolated, self-contained models that do not rely on third-party infrastructure. This differs significantly from consumer tools like ChatGPT, where privacy assurances depend on provider policies rather than user-controlled guarantees.
Scenario 3	Reflects settings where AI is used on people by governments or corporations, often without individuals' meaningful ability to consent or limit data use.

The first two involve either locally hosted systems or AI accessed through an external API, such as a centrally hosted large language model where the provider retains user prompts and responses depending on whether the model is accessed through a free or paid version. The purpose of locally hosted models across these scenarios is to provide strong privacy guarantees, prevent data leakage, ensure isolation, and offer predictable compute resources for organizations. These are conditions that also make it possible to engineer robust guardrails, enforce governance rules, and define auditable events. These protections are far more difficult to ensure when relying solely on API-based external models, where users often exchange privacy for access, particularly in free tiers designed to collect usage data for training or market research. Even paid versions present verification challenges, as providers may assert that user data is not retained or used for model improvement, but without cryptographic verification or blockchain-based audit trails, it is nearly impossible for end users to prove adherence. As models proliferate and usage expands, these privacy risks grow.

To address them, cryptography plays a central role. Privacy-enhancing technologies such as secure enclaves, homomorphic encryption, zero-knowledge proofs, or selective disclosure mechanisms enable granular and enforceable consent rather than blanket data access. These tools ensure that users can authorize specific uses of their data while preventing unauthorized leakage, creating the foundation for verifiable data sovereignty in an AI-driven world.

CUSTOMIZABILITY

Customizability of AI refers to the ability to tailor models to specific users, organizations, and operational environments in order to improve accuracy, reduce errors, and create more meaningful interactions. At the individual level, customization enables personal AI agents that adapt to a user's preferences, history, style, and objectives. For corporations, customization can involve fine-tuning models, implementing graph structures, and building retrieval-augmented generation (RAG) systems that give an AI model access to curated knowledge bases. These approaches not only enhance performance but also significantly reduce hallucinations by grounding model outputs in verified information. Emerging research suggests that many hallucinations stem from the inherent compression dynamics of large models, offering a new perspective on why these errors occur and how they can be controlled through architectural and training adjustments.

Additionally, custom AI systems can incorporate quantitative analysis before a prompt is processed to predict the likelihood of a hallucination. When a prompt is flagged as high-risk, the system can trigger additional safeguards, such as human-in-the-loop review, alternative reasoning routes, or stricter grounding protocols. This layered approach to customization, which can span across personalization, model tuning, structured knowledge integration, and pre-processing analysis, creates AI systems that are safer, more reliable, and better aligned with user and organizational needs.

ETHICAL GOVERNANCE & RECURRENT VALIDATION

Ethical governance and recurrent validation for AI require a proactive and continuous approach that integrates risk management, oversight, and technical monitoring into every stage of an AI system's lifecycle. Instead of treating AI deployment as a one-time "build, deploy, and profit" exercise, organizations need mechanisms that continually assess potential harms, detect shifts in model behavior, and provide pathways for redress, updates, and corrective action. This involves embedding ethical safeguards directly into governance frameworks and making corresponding architectural adjustments to deployment models and runtime environments so that ethical requirements can be practically implemented rather than merely stated. Because AI performance in high-risk or sensitive domains (e.g., healthcare, finance, public safety) cannot be assumed to remain stable over time, ongoing technical validation must be both scientifically rigorous and sociotechnically grounded. Models trained on clinical, demographic, or behavioral data need adequate monitoring to ensure their outputs remain accurate, fair, and safe for the populations they serve, especially if performance naturally degrades or contexts change.

The need for recurrent validation is further underscored by the continual emergence of new security vulnerabilities and the corresponding countermeasures that must be deployed to maintain system integrity. As regulatory frameworks evolve, organizations must also adapt their AI systems to stay compliant, incorporating new rules into operational practices rather than retrofitting compliance after harm occurs.

Experience-based training is an essential part of this ethical foundation: AI must be able to demonstrate the authenticity and provenance of the data it was trained on, anchoring its outputs in reliable ground truths. Systems trained solely on digital or synthetic datasets will inevitably fail to capture the complexity of lived human experience, limiting their effectiveness and increasing the risk of harm. Ethical governance, therefore, is not a static checklist but a dynamic, recurring process that ensures AI systems remain accountable, safe, context-aware, and aligned with human values throughout their operational lifespan.

SELF-ASSESSMENT FOR AI TRANSPARENCY, PRIVACY, SECURITY, RELIABILITY

It is essential for all stakeholders to take proactive measures and develop a sense of shared responsibility to ensure transparent, private, secure, and reliable AI solutions that pave way toward a true open-source and decentralized AI future. Self-awareness can be a key factor to mobilize stakeholders to take steps toward more legitimate AI uses, while avoiding questionable and suboptimal procedures.

This section translates the earlier sections' argument into a practical instrument, proposing a Self-Assessment approach for individuals, companies, and organizations to evaluate the transparency, privacy security, and reliability of their AI uses. The thesis is straightforward: if AI is to create long-term value, it must be transparent, private, secure, and reliable in ways that can be evidenced, not merely asserted. Earlier, we traced how enthusiasm for open source and decentralized AI has outpaced what enterprises can actually guarantee; we also proposed a selective, evidence-first architecture where blockchain is used as an assurance utility: anchoring provenance, emitting tamper-evident audit events, and enabling selective disclosure through verifiable credentials

A self-assessment operationalizes that architecture. It helps teams and individuals locate their current assurance maturity and prioritize what to do next. Concretely, it asks whether one can show, on demand and without oversharing, what data went in, what happened at inference, and who is accountable for what. Where earlier sections framed the ideals (openness, decentralization) and the compromises (coordination, quality, accountability), this tool focuses on what can be proven today, mapping each control to verifiable evidence, such as provenance anchors for data and weights, event-driven audit for model, and policy changes.

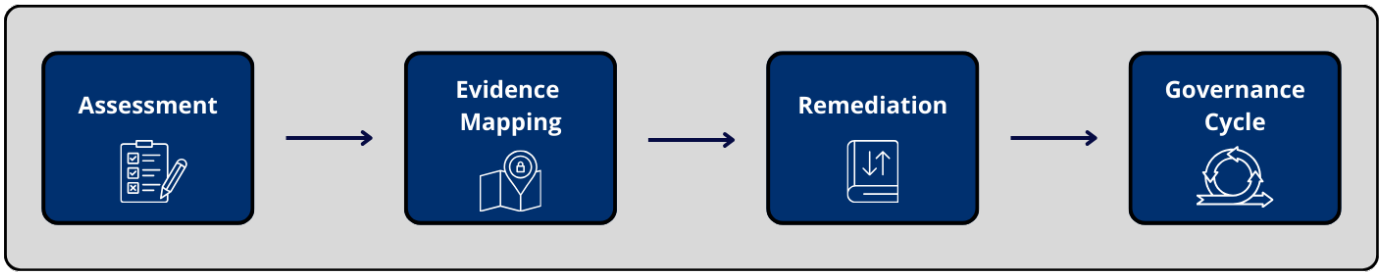
The instrument is intentionally broad in audience. Individuals running local models, SMEs fine-tuning open weights, and large enterprises orchestrating API-based copilots can all complete it. It is equally applicable to outsourced phases: labelling, fine-tuning, evaluation, even full training - by requiring machine-readable provider attestations that you can anchor into your own records. In every case, the emphasis is the same: move from claims to cryptographically backed evidence.

This full self-assessment questionnaire, which can be found in ***Annex 1***, scores AI implementations as low/medium/high in terms of transparency, privacy, security, and reliability respectively. The scoring methodology assigns points for each question to determine low/medium/high for each category (e.g., depending on answers or by yes/no answers, etc.). The final evaluation from the assessment would categorize overall AI use as fair/good/great.

The self-assessment is a tool meant to help improve AI users' awareness of what level of AI transparency, privacy, security, and reliability they are operating with, as well as present blockchain as a solution to move toward higher levels of AI transparency, privacy, security, and reliability. This is essential for users to take ownership and assess the expected accuracy of their own AI-informed decision making and outcomes, as well as how holistic their approach to AI currently is. Moreover, awareness of the resilience of AI solutions utilized can help identify any risks and take remedial steps. Eventually, this assessment framework could be tied to certifications and globally recognized standards, as a path toward compliance.

WHAT THE SELF-ASSESSMENT INSTRUMENT DOES

- Provides a structured, lifecycle-based review of AI assurance across five dimensions and produces an assurance maturity profile, plus a remediation plan indicating where web3 controls (anchors, attestations, selective disclosure, authenticity credentials) raise assurance with minimal disruption.
- Works for individuals and organizations—from a self-hosted local model to an enterprise portfolio and third-party APIs.



URGENCY OF SELF-ASSESSMENT MEASURES

As the innovation community continues to chase the next breakthrough, whether quantum advancements, new amplification tools, or novel "killer apps," there is an urgent need to communicate that responsible AI is not optional but essential. The accelerating complexity and influence of AI systems require a model of ethical governance that is not merely aspirational but structurally enforced, and this is where Web3 becomes indispensable. We need a Web3-dependent responsible AI framework because traditional AI alone cannot meet the demands of transparency, provenance, auditability, and verifiable compliance. Blockchain and related technologies provide the backbone for addressing these challenges, offering mechanisms that align with the intricacies of modern AI systems.

Adoption has outpaced assurance. Earlier sections tracked how AI moved from pilots to production across content, decision support, and domain copilots, with uptake visible across firm sizes. That same momentum exposed a gap: many programs cannot evidence the basics when challenged by a board, a regulator, or a customer—what data went in, what happened at inference, who approved the change that altered behavior. In parallel, the reference rails are hardening, with evolving regulatory requirements and standards.

Two additional pressures make the timing acute.

- **First, outsourcing and API dependence:** much of modern AI is trained, tuned, or served by third parties. Without machine-readable attestations from providers—and a place to anchor them—assurance maturity hits a ceiling, no matter how capable the model.

- **Second, data quality and sustainability:** pipelines drifting toward synthetic-on-synthetic inputs degrade silently while consuming more energy/compute. A managed ground-truth budget, verifiably sourced via web3 credentials and content authenticity tools, is the most reliable corrective (e.g., considerations for W3C DID/VC standards, C2PA, and sustainability metrics).

The Self-Assessment arrives, therefore, as a timely instrument: it internalizes this report's conceptual arc: from ideals, through compromises, to pragmatic architecture, and converts it into actionable diagnostics. It helps teams identify their current assurance maturity and shows how to move it upward using the controls already set out in this chapter, including provenance anchors, event-driven audit, selective disclosure, content authenticity, and confidential assurance. In other words, it operationalizes the central claim that responsible AI is ledger-dependent, because the forms of evidence that confer legitimacy are most credibly delivered with blockchain and related cryptography. This may also pave the way toward greater compliance as regulations like the EU AI Act, GPAI, and other requirments develop.

FROM SELF-ASSESSMENT TO A FUTURE OF RESPONSIBLE AI WITH BLOCKCHAIN TECHNOLOGY

We leave it to individuals and organizations to choose a methodology that works for best them, as long as it satisfies the criteria that we laid out in this paper. Advancing toward the future, we must resist both naive optimism (assuming that with enough data the models will eventually figure it out) and fatalistic pessimism (AGI-doom). The imperative is neither to blindly trust nor to dismiss these systems but to operate responsibly, and responsibility at scale requires cryptographic, decentralized, and verifiable infrastructure. Implementing blockchain and ledger technologies is therefore not a peripheral enhancement; it is a foundational element of how AI must function going forward. As part of this shift, it becomes increasingly important to evaluate the attributes, limitations, and governance approaches of widely used systems and large models, ensuring that they align with principles of ethical, accountable, and transparent AI.

This is precisely where blockchain as enabler matters, as verifiable provenance and tamper-evident audit are the shortest path toward trustworthy AI at scale. It is not necessary to record sensitive data, or any data at all, on chain, but instead it is necessary to anchor facts about the data and the model lifecycle. It is not necessary to make everything public, but instead it is necessary to implement a selective disclosure so third parties can check claims without seeing raw inputs. We don't need a fully decentralized stack to benefit; we instead need the minimum cryptographic commitments that make your system verifiable and replayable when it counts. Blockchain technology strengthens AI systems precisely at the data layer, where responsible AI challenges may be most acute. Regulations such as the EU AI Act, which requires foundation models to disclose training data sources, categorize high-risk systems, and specify expectations for data governance and copyright transparency, illustrate why blockchain-based verification will become increasingly necessary. Organizations operating in the EU and beyond can determine how blockchain can help them meet these requirements, from automated compliance proofs to data-lineage tracking.

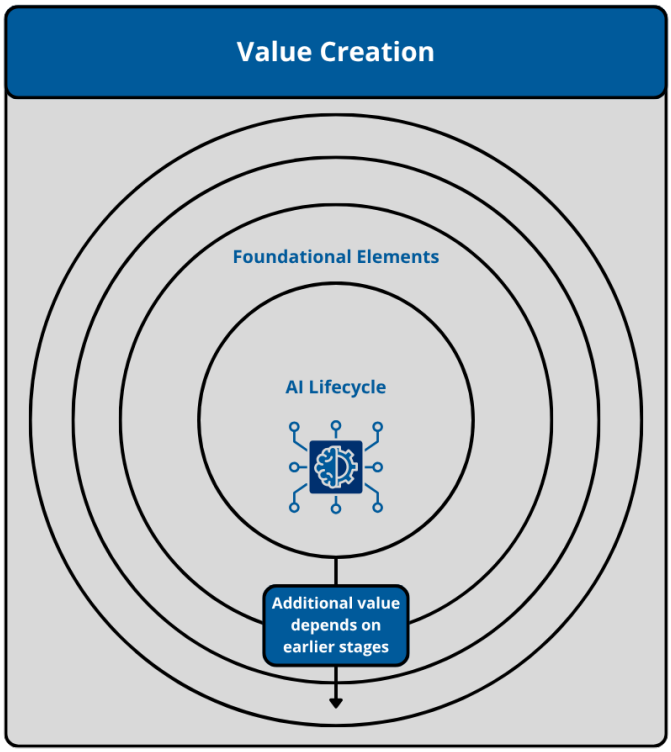
One approach we propose is to view the value that blockchain adds to AI implementations in three tiers:

Good
Ensures conformance with relevant rules and standards

Better
Introduces ledger-enabled guardrails that enforce policies and protect against misuse

Best
Applies blockchain capabilities strategically to create and deploy AI systems that are verifiably responsible, resilient, trustworthy, and aligned with human and environmental well-being

An AI lifecycle approach to unpack the results of the self-assessment into actionable strategies may be beneficial. A long-term, phased approach can be considered for moving from self-assessment to a future of responsible AI, supported by blockchain technology. The process begins by defining a clear methodology, then building the backend infrastructure, followed by the frontend interface, and finally connecting the components and executing the necessary integrations. This progression can be visualized below:



AI Lifecycle

DATA COLLECTION
(ground truth, synthetic, 3rd-party data)

TRAINING DATA

Risks	Blockchain's Role	Self-Assessment Opportunites
• Unverified or low-quality data • Bias, demographic skews • Data provenance uncertainty • IP/copyright violations • Privacy leakage from training sets	• Immutable provenance tracking (source, timestamp, rights) • Cryptographic proofs or dataset authenticity • Consent receipts & selective disclosure • Verification of lawful data usage • Audit trails for outsourcing of labeling/data	• Data diversity & representativeness • Dataset licensing & compliance • Synthetic vs. real data balance • Validate consent and privacy controls • Data minimization practices

MODELING
(training runs, fine-tuning, parameter updates, weight generation)

Risks	Blockchain's Role	Self-Assessment Opportunites
• Hallucinations from over-compression • Hidden biases in weight formation • Poisoned or manipulated training data • Undetected degradation across versions • Unverifiable claims about model behavior	• Version-controlled, tamper-proof model checkpoints • Cryptographic attestation of training events • Mechanistic interpretability records anchored on-chain • Proofs of training conditions • Secure multi-party training records	• Training documentation completeness • Hyperparameter impact on safety • Version history for improvements or regressions • Test model for bias, draft, adversarial susceptibility • Validate model grounding practices (RAG, graphs, etc.)

COMPUTATION / INFERENCE

(runtime behavior, prompt processing, context usage)

Risks	Blockchain's Role	Self-Assessment Opportunites
• Prompt injection and input manipulation • Use of user data outside approved scope • Lack of transparency in real-time decisions • Non-reproducible model behavior • Runtime vulnerabilities (e.g., jailbreaks)	• On-chain logging of inference events (privacy-preserving) • Policy enforcement via smart contracts • Selective data disclosure using 2k proofs • Proving that "data was NOT used" (negative disclosure proofs) • Runtime guardrails encoded as verifiable rules	• Prompt safety screens • Runtime privacy architecture • Adherence to data-sovereignty requirements • Hallucination-risk scoring mechanisms • Validate human-in-the-loop escalation paths

OUTPUT & ACTION

(generated results, decisions, predictions, automation)

Risks	Blockchain's Role	Self-Assessment Opportunites
• Harmful or inaccurate outputs • Unsafe automation based on faulty results • Copyright/IP misattribution • Lack of explainability • Weak accountability for downstream use	• Cryptographically verifiable output signatures • Traceability from output → model → training data • Smart-contract-based safety policies • Immutable audit logs for decisions & actions • Attribution tracking for copyrighted materials	• Accuracy & reliability metrics • Explainability and justification quality • Alignment with organizational policies • Validate that automation triggers comply with governance • Check for differential impacts across user groups

FEEDBACK, DRIFT & RETRAINING

Risks	Blockchain's Role	Self-Assessment Opportunites
• Model drift & performance decay • Feedback loops reinforcing bias • Shadow AI systems evolving without oversight • Lack of transparency in modification history	• Immutable ledger of all updates • Governance-driven "audit events" • Drift detection proofs (performance logs anchored to chain) • Verifiable model lineage across retraining cycles • Regulatory compliance-evidence submissions	• Retraining justification and criteria • Drift detection systems • Ensure governance process compliance • Validate periodic ethical and security assessments • Check tracking of human-centered & earth-centered metrics

Importantly, this framework avoids a simplistic view of "responsible vs. irresponsible" AI and instead emphasizes a layered value creation approach, where each step strengthens the overall system's reliability, sustainability, and alignment with human goals. Value in this context means the ability to achieve desired outcomes without unnecessary harm, excessive consumption of energy or compute resources, or reliance on low-quality or inappropriate data inputs.

Incorporating human-centered and earth-centered value creation is critical. These principles are not in opposition to enterprise or government priorities. Rather, they reinforce each other and serve as force multipliers when properly aligned. Yet in today's digital ecosystem, where individuals are represented as data profiles spanning thousands of parameters, significant risks arise when AI systems are trained primarily on synthetic data or when they retain uncorrected errors indefinitely (e.g., AI models cannot "forget" mistakes not identified as such). Much AI model training is outsourced, but without assurances of secure and trustworthy processes, organizations lack visibility into whether their models are built on authentic, ethically sourced information. The future of responsible AI depends on securely developed ground truths, using data that reflects real lived experiences and is anchored in verifiable provenance. Achieving this level of trust is not possible without robust web3 infrastructure.

Continued industry engagement can also help organizations ask the right questions about sustainability, ethical data sourcing, and the balance of inputs such as synthetic data, third-party datasets, and the minimum amount of certified ground-truth information required, ultimately validated through web3 mechanisms.

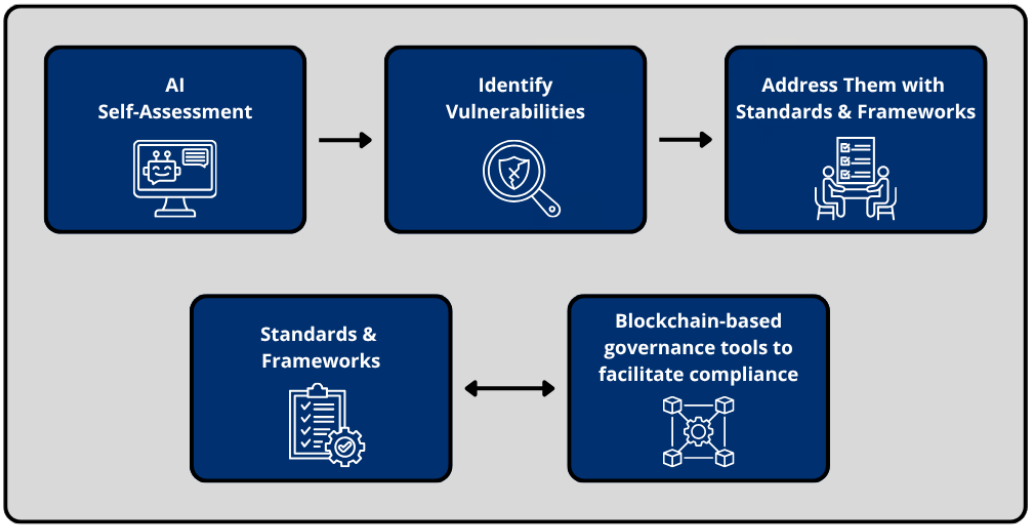
STANDARDS ALIGNMENT

Ensuring a future toward open-source and decentralized AI at scale requires a rigorous examination of existing standards and frameworks. While existing standards and frameworks for responsible AI do not generally call for blockchain explicitly, we recommend including an assessment of how they can be enhanced through blockchain technology. This involves identifying relevant AI standards and frameworks, and mapping them against decentralized solutions that strengthen provenance, auditability, governance, and verifiable compliance. Organizations can also make use of risk assessment tools, including the proposed Responsible AI Self-Assessment in Annex 1, and evaluate them against global responsibility standards using a process-driven perspective.

Key technology standards and frameworks to consider include:

- **ISO 42001** series, which provides a governance scaffold for managing AI responsibly
- **ISO 27001** to manage and protect information assets with Information Security Management Systems (ISMS)
- **W3C Verifiable Credentials**, which support trusted contributor and evaluator identity
- **C2PA credentials**, which authenticate public outputs
- **NIST AI Risk Management Framework** and **Generative AI Profile**, which translate risk concepts into operational controls
- **EU AI Act** and **GPAI Code of Practice**, which introduce legally enforceable requirements around provenance, copyright diligence, systemic-risk management, and transparency
- **The Operationally Critical Threat, Asset, and Vulnerability Evaluation (OCTAVE) Methodology**[10] for managing information security risks for both large and small organizations
- **EU General Data Protection Regulation (GDPR)**, which protects EU citizens' fundamental right to data privacy. It requires organizations to obtain individuals' consent before processing their data, granting individuals rights over their own personal data, including the right to access, correct, and delete it, often referred to as the "right to be forgotten."

This alignment effort also remains consistent with the broader regulatory and standards landscape discussed earlier. Each standards component has a function that can be cross-referenced to blockchain-based governance tools (e.g., ISO/IEC 42001 for AI management systems, NIST AI RMF and Generative AI Profile for operational risk controls, and EU AI Act/GPAI for data governance, disclosure, and systemic-risk expectations). The United Nations System White Paper on AI Governance[11], for instance, provides an analysis of the UN system's institutional models, functions, and existing international normative frameworks applicable to AI governance. Ultimately, the objective is to identify which types of evidence matters most and to illustrate how blockchain and Web3 technologies can make those evidence requirements durable, tamper-resistant, and testable.



In this context, the self-assessment serves as a bridge between conceptual guidance and the practical, day-one steps that any organization can take to embed responsible AI practices into its operations. Self-assessment tools facilitate identifying and addressing vulnerabilities in assets (e.g., hardware, software, IP) and people, in order to evaluate decision trees (e.g., was an unwanted outcome intentional or accidental?), and measure the impact of a given threat. With the Responsible AI Self-Assessment questionnaire, we hope to provide a meaningful first step that can toward transparent, private, secure, and reliable AI at scale, relying on web3-based governance mechanisms.

USE CASES FOR A FUTURE OF OPEN-SOURCE AND DECENTRALIZED AI

Two areas highlighted below, which can benefit from a future of responsible open-source and decentralized AI, are healthcare and education & workforce development.

HEALTHCARE

Healthcare examples can bring color the current concerns around responsible AI: when we don't have quality data inputs, reliability plummets. Especially when an AI model has little to no knowledge on what's happening to a physical being, learning from lived experience is hugely important. As a response, organizations such as the Coalition for Health AI (CHAI)[12] have emerged to help the industry establish responsible development, deployment, and oversight practices, recognizing that transparent, private, secure, and reliable AI models must learn from real lived experiences, beyond merely clinical records and synthetic datasets. Without a grounding on human experience, some models may deliver accuracy rates as low as 7%, which is especially concerning when lives are at stake. Historical examples like the nutritional "Food Pyramid," which had clear flaws, underscore the risks of suggestions and guidance that are based on weak or irrelevant data.

Today's digitization of health data, alongside the growth of data marketplaces, add new risks when the diversity of lived experiences that influence health outcomes is not captured into medical records. Open-source and decentralized AI models can help fill this gap by enabling individuals to contribute verified personal data while retaining control of their own data, with the added opportunity for economic incentives to sustain systems that truly reflect human diversity and lived experiences.

We reference a recent case where a man went blind after being prescribed GLP-1 medicines (e.g., Ozempic), illustrating serious adverse events tied to medications when certain medications interact differently across individuals. In this GLP-1/ Non-Arteritic Ischemic Optic Neuropathy (NAION) case, understanding a person's optic-disc size, a data point not found in standard health records, could dramatically change their risk calculation before starting a drug like Ozempic.

Therapies may fail to account for personal variation, leaving subsets of people exposed to serious side effects. The ensuing series of lawsuits[13] provide a reason for the next frontier in healthcare, with personal AI agents that act as digital fiduciaries for individuals. These agents would monitor evolving medical, scientific, and risk data in real-time and translate that data into personalized insights tailored to individuals' risks and benefits.

The traditional approach of waiting for slow scientific consensus or passive regulatory frameworks is no longer sufficient. Instead of "Ask your doctor first," open-source and decentralized AI allows a new paradigm of "Ask your agent first."

This underscores an opportunity for open-source and decentralized AI solutions to enable a major shift in healthcare approaches, evolving beyond medicine tailored to an average population, and instead focusing on approaches that are tailored to specific individuals. This can be made possible with a decentralized, privacy-preserving ecosystem where individuals own their data and receive meaningful, individualized support from intelligent systems built on verified ground truths.[14]

To achieve a future vision of personalized health intelligence,[15] AI and blockchain technology come together in a broader Web3 context to ensure data provenance, secure consent, decentralized identity, and mechanisms that allow individuals to benefit economically from sharing their data. Within this model, agents acting as digital fiduciaries can manage consent, privacy protections, and deliver personalized intelligence. A vision of digital health solutions, built much like a Waze model for traffic, with aggregated and anonymized lived experiences as inputs, would allow individuals to navigate complex medical decisions with tailored and context-specific guidance.

Open-source and decentralized AI solutions offer an unprecedented level of transparency, auditability, and adaptability to high-stakes medical environments, allowing clinicians and regulators to inspect how models work, validate their reasoning, and customize them for local needs. Decentralized architectures enable hospitals to run models locally, while preserving privacy complying with regulations, and supporting federated learning collaborations across institutions. Broadly shared goals like cancer detection or sepsis prediction can become shared tasks without sharing raw data.

- **Open-source frameworks can also strengthen public health research, decision support, integration of electronic health records, and even care delivery in multiple languages.**
- **Decentralized systems can enhance pharmaceutical R&D, supply-chain integrity, and surveillance of adverse drug events.**

Together, AI and blockchain technologies provide a foundation for a more equitable, interoperable, and trustworthy healthcare ecosystem. They minimize reliance on black-box systems, support global innovation, and ensure that AI remains aligned with human safety, individual context, and clinical reality.

OPEN SOURCE AND DECENTRALIZED AI FOR EDUCATION & WORKFORCE DEVELOPMENT

A future of open-source and decentralized AI at scale offers powerful benefits to education systems and the workforce by enabling transparent, equitable, and adaptable technology solutions to support diverse learning and labor environments.
For education and professional training, AI can scale personalized learning, which has proven to be the most effective (e.g., in person tutoring), in a way that adapts to each student's progress. Hyper-personalized education, once accessible only through private tutors, can now be scaled through AI. In an AI-driven world where students are already relying on tools like ChatGPT concerns have arisen around jumping from not knowing to having the answer without actually learning. The core challenge is that learning is a process that requires a necessary struggle. The solution is not to reject AI but to integrate it as a personalized guide through micro-struggles, helping students build understanding rather than merely produce answers.

- **Open-source models** can be inspected, customized, and aligned with local curricula and cultural contexts, helping educators ensure that AI-powered tutoring, grading systems, and content-generation tools operate without hidden bias and reflect institutional values.
- **Decentralized AI** reinforces privacy by allowing schools to run models locally, keeping student data on-site and ensure compliance with data protection and other laws.

These solutions can be especially transformative for expanding global access. Open-source models can be freely modified, translated, and deployed on low-resource devices, allowing underserved communities to benefit from solutions like AI tutors, literacy tools, STEM assistants, and vocational programs, even without reliable internet. This can reduce dependence on proprietary vendors and foster innovation through community collaboration on lesson plans, assessments, and specialized modules.

Moreover, students, who face pressure to prioritize grades in traditional systems, can be encouraged to shortcut learning with AI. In a future of open-source and decentralized AI, they increasingly need credentials with real utility, which provide proof of skills, which they can also control and share selectively. Soft skills like insight, communication, and emotional intelligence also become essential, as these cannot be replaced by AI.

In this context, educators must shift their goals from helping students pass tests to helping them internalize any given subject matter. Instead of treating AI as cheating, teachers can benefit from incorporating AI tools into assignments, requiring students to disclose prompts, validate outputs, and learn the strengths and limitations of these systems. Because AI often fabricates references or delivers inconsistent results across prompts, students must be taught how to critically evaluate these outputs rather than accept them as truth.

In the workforce, employers increasingly expect employees to use AI tools to be more effective, but foundational knowledge and insight remain irreplaceable. While AI surpasses humans in certain tasks (e.g., data retrieval, calculation, pattern recognition), wisdom remains uniquely human, grounded in experience, emotion, judgment, and context. As employers adopt AI broadly, they will hire and reward people for wisdom, creativity, and adaptability rather than brute information recall.

As for workforce development in particular, open-source and decentralized AI systems are essential for building transparent, fair tools for hiring, skills assessment, and workforce planning, while ensuring sensitive employee data remains within secure environments. These solutions can democratize reskilling by allowing workers to run AI-powered learning companions or career planning tools privately on personal devices. Open-source ecosystems allow employers, trade schools, unions, and public agencies to collaboratively develop training materials and interoperable skills frameworks that prepare the workforce for evolving themes such as cybersecurity, clean energy, advanced manufacturing, and AI engineering. These solutions reduce reliance on closed platforms, support accountable decision-making, expand access to learning, and protect worker and student autonomy.

With respect to the changing landscape of work, AI is rapidly transforming the economy by automating tasks across industries, changing the way employment will provide meaning, stability, and economic inclusion. As automation accelerates, societies must rethink what constitutes value, moving beyond "work equals value" toward models that include data, intelligence, and lived experience as legitimate economic contributions. This shift requires new policy frameworks, safety nets, and measurement systems that acknowledge that traditional labor cannot remain the sole foundation for income or mobility. Personal data, identity, and real-world experience will become increasingly important economically, with self-sovereign identity models where individuals can own their data and be rewarded for sharing it. Local organizations and governments, such as US states, are therefore faced with the need to invest in systems that help individuals retain agency and benefit from emerging value flows, to reduce the risk of being excluded from new economic incentive structures.[16]

Across education, employment, and economic policy, the central message is that open-source and decentralized AI—combined with thoughtful governance and new value frameworks—can create a future in which people are empowered rather than displaced, supported rather than surveilled, and able to thrive in a rapidly evolving digital landscape.

CONCLUSION: RECOMMENDATIONS FOR BLOCKCHAIN TO INCREASE TRANSPARENCY, PRIVACY, SECURITY, AND RELIABILITY OF AI

Industry self-assessment and customer assessments are important decision-making factors for engaging with AI services. In the journey toward increasing levels of open-source and decentralization at scale, which ultimately point toward reliability and governance, there is a need and a market for blockchain solutions in the AI space.

To bridge ideals of open-source and decentralized AI with enterprise and business reality, GSMI recommends a three-layer assurance architecture that uses blockchain alongside privacy-enhancing technologies (PETs). For instance, there's no reliable AI without blockchain being part of the provenance, making sure data going into AI models is reliable. The aim is not to "put everything on chain," but to anchor truth about the things that matter most: what was used, what changed, what happened, and who attested to it. Ultimately, the goal is for AI-driven informed decision making to increase in trustworthiness, toward positive outcomes for humanity and society as a whole. The interplay between AI and blockchain technology can, in this way, support the industrial and digital revolution taking place, in a responsible way.

Layer A: Provenance & Policy (Before You Ship)

- Establish a lineage registry for data, weights, prompts, and retrieval corpora, recording hashes, licenses, consents, and jurisdictional constraints.
- Use W3C Verifiable Credentials (VC 2.0) to identify and authorize contributors, evaluators, and auditors, with selective disclosure (e.g., SD-JWT/COSE) to minimize data exposure (e.g., considerations with respect to W3C VC 2.0; SD-JWT/COSE specs arise).
- Where copyright and licensing matters, bind license checks and opt-out/opt-in policies to artifacts before training or fine-tune

Layer B: Event-Driven Audit (As You Ship and Operate)

- Define audit events up front: model version changes; dataset, index, or safety-policy updates; system prompt or template modifications; and anomaly flags.
- For each event, emit a signed, time-stamped record whose commitment is anchored to a ledger, enabling independent replay of inference context (model identifier, policy bundle, retrieval set hash) without revealing personal data. Considerations arise with respect to event schema, anchoring approach, and privacy notes.

Layer C: Confidentiality & Authenticity (While You Communicate)

- Enforce purpose-bound access and consent using verifiable policies and PETs (TEEs, MPC, ZK; with FHE treated as an augment as standards mature).
- For external content, embed C2PA credentials and—where appropriate—anchor issuer attestations to provide a durable authenticity trail for downstream stakeholders.

EXAMPLE OPERATING MODELS

- Enterprise-hosted (self-managed). Maximum control over lineage, audit, and confidentiality; strongest fit for regulated workloads.
- Enterprise on open-weights (self-hosted). Prioritize supply-chain integrity (signed weights, tokenizers, loaders) and reproducibility.
- API-based (frontier third-party). Treat provider claims as attestations: require machine-readable statements on data handling, model updates, and incident response; bind them into your audit log and map to ISO/IEC 42001 and NIST controls (e.g., ISO/NIST clause mapping would be helpful). For EU exposure, incorporate GPAI Code of Practice commitments (e.g., considerations for code provisions arise).

What "good" looks like (outcomes over means):

- **Auditability:** measurable coverage of anchored lifecycle events; time to forensic replay (e.g. considerations for KPI ranges arise).
- **Data Control & Sovereignty:** proportion of inferences with license/consent-conformant inputs and selective disclosure.
- **Security & Privacy:** PET coverage, leak incidents, attested execution rates.
- **Reliability Under Change:** reproducibility rates, drift detection lead time, rollback success.
- **Governance & Compliance:** alignment with ISO/IEC 42001, NIST GenAI Profile, and EU GPAI expectations.

OPEN SOURCE & DECENTRALIZATION — OPEN QUESTIONS

1. **Minimum Provenance:** What minimum viable provenance (data, weights, prompts, retrievals) would make an organization comfortable for automating higher-stakes decisions?

2. **Placement of the Ledger:** Which audit events belong on-chain (or anchored) versus off-chain logs, and how should sensitive data and trade secrets be protected in each model?

3. **Open vs. Centralized Trade-offs:** Which "open" attributes (inspectability, reproducibility, verifiable contribution credits) deliver the most assurance even when hosting remains centralized?

4. **Decentralized Economics:** How should incentives reward quality of contributions (data cleanliness, red-team value, evaluation rigor) rather than mere activity?

5. **GPAI Readiness:** For entities exposed to the EU market, what evidence packages (model cards, safety evaluations, copyright diligence) will they prepare—and which commitments of the GPAI Code of Practice will they prioritize and adopt?

6. **Continuous Validation:** What cadence of drift testing, safety evaluation, and red-teaming is appropriate for your risk class, and how will results be anchored for audit?

ANNEX 1: RESPONSIBLE AI SELF ASSESSMENT QUESTIONNAIRE

A. PURPOSE

This questionnaire helps organizations assess the transparency, privacy & security, and reliability & governance of a specific AI system or use case. It is intended as a practical maturity check and input into risk management, not a formal audit.

B. HOW TO USE THIS QUESTIONNAIRE

1. **Scope**
   Complete the questionnaire for one AI system/use case at a time (e.g., "customer support chatbot", "credit risk model", "internal coding assistant").

2. **Response scale (for all scored questions)**
   For each question, select one:

   0 – Not in place / unknown
   No evidence, not implemented, or unknown.

   1 – Partially in place
   Implemented in some areas, informal, or incomplete.

   2 – Fully in place and documented
   Implemented, documented, and used consistently.

3. **Scoring**
   i. Each question is tagged to one of the three attributes:

   T = Transparency
   S = Security & Privacy
   R = Reliability & Governance

   ii. For each attribute:

   a. Add up the points for questions in that attribute.
   b. Calculate:

$$\text{Attribute score } (\%) = \frac{\text{points obtained}}{\text{maximum possible points (excluding N/A)}} \text{ x } 100$$

   iii. Interpret attribute scores as:

   0–39% = Low
   40–69% = Medium
   70–100% = High

4. **Overall evaluation (illustrative mapping)**

   Fair – No attribute is Low; at least one is Medium.
   Good – At least two attributes are High; none is Low.
   Great – All three attributes are High.

   Organizations may adjust thresholds and labels to fit their own risk appetite.

E. N/A answers

i. If a question is not applicable, mark N/A and exclude it from the maximum possible points for that attribute.

C. SECTION 0 – CONTEXT & CRITICALITY (NOT SCORED)

These questions provide context for interpreting the scores.

C1. System description
Briefly describe the AI system/use case (purpose, main users, and decisions it supports).

C2. Type of decision
Select the primary decision type:

- □ Advisory / decision support only
- □ Automated decision with human approval or override
- □ Fully automated decision with limited or no human intervention

C3. Potential impact of failure or misuse
What is the plausible worst case impact if the system fails or behaves incorrectly?

- □ Low – inconvenience, minor process inefficiencies
- □ Medium – financial or operational impact, reputational concerns
- □ High – impact on safety, rights, access to essential services, or legal exposure

C4. Data types involved (tick all that apply)

- □ Public / open data
- □ Internal non personal operational data
- □ Personal data
- □ Sensitive or special category personal data
- □ Children's data
- □ Trade secrets / highly confidential business data
- □ Other (describe): _____

D. SECTION 1 – TRANSPARENCY (T)

T1. Documented data sources
Training and operational data sources (including owners, licences, and collection methods) are documented.

0 / 1 / 2 / N/A

T2. Provenance & limitations
For each key dataset, provenance, known gaps, and limitations (e.g., coverage, demographic skew, time period) are identified and recorded.

0 / 1 / 2 / N/A

T3. Synthetic data use
The proportion and role of any synthetic data are known, justified, and tested so they do not materially distort real world performance.

0 / 1 / 2 / N/A

T4. Data quality management
There is a defined process to detect, track, and remediate data quality issues (e.g., missing, erroneous, outdated data) for both training and live data.

0 / 1 / 2 / N/A

T5. Performance across groups/contexts
Model performance is evaluated across relevant user groups or contexts, and material performance gaps are identified.

0 / 1 / 2 / N/A

T6. Fairness & bias measures
Fairness or bias metrics appropriate to the use case are defined, periodically measured, and action is taken when thresholds are breached.

0 / 1 / 2 / N/A

T7. Model assumptions & limitations
Key assumptions, intended use, and known limitations are documented in an artefact accessible to relevant technical and non technical stakeholders.

0 / 1 / 2 / N/A

T8. User transparency about AI use
Users or affected individuals are informed when AI is used in a way that meaningfully influences outcomes affecting them (e.g., decisions, recommendations).

0 / 1 / 2 / N/A

T9. Explainability & support
There is a clear, understandable explanation of how the system reaches outcomes (or why it cannot be fully explained) and how users can obtain support or clarification.

0 / 1 / 2 / N/A

E. SECTION 2 – SECURITY & PRIVACY (S)

S1. Security-by-design requirements
Security requirements specific to this AI system (including data, model, and infrastructure risks) are defined and integrated into design and architecture decisions.

0 / 1 / 2 / N/A

S2. Access control & logging
Access to models, training data, and configuration is restricted (e.g., role based access control) and administrator actions are logged and reviewed.

0 / 1 / 2 / N/A

S3. Protection of data in transit and at rest
Data used for training and inference is protected in transit and at rest (e.g., encryption, network segregation, key management).

0 / 1 / 2 / N/A

S4. Data protection compliance
The system's use of personal data complies with applicable data protection laws and internal policies (e.g., a DPIA or equivalent has been completed where needed).

0 / 1 / 2 / N/A

S5. Data minimisation & retention
Personal/sensitive data used for training and operation is minimised, and retention/deletion schedules are defined and implemented.

0 / 1 / 2 / N/A

S6. Data leakage controls
There are technical and procedural controls to prevent leakage of sensitive information via prompts, logs, model outputs, or third party providers.

0 / 1 / 2 / N/A

S7. AI specific threat mitigations
The system has been assessed for AI specific threats (e.g., prompt injection, model exfiltration, data poisoning), and appropriate mitigations are in place.

0 / 1 / 2 / N/A

S8. Patching & dependency management
Patching and vulnerability management covers AI related components (frameworks, libraries, model artefacts, dependencies) on a defined schedule.

0 / 1 / 2 / N/A

S9. AI incident response
There is a documented and tested incident response playbook covering AI related incidents (detection, triage, containment, communication, and learnings).

0 / 1 / 2 / N/A

F. SECTION 3 – RELIABILITY & GOVERNANCE (R)

R1. Pre deployment testing & validation
The model has been tested with representative data, including edge cases and stress tests, before deployment.

0 / 1 / 2 / N/A

R2. Defined performance targets
Baseline performance and quality metrics (e.g., accuracy, error rates, latency, business KPIs) are defined, agreed, and documented for this system.

0 / 1 / 2 / N/A

R3. Monitoring for drift and anomalies
The system is monitored in production for performance degradation, data drift, and anomalous or unsafe outputs.

0 / 1 / 2 / N/A

R4. Change & rollback process
There is a defined process for updating, retraining, and rolling back models and associated data pipelines, including approval steps and testing.

0 / 1 / 2 / N/A

R5. Human oversight for higher risk decisions
For higher impact use cases, appropriate human oversight is in place (e.g., human in the loop or human on the loop) with clearly defined roles and decision rights.

0 / 1 / 2 / N/A

R6. Ability to challenge or appeal
Users or affected individuals can challenge or request review of AI influenced decisions, and such cases are tracked and analysed.

0 / 1 / 2 / N/A

R7. Clear accountability
Accountability for the AI system is explicitly assigned (e.g., business owner, technical owner, risk/ compliance contact).

0 / 1 / 2 / N/A

R8. Governance oversight
An AI governance or risk body (or equivalent) reviews high risk AI use cases at defined points (e.g., design, pre launch, periodic review).

0 / 1 / 2 / N/A

R9. Periodic re assessment & improvement
The system and its controls are periodically re assessed (e.g., annually or after significant change), with documented improvement actions and tracking to closure.

0 / 1 / 2 / N/A

G. SUMMARY & NEXT STEPS (OPTIONAL TEMPLATE)

After completing the questionnaire:

Transparency (T) score: ___ / ___ → ___ % → □ Low □ Medium □ High
Security (S) score: ___ / ___ → ___ % → □ Low □ Medium □ High
Reliability (R) score: ___ / ___ → ___ % → □ Low □ Medium □ High

Overall assessment (using your chosen thresholds):

- □ Fair
- □ Good
- □ Great

Top 3 improvement actions for this AI system:

1. _____
2. _____
3. _____

ENDNOTES

AI CONVERGENCE

1 GSMI 5.0: Use Cases, Foundation Models, and Key Principles for Growth: https://www.gbbc.io/uploads/reports/gsmi50/AI-&-Blockchain-Stand-Alone.pdf; GSMI 4.0 AI Convergence Foundations: https://www.gbbc.io/uploads/reports/Standalone-AI-GBBC-GSMI-4.0-Update.pdf
2 https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Use_of_artificial_intelligence_in_enterprises
3 https://www.iso.org/standard/42001
4 https://www.nist.gov/itl/ai-risk-management-framework
5 https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence
6 https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act
7 https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai
8 https://venturebeat.com/security/report-finds-82-of-open-source-software-components-inherently-risky
9 https://openssf.org/blog/2025/01/23/predictions-for-open-source-security-in-2025-ai-state-actors-and-supply-chains/
10 https://insights.sei.cmu.edu/library/introduction-to-the-octave-approach/
11 https://docs.google.com/viewer?url=https://unsceb.org/sites/default/files/2024-04/United%2520Nations%2520System%2520White%2520Paper%2520on%2520AI%2520Governance.pdf
12 https://www.chai.org
13 https://www.linkedin.com/pulse/ozempic-lawsuits-have-arrived-personal-ai-revolution-piniewski-md-wgzwc/
14 This vision can be structured through a three-zone model for processing data with open-source and decentralized AI: Zone I: the individual's "home-base" where their personal lived experience is collected and managed; Zone II a correlation engine that aggregates anonymized data contributed from many individuals to find correlations, patterns, and risk/benefit insights; Zone III is the commercial/research zone, where insights are used for broader studies, industry, or public-good applications, while preserving privacy.
15 https://www.linkedin.com/pulse/personal-data-economies-waze-future-personalized-piniewski-md-sckzc/
16 https://www.linkedin.com/pulse/world-ai-unemployment-states-must-redefine-value-piniewski-md-j6anc

GBBC