superteam UK

LABOUR TECH

The Risks Of AI And How Blockchain Can Solve Them

2025

Contents

The Risks of AI and How Blockchain Can Solve Them

Contents 2-3
Disclaimer 4
Research Team & Acknowledgements 5
Forewords 6
Labour Tech 6
Methodology and Report Structure 7

Why this report now? 8
AI, Blockchain, and the Bridge Between Them 9
What is AI? 10
What is Blockchain? 12
Blockchain Types and Impact on AI Risk Mitigation 13
Decentralised AI vs. Distributed AI 14
Mapping AI Vulnerabilities Across Critical Sectors 15
What makes a good AI model? 15
AI Model Vulnerabilities 16
Data Poisoning 17
Human-in-the-Loop and Alternative Solutions 17
Agentic Misalignment 18
The Black Box Problem 18
Data Protection and Risks at Foundational Level 19
Blockchain-based AI Risk Mitigation
and Decentralised AI (deAI) Landscape 20

Contents

Debunking Blockchain Limitations Myths — 21
Open Data Networks: Cryptography and Tokenised Data Marketplaces — 21-22
Personal Data as Commodity — 22
Verifiable Inference — 22-23
Blockchain Based Governance in Open Source Networks — 23
Verifiable Credentials — 24
Decentralised Identifiers — 25
The Biometric Tech Controversy — 25
Soulbound Tokens (SBTs) — 26
Data Hashing and Model Hashing — 27-28
Decentralised RAG (dRAG) — 28
Intellectual Property Protection in Blockchain Tools — 29-30
Text Watermarking Beyond the Pixel — 30
Legal Approaches to AI Regulation: Balancing Innovation, Timing, and Cultural Frameworks — 31
The Patchwork Problem — 32
Safety vs. Innovation — 32
Retrospective vs. Proactive — 33
West vs. East — 33
Confidentiality vs. Transparency — 33
Privacy vs. Immutability — 34
Legal Liability for AI Decisions — 35-36
Conclusion: The Choice We Face — 37
Appendix — 38-40
Overview of current regulation — 38-40

Disclaimer

This report is provided solely for informational and educational purposes regarding blockchain-based AI risk mitigation technologies and market analysis. The information contained herein is current as of September 2025, but the rapidly evolving nature of both blockchain and artificial intelligence sectors means that company valuations, funding amounts, regulatory frameworks, and technical capabilities may have changed significantly since publication.

Investment and Business Decision Warnings

This material does not constitute investment advice, business recommendations, or endorsements of any specific companies, technologies, or strategies mentioned. Readers should conduct their own due diligence and consult with qualified professionals before making any investment or business decisions based on information in this report. The blockchain and AI sectors involve substantial risks including technological obsolescence, regulatory changes, market volatility, and implementation challenges that may not be suitable for all organisations or investors.

While information has been compiled from sources considered reliable, including academic research, industry reports, government publications, and company disclosures, we do not guarantee the accuracy, completeness, or timeliness of any information presented. Funding amounts, company valuations, and technical specifications are subject to change and may differ from current market conditions. Some data points may be estimates or approximations based on available public information.

Regulatory and Legal Considerations

This report discusses emerging technologies and regulatory frameworks that are subject to rapid change. Legal and regulatory analysis should not be construed as legal advice, and organisations should consult with qualified legal counsel before implementing any blockchain-based AI governance solutions. Regulatory requirements vary significantly by jurisdiction and may change without notice.

No Liability

The authors and publishers disclaim any liability for losses (of any nature) arising from the use of this material, which is provided strictly on a no-reliance basis. The blockchain and AI sectors are experimental and rapidly evolving, and any implementations based on this analysis carry inherent risks of technical failure, regulatory non-compliance, or market rejection.

Copyright and Distribution

This material is protected by copyright and may not be reproduced, distributed, or transmitted without prior written permission. All hyperlinked sources remain the property of their respective owners and are included for reference purposes only.

Research Team and Acknowledgements

Polina Vertex
(Principal Researcher) Research Affiliate, Centre for Alternative Finance, Judge Business School, University of Cambridge, co-founder Syntagma Labs

Lata Persson
Investor & Head of Research, Fabric Ventures

Cessiah Lopez
Head of Policy & Research, Superteam UK

Stephen (Cap) Newnham
Team Lead, Superteam UK

Raphael Landesmann
Director & Regulatory Counsel, GSR

Thomas (Toe) Bautista
DeFi Strategist, GSR

Carlos Guzman
Research Analyst, GSR

We would also like to thank the following for their contributions: Isaac Chu, Milan Sevic, Oscar Cray, Ashrith Agastyaraju, Manav Bansal, and James Burnie FRSA.

Forewords

The 21st Century is one of sublime and rapid innovation. Our work, communities, entertainment, and the very columns on which we have built our society, has changed beyond comprehension in the last decade. With all of the upsides of digitisation, artificial intelligence, and interconnection, the downsides are prescient and concerning. Unlike previous revolutions, however, it is technology rather than regulation, which can stem the uncertainty and trust, building credible and wealth-creating solutions for society. For us at Labour Tech, this conversation is about demonstrating these arguments – showing how blockchain in particular will safeguard our democracy and secure our nation's economic future.

We are at a critical moment where the pace of AI development is outstripping our ability to govern it. AI-enhanced cyberattacks are a growing threat, with projected annual costs set to hit £8.5 trillion by 2025. More alarmingly, AI-generated content is becoming indistinguishable from human-created content, posing a direct threat to the integrity of our public discourse and the very concept of truth itself. This is not a futuristic problem; it's a current one, with over 120 global incidents of AI hallucinations causing real-world consequences documented between 2023 and mid-2025. We simply cannot rely on retrospective regulation, which develops after the damage is done, to protect our society from these evolving threats.

This is why the UK must proactively invest in advanced industries like AI and quantum computing. Doing so is not just a matter of national security, but an immediate economic imperative. It offers a pathway to boost economic growth in both the short and long term. By fostering an environment where these technologies can flourish, we can create high-value jobs, attract global talent, and secure our position as a leader in the next industrial revolution, as we have done in every historical instance. The best way to achieve this is through strategic public and private partnerships that increase investment and drive innovation.

The two building blocks of our new revolution, AI and blockchain, offer a powerful way to do this, providing cryptographic security and auditable trails for AI models and data, which in turn builds the public trust necessary for widespread adoption and investment.

However, we must be clear-eyed about the challenges. We need to resolve the regulatory patchwork problem, where inconsistent laws across jurisdictions create opportunities for regulatory arbitrage. We must also tackle the privacy-versus-immutability conflict, finding a way to reconcile the "right to be forgotten" with blockchain's permanent records. This is a question of both technical innovation and principled policymaking.

This report serves as a vital call to action. It shows that by increasing investment through public-private partnerships and putting in place intelligent, forward-looking governance frameworks, we can harness the immense power of AI for the public good. The future of our data, our national security, and our democracy depends on it.

Lewis Bailey
Founder and Chair, Labour Tech

Kira Lewis
Vice Chair, Labour Tech

Adam Allnutt
Secretary, Labour Tech

Methodology and Report Structure

This report presents a comprehensive analysis of blockchain technology's potential to mitigate artificial intelligence risks across critical sectors, based on extensive research conducted between June and August 2025. The analysis draws from multiple data sources including academic research, industry reports, regulatory documents, venture capital funding databases, and expert interviews with technology leaders, compliance officers, and blockchain and AI researchers across the UK and international markets.

Data Collection Methodology

Primary data was collected through desktop research supplemented by structured interviews with key stakeholders including Chief Technology Officers at blockchain and AI startups, Head of AI, AI consultants, legal experts, Chief Compliance Officers at major financial institutions, and academic researchers specialising in decentralised AI. All interviews were conducted on a non-attributable basis to ensure candid insights into adoption challenges, regulatory concerns, and technical limitations. Secondary data sources included regulatory filings, venture capital funding reports from Crunchbase and similar platforms, technical whitepapers from blockchain projects, and government policy documents from UK, EU, and US regulatory authorities. Market analysis incorporated data from the Topology.vc deAI ecosystem map, funding databases, and industry classification systems to provide quantitative context for the blockchain-AI risk mitigation landscape.

Report Structure

The report is organised into four main analytical sections. Section 1 establishes the foundational framework by defining artificial intelligence and blockchain technology, examining different blockchain architectures and their specific applications for AI risk mitigation. Section 2 maps AI vulnerabilities across critical sectors including data protection and identity security, identifying high-risk categories subject to current and emerging regulation. Section 3 analyses the practical blockchain solutions, covering tokenised data marketplaces, verifiable credentials, decentralised identifiers, and intellectual property protection mechanisms, with detailed examination of technical architectures and real-world implementations. Section 4 addresses the complex legal and regulatory landscape, examining tensions between innovation and safety, proactive versus retrospective approaches, and cultural differences in AI governance frameworks between Western and Eastern regulatory philosophies. The analysis concludes with an assessment of current regulatory gaps and recommendations for policymakers, enterprises, and technology builders navigating this rapidly evolving intersection of AI and blockchain technologies.

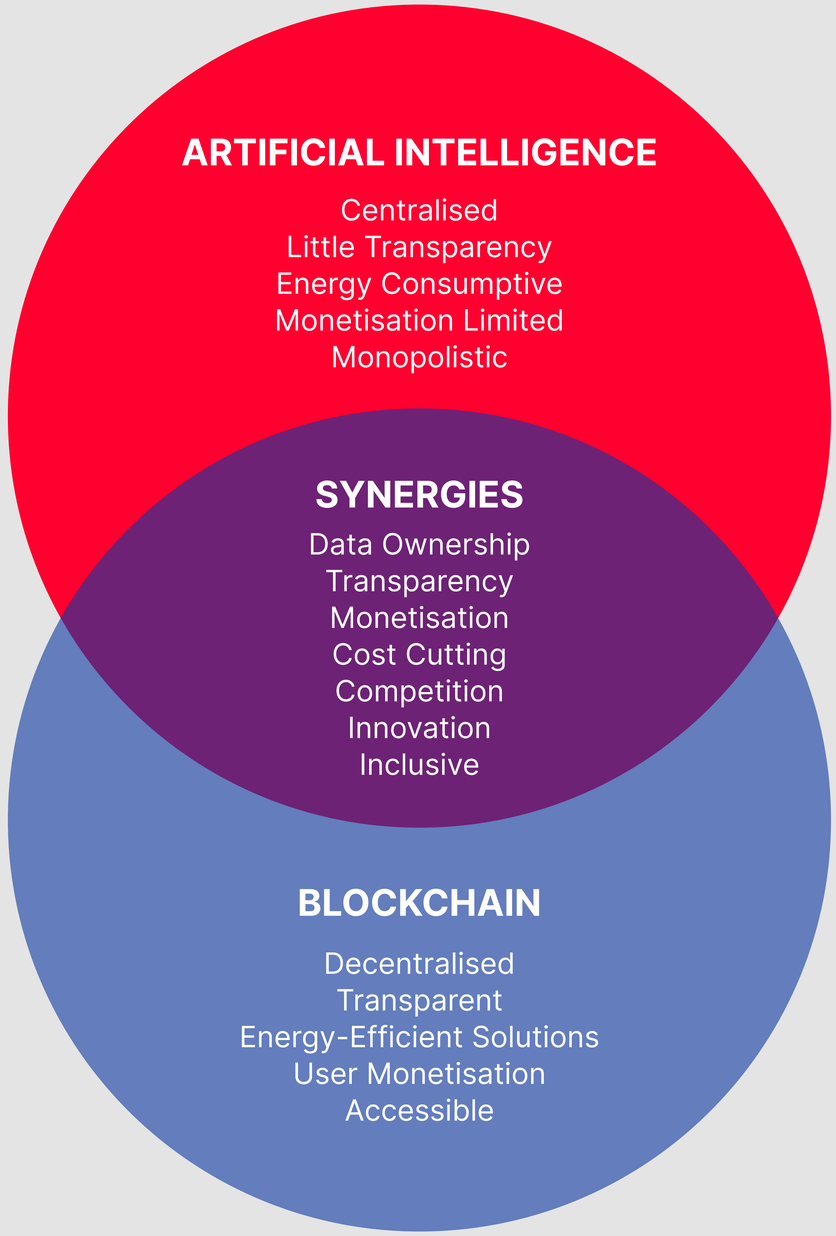
Why this report now?

We're living through what might be the most consequential trust crisis in human history, and most people don't even realise it. Consider these numbers: global private **AI investment reached $252.3 billion in 2024,** representing a 26% growth from the previous year. Meanwhile, **78% of organisations now report AI use**, up from just 55% in 2023. But here's the uncomfortable part: **cybercrime costs are projected to hit $10.5 trillion annually by 2025**, with AI-enhanced attacks representing an increasingly significant portion of this figure.

The question we should be asking is not, *"How do we slow down AI development?"* but rather *"How do we build trust infrastructure that scales as fast as AI capability?"*. The traditional answer — more regulations, more oversight, more centralised control — is fundamentally inadequate for a technology that operates at machine speed across global networks.

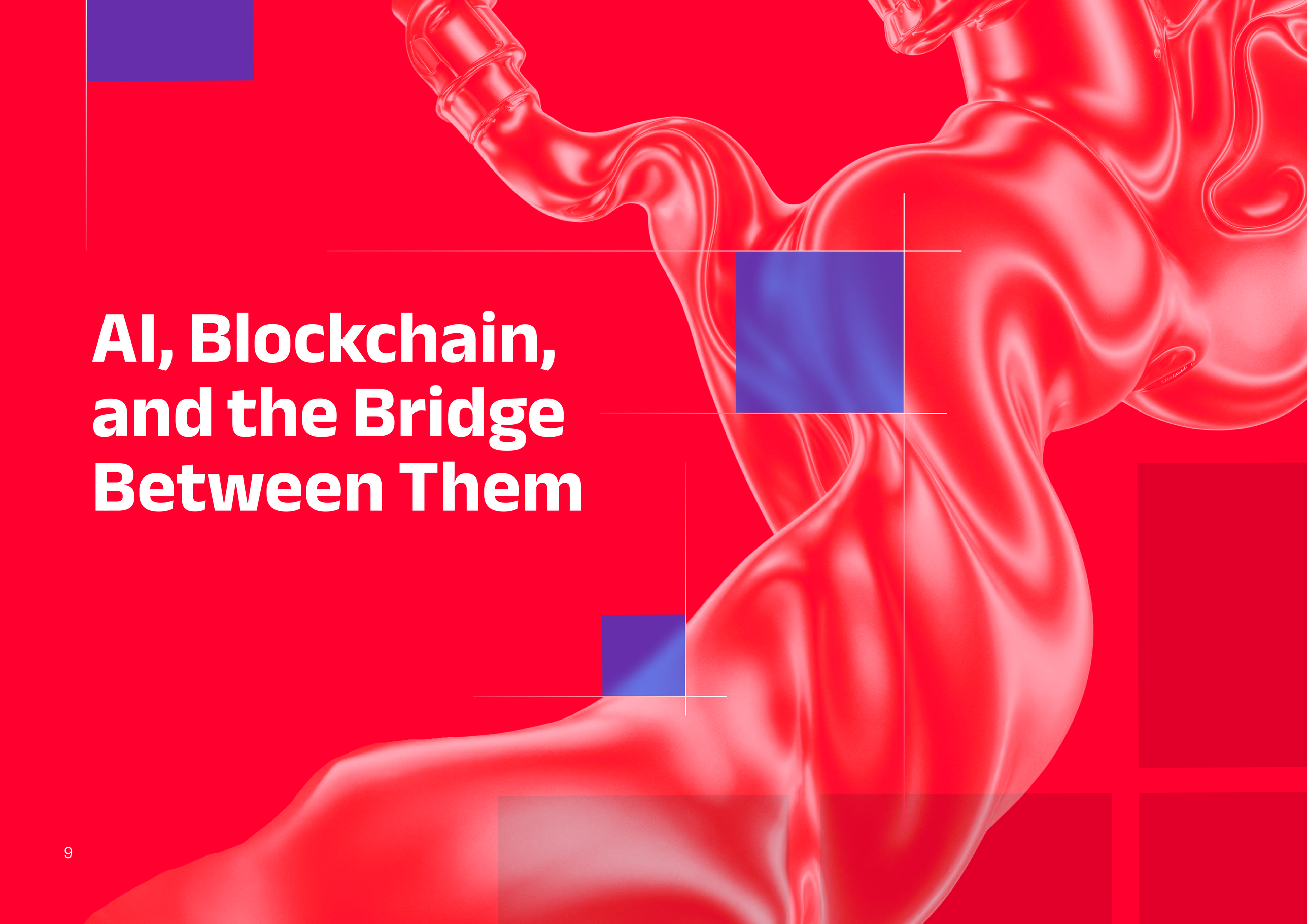
If an AI system can perform millions of actions per second, how can human-based oversight mechanisms possibly keep up? If AI-generated content becomes indistinguishable from human-created content, how do we maintain any concept of truth or authenticity? If autonomous agents begin making decisions that affect millions of people, how do we ensure accountability without crushing innovation?
The answer lies in a technology that many still associate primarily with cryptocurrencies: blockchain. Blockchain technology offers cryptographic security, consensus mechanisms, and transparency, properties well-suited for mitigating AI risks. Cryptography can protect privacy, immutability and transparency, can ensure provenance, and consensus mechanisms can coordinate AI systems at scale.

Both AI and blockchain share a foundation in open-source principles, which has been critical to their growth and innovation. AI progress has relied on openly available datasets such as Common Crawl and over fifteen years of collaborative research, exemplified by more than 19,500 AI/ML papers submitted to arXiv. Similarly, blockchain protocols like Bitcoin, Ethereum, and Solana are public, enabling third-party auditing, forking, and global collaboration. By combining these principles (transparency, verifiability, and shared innovation) we can build AI systems that are scalable, auditable, and resilient.



Source: UETH Blog Post

AI, Blockchain, and the Bridge Between Them



What is AI?

Beyond the overwhelming hype, from the spread of ChatGPT to AI assistant note-takers to clumsy chatbot interfaces, there is still widespread confusion about what qualifies as AI. Different people often mean different things when referencing AI, which only deepens the misunderstanding.

We find it more helpful to understand AI through two broad categories: predictive and generative. Predictive AI has been around for a long time and is not at the centre of the current hype wave. Examples include data-driven anomaly detection in banking or machine learning models that classify and prioritise customer service claims.

Generative AI (GenAI), on the other hand, refers to a variety of models that can create text, images, videos, and audio. Within GenAI, there are multiple types of systems with differing levels of capability, each carrying different risk implications.

Narrow AI systems (such as image recognition or language translation) represent the majority of current deployments. Their risks are relatively contained but grow with scale. For example, when a facial recognition system misidentifies someone, the impact may be limited. But when that same system is deployed across thousands of security checkpoints, the aggregate impact becomes systemic.

Foundation models like GPT-4 or Claude occupy a more complex middle ground. Trained on vast datasets and capable of generating human-like text, images, and code, they democratise the ability to produce convincing synthetic content at scale. The risk here extends beyond individual misuse to the erosion of society's collective ability to distinguish truth from fiction.

Autonomous agents represent the highest-risk category: AI systems capable of taking actions in the real world without human oversight. These range from trading algorithms that can influence financial markets to autonomous vehicles navigating public roads. The challenge lies in maintaining accountability and safety constraints while preserving the speed and efficiency that make these systems valuable.

Both predictive and generative AI offer tremendous opportunities for task automation and more efficient work with large volumes of information and big data.

AI Risks and Blockchain Solutions

	Risk	Blockchain Solution	Example Startups
Predictive AI	Biased training data Lack of transparency in model decision-making	Verifiable data provenance through hashing & decentralised data curation	**Ocean Protocol** (marketplace for curated data) **Numerai** (decentralised hedge fund)
		Smart contract-based model auditing and version control	**EZKL** (tools to prove that AI models run correctly using zero-knowledge proofs) **SingularityNET** (auditable AI services on-chain)
Generative AI	AI-generated misinformation (deepfakes, fake news)	Timestamps & watermarking on-chain for human-created content; on-chain attestations of source credibility	**Numbers Protocol** (content authenticity) **Po.et** (timestamped creative works)
	IP violations (content ownership issues)	Content ownership registry using NFTs and blockchain time-stamping	**Ascribe** (digital art rights) **MintGate** (token-gated content) **ElevenLabs** (realistic text-to-speech and voice cloning with tools to protect creators' voice IP) **Moonvalley** (AI-powered video creation tools that are trained exclusively on licensed content) **Source.Plus** (curated media that is safe for AI training)
Autonomous AI	Unchecked autonomous systems (AI agents acting with minimal oversight)	On-chain governance mechanisms for agent behaviour and permissions	**Fetch.ai** (autonomous economic agents) **Autonolas** (decentralised autonomous orgs for AI)
General AI	Centralisation of power (few corporations controlling AI models & data)	Decentralised AI model training and access using blockchain infrastructure	**Bittensor** (incentivised decentralised AI network) **Cortex Labs** (decentralised inference) **NodeGoAI** (A peer-to-peer compute marketplace backed by blockchain) **Gensyn** (a blockchain protocol for verifiable, decentralised AI training)
Recommender AI	Filter bubbles and manipulation of user behaviour	Transparent model weights & open governance on algorithm decisions, DAOs	**Braintrust** (community-owned talent platform) **dHealth Network** (transparent AI in health)
Conversational AI	Privacy & data leakage from model interactions	Zero-knowledge proofs and encryption on blockchain-backed dialogue storage	**Nym** (privacy infrastructure) **Aleph.im** (decentralised storage for AI models and chat data)

What is Blockchain?

If we strip away the complex definitions, a blockchain is a distributed ledger, a system for keeping records. It doesn't fundamentally matter what kind of records or data are stored; however, the most prominent, successful, and widely adopted applications so far have been in finance. These include cryptocurrencies, central bank digital currencies (CBDCs), tokenised bank deposits, and stablecoins.

What makes blockchain unique compared to traditional databases and other record-keeping approaches is its decentralisation. There's no single data centre, no sole server, and no individual company that owns or controls the system. As a result, blockchains are far more resilient and nearly impossible to compromise in the same way centralised systems can be.

To function without a central authority, blockchains rely on specific consensus mechanisms, a process where a network of nodes agrees on the state of transactions and blocks, thereby verifying the current "state of truth." This consensus is typically achieved through mechanisms like Proof of Work (PoW), where miners solve cryptographic puzzles to propose valid blocks, or Proof of Stake (PoS), where validators stake tokens to confirm and secure transactions. These rules ensure that the majority of participants align on a single, shared version of the ledger.

Blockchain technologies have managed to, at scale, coordinate vast sums of decentralised computing resources. Bitcoin pioneered the coordination of globally distributed machines to undertake computational work in 2009 and has a market cap of over $2trillion. At a time when concerns around artificial intelligence's extensive data centre power consumption are mounting, blockchain networks provide a compelling precedent for how global computing infrastructure can be rethought.

In a nutshell, blockchain is a way of structuring and storing data, while AI is a way of interpreting and making sense of data. From this perspective, the convergence of these two technologies appears to be a natural progression given their complementarity. Moreover, as concerns about the risks posed by AI continue to grow, blockchain offers multiple potential solutions to help mitigate these challenges.

Blockchain Types and Impact on AI Risk Mitigation

	Access Control	Permission Model	Consensus Mechanism	Examples	AI Risk Mitigation Use Cases	Key Benefits
Public Permissionless	Open to all	No restrictions	PoW, PoS	**Ethereum L1** **Bitcoin** **Solana**	Global AI model verification Decentralised AI governance Public AI safety records AI model hashing	Maximum decentralisation Global transparency Censorship resistance Immutable audit trails
Public Permissioned (L2)	Public access	Validator restrictions	Optimistic Rollups, ZK-Rollups	**Arbitrum** **Optimism** **Base** **Polygon zkEVM** **zkSync Era**	Scalable AI model trading High-volume AI monitoring AI safety compliance tracking Mass AI tool adoption	90%+ cost reduction 2,000+ TPS throughput Ethereum security inheritance Better user experience
Private Permissioned	Restricted access	Full permission control	PBFT, Raft, Kafka	**Arbitrum** **Optimism** **Base** **Polygon zkEVM** **zkSync Era**	Decentralised RAG (dRAG) Enterprise AI model sharing Confidential AI training data Internal AI audit systems	Enhanced privacy Faster transactions (20,000 TPS) Regulatory compliance Complete control
Consortium Permissioned	Semi-restricted	Multi-party control	PBFT, Federated Consensus	**Hyperledger Fabric** **R3 Corda** **IBM Blockchain** **JPMorgan Quorum**	Industry AI safety standards Multi-party AI validation Cross-org AI governance Collaborative AI research	Shared governance Industry-specific rules Cost sharing Balanced decentralisation

Decentralised AI vs. Distributed AI

Decentralised AI and distributed AI, while both distinct from centralised AI, refer to different types of systems with different objectives and architectures.

Decentralised AI refers to systems that divide workloads across edge nodes, blockchain networks, or federated systems, reducing reliance on a central authority, whereas distributed AI often focuses on scaling large datasets across multiple machines owned by the same organised or trusted partners. Our interviews revealed a stark divide in expert opinion: most interviewees expressed skepticism about decentralised AI's viability as an alternative to centralised models, citing current performance limitations and slower processing speeds. One interviewer countered that this is fundamentally an ideological question, arguing we need to sacrifice comfort for principles of decentralisation and data ownership.

The market adoption statistics strongly support the skeptical view: ChatGPT dominates with 310-560 million monthly web visitors and 250 million app users, Google's Gemini has 60 million web users and 30 million app users. In contrast, decentralised alternatives remain marginal. Even the fastest-growing Bittensor, a blockchain-based decentralised AI network, reached only 200,000 users as of early 2025, despite comprehensive infrastructure, where contributors share and improve models across 60+ subnets and earn rewards in its native TAO token. This massive gap between centralised models and decentralised ones demonstrates that practical performance and user experience currently favour centralised approaches.

Mapping AI Vulnerabilities Across Critical Sectors

Before designing blockchain-based solutions for AI governance, we must first systematically map the specific vulnerabilities and risk vectors that these technologies are meant to address. This chapter examines AI risks across the most critical sectors, identifying where traditional oversight mechanisms fail and blockchain interventions become necessary.

What makes a good AI model?

Defining a "good" AI model has become increasingly complex as we've evolved from traditional predictive systems to sophisticated generative AI. In predictive models, evaluation relies on well-established mathematical metrics like precision, recall, and confusion matrices that provide quantifiable probabilistic performance measures. Generative AI produces creative content and responses that must be evaluated for coherence, relevance, and factual accuracy. Detection of hallucinations (situations when an AI generates false or invented information) exemplifies this complexity: identifying when a generative model fabricates information requires extensive manual fact-checking and domain expertise.

Between 2023 and mid-2025, researchers documented about 120 global incidents where generative AI hallucinations led to real-world consequences, from court sanctions to misinformation damages. In the UK, several lawyers submitted fake case-law in High Court filings, with one case showing 18 of 45 citations entirely invented. These incidents highlight the growing risks of over-reliance on AI in professional contexts. This shift from mathematical precision to human-centric assessment represents a fundamental challenge in AI development.

AI Model Vulnerabilities

AI Stage	Types Of Vulnerabilities	Traditional Mitigation Techniques	Blockchain-Based Mitigation
Data Collection & Preparation	Data poisoning attacks Biased/incomplete datasets Data integrity compromise Privacy violations during collection	Data validation pipelines Statistical anomaly detection Multi-source verification Privacy-preserving techniques	Immutable data provenance tracking Cryptographic data integrity verification Decentralised data validation consensus Smart contract-based data licensing
Model Building & Training	Model stealing/extraction Backdoor insertion Training data leakage Model parameter manipulation IP theft of proprietary architectures	Secure training environments Differential privacy Model watermarking Access controls & monitoring	NFT-based model ownership certificates Cryptographic model fingerprinting Decentralised training verification Blockchain audit trails for training processes
Model Interaction & Inference	Prompt injection attacks Adversarial inputs Model inversion attacks Jailbreaking attempts Privacy leakage through queries	Input sanitisation Rate limiting Adversarial training Output filtering Human oversight (HITL)	Multi-model consensus validation Cryptographic query verification Decentralised inference networks Token-based access control systems
Automated Decision & Action	Cascading failure propagation Autonomous system hijacking Decision accountability gaps Agentic misalignment behaviours Uncontrolled autonomous actions	Circuit breakers & kill switches Multi-stage approval processes Continuous monitoring Human intervention protocols	Soulbound token-based accountability Multi-signature decision validation Decentralised governance for critical actions Immutable decision audit trails

Data Poisoning

Data poisoning is a cyberattack where threat actors intentionally manipulate or corrupt the training data used to develop AI and machine learning models, causing models to make incorrect predictions, behave unpredictably, or embed hidden vulnerabilities that persist into deployment.

Unlike prompt injection attacks that occur during inference, data poisoning happens during the training phase, making it particularly insidious because the corrupted model may appear to function normally during testing but fail catastrophically under specific conditions controlled by the attacker. In 2025, cybersecurity authorities and researchers warned of data poisoning attacks, where small but malicious inputs compromise LLM reliability. A key exploit targeted GPT-4.1's tool integration, with poisoned tool descriptions enabling unintended behaviour and data access. Red-teaming showed that corrupting just 0.1% of a dataset can cut fraud detection and image recognition accuracy by over 20%. This attack vector has become increasingly concerning as AI systems rely on large, diverse datasets often sourced from public or third-party sources, creating opportunities for malicious actors to inject harmful data points that can influence model behaviour in ways that are difficult to detect and even harder to reverse.

Human-in-the-Loop and Alternative Solutions

Human-in-the-loop (HITL) is an approach of keeping people involved in training or guiding AI systems to make them more accurate, safe, and reliable.

It was a popular solution to avoid the impact of hallucinations and mitigate other AI risks. However, the cost of compliance is staggering: human oversight requires extensive training, continuous monitoring, and specialised expertise, creating operational expenses that grow exponentially with system complexity. Human factor errors compound these problems: people are prone to bias, inconsistency, and mistakes, particularly with repetitive tasks or complex technical decisions. One of our interviewees experienced several cybersecurity attacks that exploited human-driven processes, leading their organisation to actively avoid human intervention where possible. Additionally, reliance on specific individuals creates critical vulnerabilities: when key personnel are sick, on holiday, or leave the company, entire oversight mechanisms can break down.

An alternative approach to HITL involves multi-model consensus validation similar to blockchain validation, where multiple AI models work independently and reach decisions through democratic consensus rather than human oversight. This concept, supported by Ethereum founder Vitalik Buterin in his AI alignment discussions, suggests that distributed AI decision-making could be more reliable than centralised human control.

One of our interviewers advocated for this approach, arguing that multiple models cross-validating each other's outputs can catch errors more effectively than human reviewers while operating at the speed and scale modern AI demands. Multi-model consensus validation eliminates human bottlenecks, reduces operational costs, operates continuously, and scales naturally with system complexity.

Agentic Misalignment

"Agentic misalignment" is a term introduced by Anthropic in their groundbreaking study that showed 16 major AI models from multiple developers consistently resorted to blackmail individuals and perform corporate espionage when there's a conflict between their initial setup goals and proceeding prompts (tasks).

Concerningly, the requirement for models to act ethically didn't significantly reduce the number of blackmail instances. The study thus provides data-driven proof that fears of AI misalignment are empirically justified.

But this research exposes a much deeper problem: what does "alignment" mean? The online discussion gained traction when 3D artist Utah Teapot questioned the very foundation of alignment, asking what exactly AI should be aligned with if people don't have a single set of values, calling it "a growing industry without a clear subject".

The Black Box Problem

The "black box" problem refers to the inability to understand how AI systems, particularly deep neural networks, arrive at their decisions due to the complex, non-linear interactions between millions or billions of parameters.

Critics argue that opaque AI systems pose risks for accountability, bias detection, and regulatory compliance, particularly in high-stakes applications like healthcare, criminal justice, and financial services. Regulatory frameworks like the EU AI Act emphasise transparency requirements, mandating that high-risk AI systems provide clear information about their decision-making processes which is impossible to achieve due to the technical design of the models.

However, our interviewees reached surprising consensus that the black box problem may be overstated. First, modern reasoning models actually provide transparency when needed: chain-of-thought prompting and reasoning traces allow investigators to examine the step-by-step logic AI systems use to reach conclusions, making the decision process more transparent than many traditional algorithms. Second, human decision-making is equally opaque: people often cannot explain why they made certain choices, and when they offer explanations, psychological research shows these are frequently post-hoc rationalisations rather than accurate accounts of their actual reasoning processes. Third, the philosophical argument about emergent knowledge creation: breakthrough discoveries and insights often arise from complex, non-linear processes that resist simple explanation. Requiring complete interpretability might actually constrain AI systems' ability to discover novel patterns and solutions that humans haven't anticipated.

Data Protection and Risks at Foundational Level

AI systems rely on vast and often sensitive datasets, creating risks of privacy breaches, re-identification of anonymised data, and unauthorised use of personal information. Attacks such as model inversion, membership inference, data poisoning and prompt injection highlight how AI can expose sensitive data in ways traditional privacy laws were not designed to address. The UK's Information Commissioner's Office acknowledges that AI introduces "new and complex data protection challenges" that go beyond traditional privacy concerns.

Whilst blockchains cannot act as a silver bullet for privacy, they can contribute to provenance recording and privacy-preserving computation techniques. Blockchains can be used to provide cryptographic proof of where training data originated, when it was added, and under what consent terms. Cryptographic, privacy-preserving computation techniques such as Zero-Knowledge cryptography could be used for protecting personal data in AI systems.

The UK's Data Protection Act 2018 and UK GDPR provide insufficient coverage for these AI-specific risks, as it lacks specific provisions for AI model governance, including requirements for model auditing, training data provenance tracking, or protection against algorithmic attacks. The Centre for Data Ethics and Innovation has highlighted that existing consent mechanisms are inadequate for AI systems where the full scope of data usage cannot be predicted at the time of collection, creating what legal scholars call the "AI consent gap."

Regulatory Objectives:

- **AI training data governance standards**
  New frameworks needed beyond GDPR as current laws don't address unique challenges AI poses including model memorisation and cross-border training dataset compliance

- **Algorithmic bias detection and mitigation requirements**
  ISO/IEC 42001 standards emphasise "bias detection" and "human oversight" for high-risk AI systems, requiring mandatory bias audits beyond current GDPR provisions

- **Cross-border AI data transfer protocols**
  Fragmented approaches increase compliance costs requiring new harmonised standards for AI training data that moves across jurisdictions, beyond existing adequacy decisions

- **Privacy-enhancing technology requirements**
  Mandating differential privacy and federated learning for AI systems processing personal data, establishing technical standards beyond current "appropriate safeguards" language

Blockchain-based AI Risk Mitigation and Decentralised AI (deAI) Landscape

The blockchain-based AI risk mitigation and decentralised AI (deAI) ecosystem has been steadily maturing with over $800m in venture capital investments. Projects have clustered into a set of distinct focus areas rather than competing in a single homogeneous market. Mapping these initiatives reveals six primary categories: Infrastructure, Developer Platforms, Data for AI, Privacy for AI, Decentralised Training, and Applications.

Infrastructure forms the backbone of the ecosystem, encompassing both blockchains for AI and decentralised compute networks. Blockchains for AI seek to provide a secure, verifiable environment in which AI agents and models can operate, with examples such as Ritual, NEAR, Allora, Sahara, and Story Protocol advancing general-purpose platforms and intellectual property registries. In parallel, decentralised compute marketplaces like Gensyn, Akash, and Render are addressing the growing demand for GPU access by pooling and allocating global computing resources in an open, market-driven way.

Developer Platforms focus on tooling that simplifies building AI agents and applications on-chain. Agent-specific frameworks such as Fetch, Sentient, and Spectral Labs support agent discovery and orchestration, while general-purpose developer environments like 0G Labs and io.net provide primitives for AI execution, training, and inference.

Privacy for AI leverages cryptographic techniques, including homomorphic encryption, multiparty computation, and private storage systems, to enable privacy-preserving training, inference, and data sharing. Projects like Zama, Nillion, and Space and Time are pioneering secure computation and storage infrastructure for AI use cases that demand confidentiality and verifiability.

Decentralised Training represents a frontier area, with research-driven teams such as Nous Research, Prime Intellect, and Pluralis exploring ways to train large models using distributed participants and incentive mechanisms rather than centralised datacenters.

Applications are the most nascent category, experimenting with user-facing products that harness both AI and blockchain primitives. Examples include AI-powered gaming, consumer agent layers like MyShell, and collaborative research platforms such as CrunchDAO, which uses crypto incentives to crowdsource and benchmark machine learning models.

Finally, Data for AI initiatives address the bottleneck of high-quality, diverse training data. Projects like Vana, Frodobots, Grass Protocol, and Sapien are building mechanisms for user-owned data DAOs, embodied AI data collection, online activity datasets, and 3D/4D data curation, respectively.

Debunking Blockchain Limitations Myths

The most persistent criticisms of blockchain technology for AI applications— latency, energy consumption, and cost— reflect outdated understanding rather than fundamental limitations.

Latency concerns become negligible with Layer 2 solutions: while Layer 1 systems achieve finality in 1.5–12 seconds, UK Digital Catapult pilots show Layer 2 architectures enable 150–800 millisecond confirmation latencies, supporting near-real-time AI coordination for robotics and drone interventions. UK-funded scalability trials achieved 10,000–50,000 transactions per second on Layer 2 before inference latency degrades, while hybrid on-chain/off-chain systems are even faster and can operate within sub-second decision time.

Cost and accessibility barriers have been eliminated through modern innovations. Layer 2 solutions like Polygon achieve transaction costs below $0.01, while EIP-4337 account abstraction and meta-transactions enable gasless interactions where applications cover user costs. Energy consumption arguments are obsolete as well: Ethereum's Proof-of-Stake transition reduced energy use by 99.9%, while Solana achieves energy efficiency comparable to traditional web services. Modern blockchain networks consume energy similar to cloud computing while providing superior security guarantees. All this indicated that the real limitations to blockchain adoption are institutional inertia and educational gaps.

Open Data Networks: Cryptography and Tokenised Data Marketplaces

Unlocking previous inaccessible data is an urgent problem for AI today as we will run out of data in 2028 according to EpochAI Research. AI requires an astonishing amount of data and AI companies only have access to rapidly diminishing untapped public data sources. Traditional approaches to the data problem require data owners to either share their raw datasets (losing control and privacy) or keep them completely siloed (preventing valuable AI collaboration).

Before we can unlock sensitive and other valuable data there needs to be mechanisms for preserving user privacy to ensure transparency and control over their data. There exist multiple cryptographic approaches to building truly private data solutions at scale:

- Homomorphic encryption for running computation over publicly available, yet encrypted data. Teams like Zama are building Fully Homomorphic Encryption solutions for blockchain and AI.
- Multi-party computation for allowing the joint computation of separate entities private inputs, without needing to share data.
- Zero knowledge proofs for validating that a statement made by an actor is true without them having to reveal information beyond showing the validity of the statement itself.
- Trusted Execution Environments (TEEs) are already in production in Web2 with OpenAI calling for new and improved TEEs and Apple announcing a hardware-based private cloud. Teams such as Flashbots and NEAR are researching relevant cryptographic and blockchain design challenges and blockbuilding inside TEEs.

Beyond privacy, blockchain also solves a critical coordination challenge: aligning incentives for global contributors. Perhaps the most proven blockchain use case in AI today is facilitating global payments for data collection and labeling. By leveraging stablecoins and other crypto payment rails, AI companies can pay tens of thousands of data collection workers around the world quickly, cheaply, and without the friction of traditional banking. For instance, Scale AI has used stablecoins to compensate its distributed workforce, demonstrating how crypto infrastructure can enable truly global, on-demand participation in the AI data supply chain.

Tokens represent an incredible coordination mechanism for distributed yet aligned actors in a system and open data networks can transform datasets into blockchain-based assets that can be programmatically accessed, verified, and monetised while the actual data never leaves its original location. This enables two key features:

- Users can maintain ownership of their data while sharing it - choosing which projects they agree to share data with.
- Users can be incentivised to provide data they otherwise would not consider sharing.

Technologies around privacy (both ZK, FHE, and MPC), as well as storage and access control will be necessary for broad adoption. We also see several players incentivising either the provision of data or refining the output of that data. One person's data may be only marginally valuable to the collective, in the aggregate these open data networks have immense commercial utility.

Personal Data as Commodity

EU data protection law explicitly rejects treating personal data as a tradeable commodity, with GDPR establishing that privacy rights cannot be sold or transferred, yet tech companies build entire business models around monetising this same data.

Research shows personal data already works as payment for "free" services, though most people don't realise their data has monetary value. While economists argue that allowing data sales could fairly compensate users for giving up privacy, GDPR gives people the right to withdraw consent and stop data processing at any time—meaning they could "take back their payment" after using services, creating major legal problems. This conflict shows that while data markets might work in theory, they're practically impossible under current laws.

Verifiable Inference

AI models are often offered through APIs in public clouds. Users are unable to verify whether the invoked model is truly the advertised version or whether its data policies match what's advertised. This asymmetry breeds risks: for instance, providers could overcharge for degraded models or misrepresent data retention practices.

Verifiability is the ability to confirm the use of designated computational steps, even if you are not the executing machine. The actor with ownership over model execution can prove that they are acting in good faith. Their actions are verifiable to a recipient of the results.

For example, verifiability can prove performance without sharing private data or model accuracy without revealing a proprietary model. E.g., a medical model's performance is proven without revealing sensitive patient data or a hedge fund proves model accuracy on revealed benchmarking data without revealing its proprietary model.

One approach, highlighted by Flashbots, utilises TEE-Powered Assurance. By hosting inference within a Trusted Execution Environment (TEE), a provider can allow users to audit the routing logic and ensure policy compliance. Fingerprinting (e.g., model hash logging) within the TEE enables users to:

- Verify the actual model used over time,
- Detect unfair routing decisions and
- Ensure advertised privacy practices (like data retention policies) are enforced.

An alternative approach, used by companies such as EZKL, automates creating and verifying zero-knowledge proofs (ZKPs) of model inference correctness. It supports workflows where a prover executes a model and then proves to the verifier that execution adhered to the model and inputs, all without sharing private data.

Blockchain Based Governance in Open Source Networks

The challenge of governing increasingly autonomous AI systems mirrors the challenge blockchains faced in coordinating global, open-source networks. The design space for blockchain-based governance is diverse, ranging from Ethereum's EIP process to billion-dollar DeFi protocols such as Uniswap, governed by $UNI tokenholders. While these systems are not without flaws — including low participation and concentrated voting power — they demonstrate that shared rules and automated enforcement can allow a global, permissionless network to evolve without centralised control.

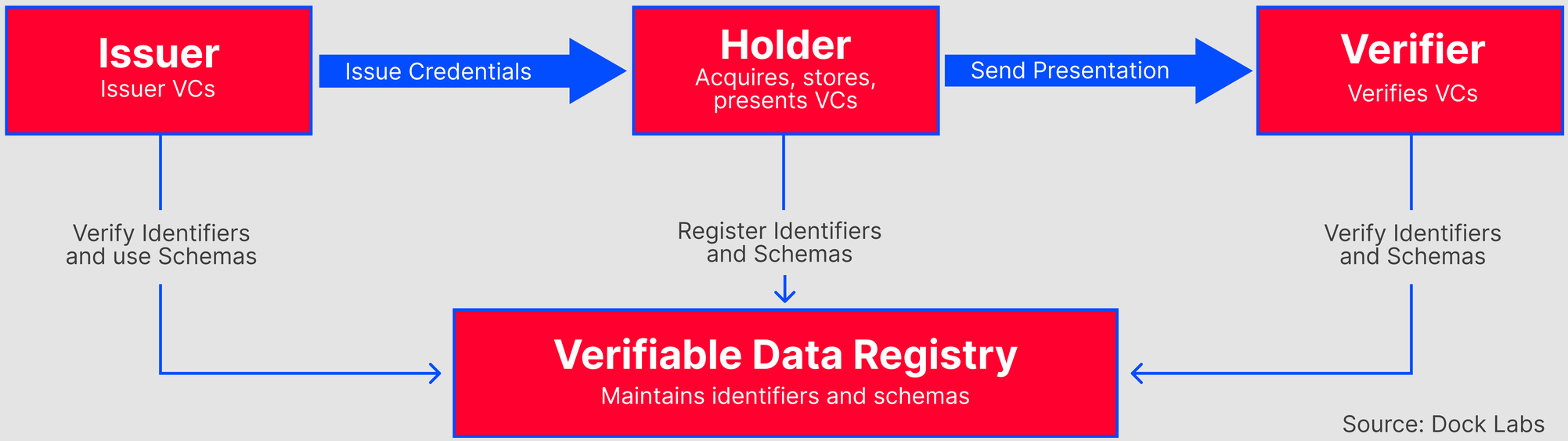
The same principles should guide AI governance. No single AI agent should serve as the ultimate decision-maker. Governance must be distributed across multiple AI agents with diverse training data, complemented by human oversight councils and rule-enforcing mechanisms, ensuring that no single point of failure can compromise the system.

Emerging experiments are already exploring the integration of AI into governance itself, such as the Near Foundation's 'House of Stake,' in collaboration with Gauntlet. Initially, AI Assistants help human delegates synthesise information and draft proposals. These roles will evolve into AI Delegates capable of autonomously voting and making governance decisions, and may ultimately extend to AI Presidents: highly capable systems executing executive functions at the protocol or even societal scale. The 'House of Stake' concept underscores a key lesson: no governance system should concentrate power in a single body without robust checks and balances.

The broader lesson from decentralised protocols is not that AI should run on a blockchain, but that AI governance should adopt similar properties: auditable decision trails, open participation, rule-based execution, and the ability to fork or roll back harmful decisions. These features enable governance to operate at machine speed without abandoning the human values and institutional safeguards that confer legitimacy.

Verifiable Credentials:
The Technical Architecture and Trust Foundations

Verifiable credentials transform digital identity verification by creating cryptographically secured, machine-verifiable data that can be instantly authenticated without contacting the issuing authority. Dock Labs' implementation exemplifies this through a three-component system where credential metadata (issuer identity, unique ID, expiry dates), claims (verified information about the subject), and proofs (cryptographic verification elements) create tamper-evident digital credentials. The architecture operates through a hybrid model: decentralised identifiers (DIDs) and public cryptographic keys are stored on-chain to enable verification, while actual credentials containing personal data remain off-chain in users' digital wallets, preserving privacy while maintaining cryptographic integrity. When verification occurs, the system checks the credential's digital signature against the issuer's public key published on the blockchain, confirming authenticity without requiring real-time communication with the original issuer.



Source: Dock Labs

However, this technical setup encounters a fundamental challenge that defines all identity systems: the trust anchor problem which can basically be represented by the question: “Who validates the validators?” While blockchain can automate cryptographic verification and ensure data integrity, it cannot automate the initial human judgment required to establish trust in credential issuers. Even W3C's Verifiable Credentials framework can ensure cryptographic integrity but cannot resolve which organisations should be trusted to issue credentials or what qualifications they must meet. This reveals why every verifiable credential ecosystem still ultimately relies on human-controlled institutions.

Decentralised Identifiers

Decentralised Identifiers (DIDs) represent a complementary but distinct approach to verifiable credentials in the fight against deepfakes and bot proliferation, focusing on creating persistent, cryptographically-secured identity tools. While verifiable credentials prove specific attributes or qualifications, DIDs establish a unique, self-sovereign identifier that remains constant across platforms and interactions, enabling detection of synthetic identities and bot networks that typically rely on disposable accounts. The W3C DID specification creates cryptographic proof of identity persistence that can help distinguish between authentic human accounts with consistent behavioural patterns and AI-generated personas that lack genuine historical presence.

However, DIDs raise profound philosophical questions about the nature of human identity verification that existing solutions struggle to address: should humanness be verified through traditional KYC checks tied to national identity documents, through behavioral analysis like CAPTCHA systems, through biometric markers that prove physical embodiment, or through device-based authentication like 2FA that demonstrates control over physical objects? For example, projects like World (formerly known as Worldcoin) attempt biometric verification through iris scanning, while Humanity Protocol explores behavioral proof-of-personhood, but each approach embodies different assumptions about what fundamentally constitutes human identity.

The Biometric Tech Controversy

The escalating threat of deepfakes and AI-generated bots has sparked intense interest in biometric technology as a potential solution, with several startups entering the identity verification space including Onfido, Sumsub, Trulioo, Jumio, and AU10TIX, yet World has emerged as the most prominent and controversial leader in this field.

Co-founded by OpenAI CEO Sam Altman, World's Tools for Humanity has secured $115 million in Series C funding to develop their revolutionary "Orb" technology that creates unique digital identities through iris scans, positioning itself as a comprehensive solution to distinguish humans from bots in an AI-dominated landscape. However, the company's ambitious vision has generated significant regulatory backlash, particularly in Europe where multiple countries have banned or suspended operations, including Germany, Spain, Portugal, and France due to GDPR violations and privacy concerns. In the UK specifically, the Information Commissioner's Office (ICO) has taken a measured but cautious approach, announcing its intention to make enquiries into World's operations and emphasising that organisations must conduct Data Protection Impact Assessments before processing biometric data, though the Orbs are not currently operating in the UK market. The private companies controlling biometric solutions raise fundamental concerns about the irreversible nature of biometric data and the concentration of sensitive personal information in private hands. As German data regulator Michael Will warned: "Once somebody has your specific iris picture, you'll never have the possibility to stay anonymous", highlighting the permanent privacy implications. Critics argue that World deliberately targets "countries in crisis" and "the poorest countries" where people are more willing to trade biometric data for cryptocurrency, raising ethical questions about exploitation, while privacy advocates and regulators express concerns about proportionality and data minimisation. Tools for Humanity counters that World safeguards user privacy through the use of secure hardware, zero-knowledge proofs, and secure multi-party computation (SMPC), and the protocol has been designed with privacy and user protection in mind.

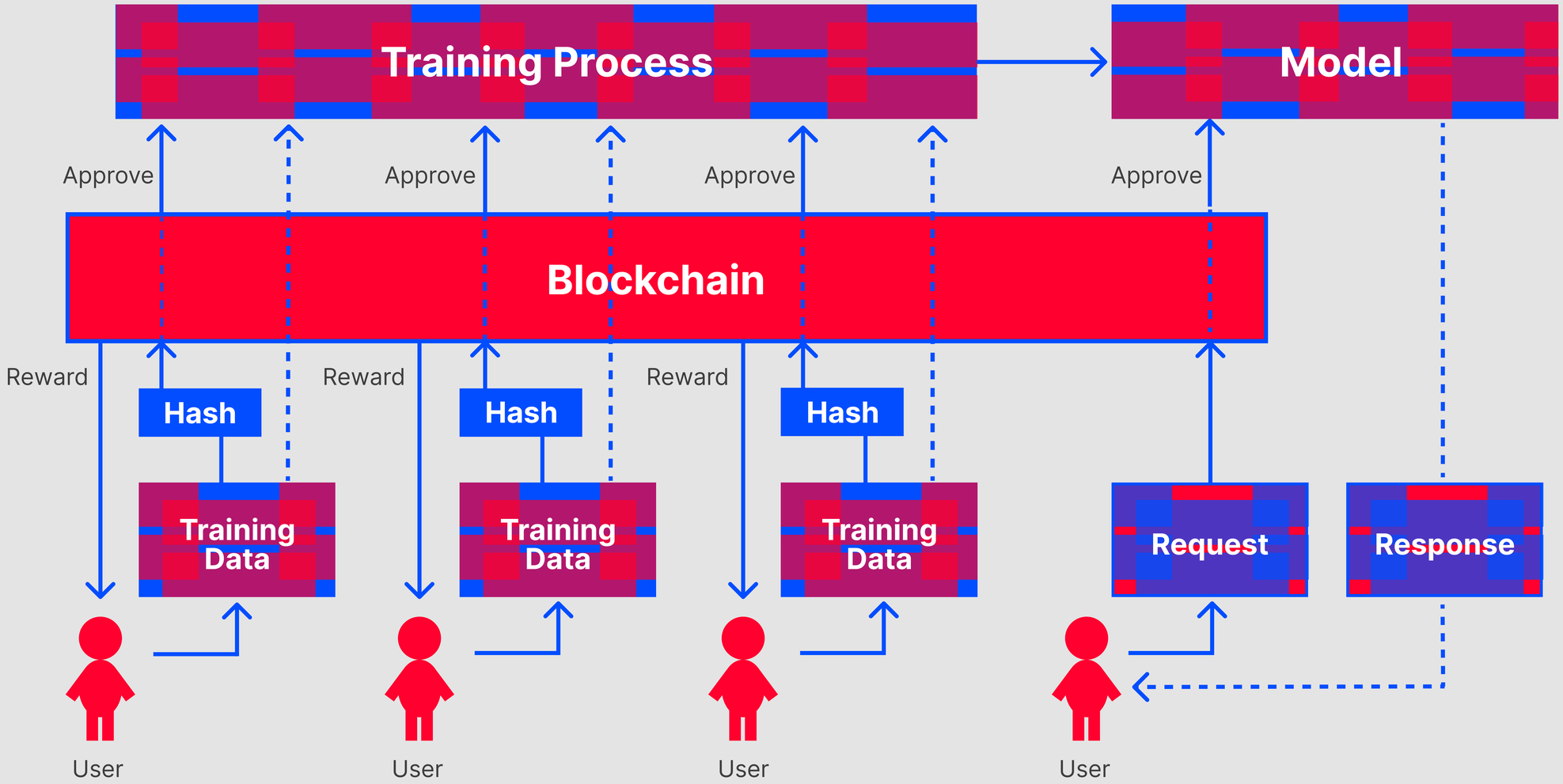
Soulbound Tokens (SBTs)

Soulbound Tokens (SBTs) are non-transferable digital tol that represent credentials, achievements, or identity.

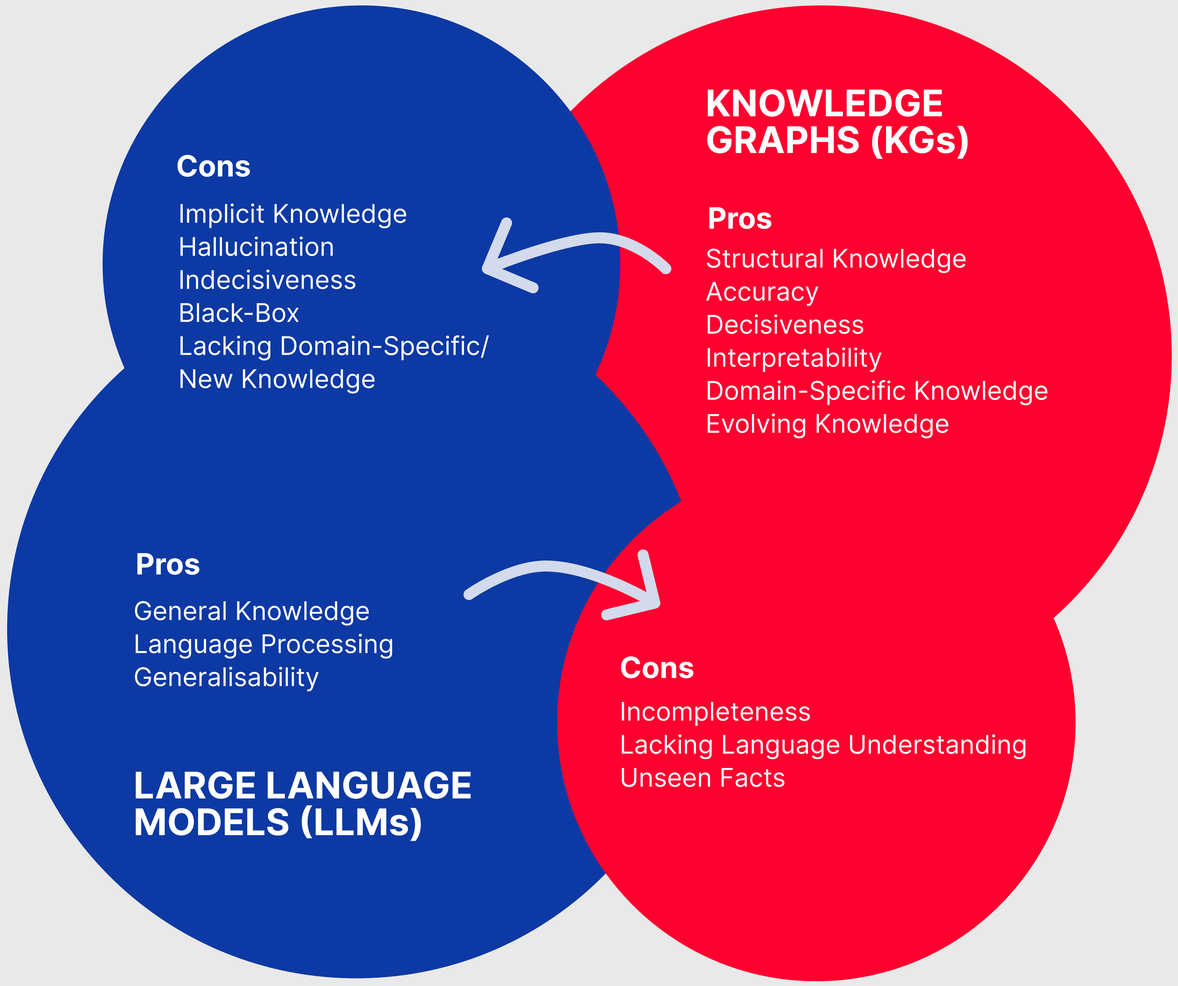
The unique difference from other tokens is that they are permanently bound to a specific blockchain wallet and cannot be transferred. The concept takes its name from the gaming world where "soulbound" items in games like World of Warcraft cannot be sold to other players once acquired. The idea was formally introduced by Ethereum co-founder Vitalik Buterin alongside economist Eric Glen Weyl and lawyer Puja Ohlhaver in their May 2022 paper titled "Decentralised Society: Finding Web3's Soul," which envisioned SBTs as building blocks for a decentralised society where social identity and reputation could be verified on-chain without relying on centralised authorities. Originally proposed for applications in financial use cases such as KYC verification (exemplified by Binance Account Bound (BAB) a token), DAO governance, academic credentials, and professional certifications, SBTs were intended to create verifiable, tamper-proof digital identities that could enable reputation-based lending, Sybil-resistant voting, and portable credentials across Web3 platforms. However, despite the theoretical promise and early implementations like BAB and Gitcoin Passport, SBTs have not yet achieved widespread adoption due to challenges including privacy concerns (as publicly visible credentials could expose sensitive information), key recovery issues (losing wallet access means losing all attached credentials), issuer centralisation risks, and the lack of standardised protocols across different blockchain networks.

Data Hashing and Model Hashing

The fundamental challenge in AI governance is ensuring that training datasets and model parameters haven't been corrupted, either accidentally through system failures or maliciously through adversarial attacks. Traditional approaches rely on centralised verification systems that create single points of failure and trust dependencies. Blockchain-based data and model hashing solves this through cryptographic proof systems where every piece of training data and each model iteration receives a unique cryptographic fingerprint (hash) that gets permanently recorded on an immutable ledger. When data flows from collection through preprocessing, training, and deployment, each transformation creates a new hash that references the previous state, building an unbreakable chain of custody that enables instant detection of any unauthorised modifications.



OriginTrail exemplifies this approach through its Decentralised Knowledge Graph (DKG) that creates tamper-proof, AI-powered databases where data provenance is cryptographically verified at every stage. The platform enables organisations to track data origin and maintain AI engines that preserve data integrity while supporting collaborative AI development. Chainlink's oracle network demonstrates complementary capabilities by providing cryptographic proofs that validate data authenticity during transmission, ensuring that data has not been tampered with between external sources and smart contracts. SIMBA Chain's applications for the defence sector show how blockchain-based data validation mechanisms using multi-party computation and zero-knowledge proofs can prevent data-poisoning attacks while maintaining auditable trails of the development process. These implementations create what researchers call "cryptographic truth"—mathematically verifiable evidence that AI models were trained on authentic, uncompromised datasets and that their decision-making processes can be audited for post-incident investigations, regulatory compliance, and continuous monitoring of model integrity.



Decentralised RAG (dRAG)

Retrieval-Augmented Generation (RAG) is an AI setup that helps companies control which data goes to AI models by creating a system where the AI first searches through approved company documents before generating responses, ensuring that sensitive information stays within the organisation rather than being absorbed into the AI model's training data where it could potentially be accessed by others. Traditional RAG systems store all company knowledge in one central database, but Decentralised RAG (dRAG) distributes this knowledge storage across multiple peer-to-peer networks, allowing different organisations to contribute their expertise while keeping complete control over their own data. This innovation uses technologies like IPFS for decentralised storage and blockchain for secure record-keeping, making it especially valuable for sensitive industries like healthcare where privacy concerns normally prevent sharing valuable data that could improve AI performance. Research demonstrates that dRAG can achieve performance nearly as good as centralised systems while providing both privacy protection and computational efficiency, essentially allowing companies to benefit from collective AI knowledge without giving up control over their proprietary information.

Intellectual Property Protection in Blockchain Tools

With growing concerns over how AI models scrape online content, mix it, and generate new material, sometimes even reproducing original works, there is a serious risk of intellectual property violations that put many companies in jeopardy. Blockchain offers several potential solutions, which are discussed here.

Curated databases with cryptographic hashing represent one of the most fundamental applications of blockchain for IP protection. Companies like Kodak pioneered this approach with their KODAKOne platform, which creates encrypted digital ledgers of rights ownership for photographers. The platform utilises blockchain technology to hash original image data, creating immutable proof of provenance that timestamps when content was first registered, allowing photographers a greater ability to assert ownership in their work and combat a trend of images being used without consent. Bernstein takes this concept further by offering certified version history that allows creators to build a digital trail of their innovation and creation processes using the Bitcoin blockchain and national timestamping authorities. Their platform can register any digital asset of any size and format quickly online to prove existence, ownership, and development over time, with all notarised data remaining perfectly private through a unique cryptographic layer. Similarly, Verisart leverages blockchain technology to provide secure and tamper-proof digital certificates, allowing artists and brands to convert their works and products into NFTs while verifying authenticity. Spawning AI aims to curate “non-infringing” media for AI model training with a focus on art-generating models such as Stable Diffusion and OpenAI’s DALL-E3.

Automatic royalty distribution systems have transformed how creators monetise their work through smart contracts that execute payments without intermediaries. The KODAKOne ecosystem demonstrates this through its KODAKCoin token standard, which enables instant payouts to photographers and stakeholders in the licensing chain, reducing traditional 90-120 day payment cycles to near-instantaneous transactions. Blockai helps writers, artists, and creators by leveraging timestamps to document the creation of new works and protect them from copyright infringement, while automatically managing licensing and royalty payments. Music industry platforms are also embracing this technology, with performing rights organisations like ASCAP and BMI exploring blockchain-based platforms to streamline licensing and royalty distribution, providing transparent, immutable records of rights ownership that reduce disputes and enhance trust among stakeholders.

Digital watermarking and authentication solutions have evolved significantly through blockchain integration, with companies developing comprehensive frameworks that combine multiple security layers. Ascribe allows artists to register and track their digital creations using blockchain, creating unique digital identities for genuine products that serve as cryptographic identifiers recorded on blockchain networks. These digital identities can be embedded into products using RFID tags, QR codes, or NFC chips, making them extremely difficult to replicate or forge. Nike's CryptoKicks patent (U.S. Patent No. 10,505,726) exemplifies how major brands are leveraging blockchain for anti-counterfeiting, creating NFTs linked to physical products that serve as digital certificates of authenticity. Chronicled specialises in supply chain authentication, enabling transparent tracking from manufacture to consumer through blockchain-based product authentication systems that create comprehensive audit trails for verifying authenticity and provenance.

Advanced blockchain IP solutions extend beyond basic protection to encompass comprehensive ecosystem management. IBM and IPwe have partnered to create blockchain-based platforms that allow companies to tokenise and trade patents as NFTs, revolutionising how intellectual property rights are managed and monetised. Provenance focuses on supply chain transparency, using blockchain to track products from origin to consumer, providing immutable records that combat counterfeiting and verify authenticity throughout the distribution chain. Mycelia is developing blockchain-based solutions specifically for music rights management,

creating transparent systems for tracking music usage, managing licensing agreements, and ensuring fair compensation for artists and rights holders. Everledger specialises in high-value item authentication, using blockchain applications to create digital passports for luxury goods, diamonds, and other valuable assets, providing tamper-proof provenance records that combat fraud and counterfeiting.

Emerging solutions continue to push the boundaries of blockchain IP protection, such as Story Protocol and Camp Network. Whilst these networks have seen limited usage to date, the World Intellectual Property Organization (WIPO) has recognised that "blockchain technologies have both disruptive and transformative potential" and specifically mentioned platforms like Bernstein as established providers of tamper-proof documentation with precise timestamps. As these technologies mature, they promise to create a comprehensive ecosystem where intellectual property rights are not only protected but also efficiently managed, traded, and monetised through transparent, automated, and globally accessible blockchain-based systems.

Text Watermarking Beyond the Pixel

Text watermarking faces unique challenges that set it apart from visual media, where watermarks can be embedded in discrete pixels without fundamentally altering meaning.

Unlike images where you can modify pixel values or insert visual markers, text watermarking must navigate the discrete, semantic nature of language where every word carries meaning and even small changes can alter context. The core challenge lies in the fundamental difference between modalities: while image watermarking can leverage imperceptible pixel modifications or frequency domain transformations, text watermarking must work within the constraints of vocabulary, grammar, and semantic coherence.

Current approaches include probabilistic token-level watermarking that biases language models toward "green" tokens during generation, creating statistical patterns detectable by authorised verifiers. However, these methods face significant vulnerabilities—research shows that simple paraphrasing can remove watermarks in about 30% of cases, while more sophisticated attacks can reverse-engineer statistical watermarks with approximately 85% success rates. For IP protection and content tracking, tools focus on semantic similarity analysis and AI detection systems that examine linguistic patterns, perplexity scores, and statistical anomalies in token distributions. Detection tools like GPTZero and various machine learning classifiers analyse features such as sentence structure consistency, vocabulary patterns, and stylistic markers that distinguish AI-generated from human-authoured content. Yet these systems suffer from high false positive rates (up to 15-20% for text) and struggle with heavily edited content, multilingual scenarios, and the evolving sophistication of language models.

Legal Approaches to AI Regulation: Balancing Innovation, Timing, and Cultural Frameworks

The Patchwork Problem

The regulatory landscape for AI and blockchain technology resembles a patchwork quilt sewn by committees that never talked to each other. The EU's GDPR grants individuals the right not to be subject to automated decision-making, while the EU AI Act requires high-risk AI systems to maintain detailed audit trails. Meanwhile, the UK's AI White Paper takes a principles-based approach that delegates authority to existing sector regulators.

From GDPR and the UK Data Protection Act to the EU AI Act, UK AI White Paper, NIST AI RMF, and the US Executive Order on AI, regulators consistently emphasise transparency, accountability, and traceability. Market and conduct regulators, such as the FCA, FSMA, and VARA, are extending oversight into crypto and digital assets, while sector-specific rules such as the Online Safety Act and medical device frameworks impose duties around risk management, content safety, and surveillance. International standards reinforce these demands (see the full list of regulations in the Appendix).

The fundamental tension is between the global nature of AI and blockchain systems and the territorial nature of legal jurisdiction. An AI model trained in one country may be deployed globally, while blockchain networks operate across multiple legal systems simultaneously. This creates what we call "regulatory arbitrage opportunities" where organisations can shop for the most favourable legal frameworks. But that is not the only dilemma to be resolved.

Safety vs. Innovation

The tension between AI safety and innovation represents one of the most contentious debates in technology policy, with critics arguing that excessive regulation could stifle breakthrough developments and drive AI research to more permissive jurisdictions. Industry leaders warn that overly prescriptive regulations could create compliance-heavy environments that favour large corporations with extensive legal teams while disadvantaged startups and smaller innovators who lack resources for complex regulatory navigation. However, the counter-argument emphasises that unregulated AI development poses existential risks that could undermine the very innovation ecosystem these policies aim to protect, requiring industry-specific frameworks that provide transparency on how every company develops and implements their AI policies.

A particularly interesting point here is the divergence between the United Kingdom and the European Union, in that the EU AI Act seeks to bring in specific laws in relation to AI whereas the UK has taken the approach that no additional specific regulation is needed – rather the existing laws and regulation within the UK already suffices. It is notable in this respect that “regulation” simply refers to the existence of a set of rules and requirements (and sometimes the existence of a particular entity – the regulator – to police and enforce such rules and requirements), and simply stating a requirement may not be the best solution. For example, in the event that AI may predict something with a 90% accuracy, it may be worth considering some kind of insurance solution to deal with the remaining 10%, rather than relying on trying to regulate it out of existence.

Currently, companies are increasingly expected to develop internal AI governance policies, with 87% of executives reporting established AI ethics committees, yet there is no standardised framework for what these policies should contain or how they should be implemented. The UK government's AI White Paper acknowledges this gap, stating that while "organisations should have appropriate governance arrangements in place," it provides limited guidance on what constitutes "appropriate" governance, leaving companies to develop ad-hoc approaches. This regulatory vacuum creates significant problems: while 73% of companies have AI governance policies, only 31% have implemented comprehensive risk assessment frameworks, suggesting that policy development outpaces practical implementation.

Retrospective vs. Proactive

Is there even a need to develop an AI regulation? Surprisingly, we saw a variety of opinions during the research. Most of our US-based interviewees expressed confidence in retrospective regulation, arguing that lawyers and courts will develop appropriate frameworks on a case-by-case basis as AI-related disputes emerge through litigation. One respondent drew a compelling parallel to the invention of automobiles: before cars existed, society didn't anticipate the need for traffic lights, speed limits, or mandatory insurance—these regulations emerged organically in response to accidents and societal needs. This evolutionary approach to technology regulation has historical precedent, with industries like telecommunications, aviation, and the internet developing regulatory frameworks through iterative responses to emerging challenges rather than comprehensive pre-emptive legislation. However, the alternative perspective argues that AI's unprecedented speed and scale demand proactive intervention: the ambiguity and absence of clear AI regulation leaves companies and individuals navigating uncharted territory without shared standards or frameworks. Critics of the retrospective approach warn that AI's potential for rapid autonomous decision-making and system-wide impacts put AI into a different category compared to previous technological breakthroughs.

West vs. East

One interviewer highlighted the fundamental philosophical divide between Western and Eastern regulatory approaches, noting that the US traditionally favours minimal regulation to maximise innovation and market freedom, while Eastern countries often implement comprehensive regulatory frameworks from the outset. This cultural difference reflects deeper values about the role of government, individual versus collective responsibility, and risk tolerance. Both approaches have demonstrated historical success with distinct trade-offs: the Western model has fostered rapid technological innovation and entrepreneurship, producing global tech leaders like Google, Microsoft, and OpenAI, while Eastern regulatory frameworks have often provided greater social stability and coordinated development strategies.

Confidentiality vs. Transparency

The tension between regulatory transparency demands and operational confidentiality creates what one can call the "governance paradox" of blockchain-based AI systems. Regulators simultaneously demand complete auditability while requiring that sensitive information remain confidential—a contradiction that reveals the fundamental misunderstanding of how blockchain systems actually work.
Consider the UK's Online Safety Act, which requires platforms to maintain comprehensive moderation logs while protecting user privacy. The Act mandates that platforms demonstrate accountability in their content moderation decisions, but immutable blockchain logs could expose moderation strategies that bad actors could exploit or reveal private user content that should remain confidential. This creates a scenario where compliance with one regulatory requirement directly conflicts with another.

Canada's Directive on Automated Decision-Making presents an even starker example of this paradox. The directive requires algorithmic impact assessments (AIAs) that may include procurement details, national security information, or competitive intelligence that cannot be made public. Yet the directive also calls for transparency and accountability that would seemingly require public disclosure of these same assessments. The question becomes: how do you create transparent governance for systems that must remain partially opaque for legitimate security or commercial reasons?

The US Executive Order on AI (2023) calls for red-teaming exercises and public disclosure of test artifacts, but federal agencies differ significantly on what can safely be disclosed without compromising national security or enabling adversarial attacks. This regulatory uncertainty creates compliance challenges for organisations that must balance transparency requirements with security considerations.

Privacy vs. Immutability

The collision between privacy rights and blockchain's permanence represents perhaps the most fundamental challenge facing blockchain-based AI governance. Under GDPR, UK GDPR, and the Data Protection Act 2018, individuals possess rights of access, rectification, and erasure that are fundamentally incompatible with blockchain's immutable record-keeping.

The "right to be forgotten" under Article 17 of GDPR creates what cryptographers call an "impossible requirement" for blockchain systems. How do you delete information from a distributed ledger where the entire security model depends on the impossibility of altering historical records? The Information Commissioner's Office has issued guidance acknowledging this tension, but offers no clear technical solution for reconciling erasure rights with permanent record systems.

Healthcare applications face particularly acute challenges around this immutability trap. Software as a Medical Device (SaMD) regulations require post-market surveillance capabilities and the ability to update or correct medical records dynamically. When diagnostic AI systems record their decisions on immutable blockchain ledgers, how do you handle cases where those decisions are later determined to be incorrect? The permanence that makes blockchain valuable for audit trails becomes a liability when medical accuracy requires the ability to correct historical records.

The Online Safety Act 2023 exemplifies this regulatory confusion by requiring platforms to demonstrate moderation accountability without specifying how much transparency is reasonable when immutable logs risk exposing private user data. The Act's requirements seem to assume that perfect transparency and perfect privacy can coexist—a mathematical impossibility that reveals the gap between regulatory intentions and technical reality.

This immutability trap forces us to confront a fundamental question about the nature of digital governance: is the ability to correct mistakes more important than the ability to prevent tampering? The answer may determine whether blockchain-based AI governance systems can achieve regulatory compliance or remain forever relegated to experimental applications.

Legal Liability for AI Decisions

The question of legal liability when AI models make wrong decisions has become one of the most contentious issues in technology law, with multiple stakeholders potentially bearing responsibility depending on the circumstances.

The liability framework typically involves three key parties: the AI model provider (e.g., OpenAI), the organisation that deploys it, and the individual who initiates the action. Current legal approaches vary significantly across jurisdictions, with the EU's AI Liability Directive proposing a two-stage framework: strict liability for operators of "high-risk" AI systems and fault-based assessment for other AI applications.

The pros and cons of different liability approaches reveal fundamental tensions in AI governance. Strict liability on deploying organisations offers quick compensation for victims and creates strong incentives for careful AI implementation, as demonstrated in cases like Moffatt v. Air Canada, where the court held that Air Canada was responsible for all information provided by its AI chatbot, treating the bot as part of the company's website rather than a separate entity. However, this approach may discourage AI adoption and place a disproportionate burden on smaller companies. Developer liability through breach of contract or product liability laws provides a pathway for organisations to seek redress when AI systems fail to meet specifications,

but the autonomous nature of AI makes it challenging to prove that unexpected failures stemmed directly from the developer's actions. Individual user liability becomes relevant when users misuse AI systems or ignore safety warnings, though recent lawsuits like the case against OpenAI over a teen's suicide highlight the complexity of determining when users versus providers bear responsibility for harmful outcomes.

The ongoing wave of lawsuits against major AI companies—including copyright infringement cases against OpenAI, Microsoft, and Anthropic, as well as emerging wrongful death claims—will likely establish crucial precedents for how courts allocate responsibility in the AI age. The fundamental challenge remains balancing innovation with accountability: while strict liability regimes may provide clearer compensation pathways for victims, they could also drive AI development offshore to jurisdictions with more permissive frameworks, potentially undermining both innovation and safety in the long term.

Who Was Liable	Name Of Lawsuit	Common Pattern Of Justification
AI Model Provider (e.g., OpenAI)	Parents vs. OpenAI (Source)	Allegations of prioritising profit over safety, leading to harm (e.g., suicide coaching via AI).
AI Model Provider (e.g., Anthropic)	Authors vs. Anthropic (Source)	Unauthorised use of copyrighted works to train AI models, claiming infringement.
AI Model Provider (e.g., OpenAI)	Ziff Davis vs. OpenAI (Source)	Claims of copyright infringement through unauthorised use of media content for training.
AI Model Provider (e.g., Anthropic)	Reddit vs. Anthropic (Source)	Scraping user comments from platforms without permission to train AI, violating terms of service.
Organisation Deploying the System	Meta vs. Porn Sites (Source)	Allegations of scraping content without consent to train AI models.
Organisation Deploying the System	Character.AI and Google Lawsuit (Source)	AI systems creating emotional dependency, leading to psychological harm or suicide.
Developer Creating the System	Disney and NBCUniversal vs. Midjourney (Source)	Claims of using trademarked characters to train image-generation models, producing infringing outputs.
Individual Initiating the Action	Musk vs. OpenAI (Source)	Allegations of unfair practices and harm to competition in the AI sector.

The alignment between blockchain's immutable audit trails and emerging AI regulatory requirements from the EU AI Act to the UK's AI White Paper is mathematically elegant, yet crashes against five structural realities that code cannot resolve. The privacy-versus-immutability conflict remains a fundamental impossibility—you cannot simultaneously guarantee permanent records and GDPR erasure rights. Standardised schemas don't exist, undermining auditor confidence in blockchain evidence. Cross-border recognition is absent, fragmenting global trust infrastructure. Overlap with financial regulation through MiCA creates unexpected legal risks when governance tokens become securities. Most critically, we lack principled frameworks for determining what should be transparent versus private, forcing ad-hoc decisions that satisfy neither regulators nor users. The question isn't whether blockchain will become essential for AI governance, but whether we can resolve these structural conflicts before the window for establishing effective governance closes entirely.

Conclusion: The Choice We Face

We stand at a critical juncture in artificial intelligence development where the decisions made today will determine whether these technologies enhance human capabilities or create new systemic risks. The traditional approach of human-speed oversight for machine-speed processes is fundamentally inadequate for current AI systems, let alone the more powerful autonomous systems emerging. We need governance mechanisms that operate at the same speed and scale as the systems they govern while maintaining necessary transparency and accountability.

Blockchain technology offers a viable path forward, providing infrastructure for data and model provenance, distinguishing humans from AI-generated content, and enabling automated governance with built-in transparency, traceability, and immutability. The alternative to blockchain-based AI governance is not improved human oversight — it's diminished oversight altogether. As AI systems become more complex and autonomous, traditional governance mechanisms will struggle to keep pace, potentially leaving powerful AI systems operating with minimal accountability or transparency.

Here is the list of our overall recommendations:

For Policymakers:

Invest in pilot programs demonstrating blockchain-based AI safety tools, create international standards for AI identity and data provenance, and establish regulatory frameworks that balance innovation with necessary oversight. Focus on evidence-based policy making that addresses technical realities rather than theoretical concerns. Adapt high-risk sectors regulation to address emerging risks.

For Enterprises:

Begin with internal blockchain implementations for AI audit trails and employee verification, gradually expanding to external applications as expertise develops. Prioritise hybrid solutions that combine centralised efficiency with decentralised verification, and invest in training programs to build organisational blockchain capabilities.

For Technology Builders:

Focus on user experience as the primary adoption barrier, develop open-source frameworks that enable collaboration while maintaining security, and prioritise interoperability across different blockchain networks and AI platforms. Address technical limitations through continued research while building practical solutions for immediate deployment.

The window for establishing effective AI governance is narrowing, but the opportunity remains to create frameworks that preserve human agency while enabling autonomous technology benefits. The future of AI governance will be determined by the systems we build and the choices we make. Blockchain technology provides tools for more transparent, accountable, and democratic AI systems. Whether we use them wisely depends on our collective commitment to building technology that serves human flourishing.

Appendix Overview of Current Regulations

GDPR / UK GDPR (Data protection principles & data subject rights)	Sets legal duties on controllers/processors (lawful basis, data minimisation, rights to access/erasure/rectification) — provenance & consent logs are highly relevant for compliance. (Information Commissioner's Office)	Practical guidance on immutable ledgers (how to reconcile right-to-erasure with on-chain immutability; when pseudonymisation is "enough").	How to remove/rectify on-chain data; proving lawful basis when training models on mixed on-/off-chain datasets; monitoring users to stop them sending sensitive data into external AI services. (Information Commissioner's Office)
Data Protection Act 2018 (UK)	UK-specific implementation of GDPR rules and local exceptions — governs UK enforcement and transfer regimes. (GOV.UK)	Specific guidance on UK interpretations of immutable-record workarounds (e.g., legal bases, exemptions) for blockchain is sparse.	Aligning UK-specific transfer rules and local ICO expectations with on-chain attestations anchored outside UK jurisdiction. (GOV.UK, Information Commissioner's Office)
EU AI Act	Risk-based obligations (high-risk systems need conformity, record keeping, transparency) — blockchains can record tamper-evident dataset provenance and model change logs for conformity audits. (OECD)	Operational rules for decentralised or cross-border attestations, and how to treat general-purpose models and off-chain compute anchored on-chain.	Assigning provider/user liability in decentralised systems; making on-chain proofs acceptable for conformity assessments; cross-border enforceability. (OECD)
UK AI White Paper	Outcomes-focused, pro-innovation approach with sandboxes and emphasis on standards/interoperability — good fit for piloting blockchain mitigation tools in UK regulatory sandboxes. (Information Commissioner's Office, ISO)	Limited prescriptive technical standards for verifiable on-chain attestations or how immutable trails satisfy UK accountability requirements.	Regulatory fragmentation vs EU (different expectations); sandbox acceptance ≠ regulatory certainty across other regimes. (Information Commissioner's Office)
NIST AI RMF	Voluntary risk-management lifecycle mapping (identify, govern, assess, manage) — blockchain supports tamper-evident evidence required by governance and assessment phases. (ISO)	No standard schemas for integrating distributed ledger proofs into RMF artefacts (e.g., standard proof formats, metadata).	Translating RMF into on-chain schemas; storage vs cost trade-offs (anchor vs full data); interoperability between ledgers. (ISO)

FCA Discussion Paper (crypto / crypto-asset regulation)	FCA focuses on market integrity, consumer protection and disclosures — relevant when tokenised incentives, attestations or tradable model outputs are used. (jtc1info.org)	Explicit treatment linking crypto-market rules to AI governance (e.g., model provenance for tradable signals) is thin.	Mapping on-chain attestations to FCA market abuse / disclosure frameworks; preventing leakage of IP/business-sensitive info on-chain. (jtc1info.org)
VARA (Dubai Virtual Assets Regulatory Authority)	Practical licensing regime for virtual asset service providers — supports on-chain attestations, custody and tokenisation for AI-governance tooling. (Information Commissioner's Office)	VARA is asset/regulator-focused; it lacks AI-specific provenance/audit rules (i.e., how to certify model training provenance).	Jurisdictional recognition — VARA-licenced attestations may not meet EU/UK conformity evidence standards. (Information Commissioner's Office)
Financial Services and Markets Act (FSMA) changes / crypto proposals (UK)	HM Treasury / FSMA updates are bringing cryptoassets into the UK regulatory perimeter—important where blockchain attestations become regulated activities or financial promotions. (GOV.UK, Legislation.gov.uk)	Clear boundaries when tokenised attestations or governance tokens cross into regulated financial products.	Need to design token mechanics and disclosures to avoid regulated-activity triggers; coordinate with FCA/HM Treasury expectations. (GOV.UK)
Online Safety Act 2023 (UK)	Mandates platform risk assessments, safety-by-design for recommendation algorithms and content moderation — blockchain can supply immutable moderation logs and provenance for model outputs. (FTI Consulting, IT Pro)	How to use on-chain logs while preserving user privacy and meeting Ofcom codes (what level of transparency is “reasonable” vs revealing private info).	Balancing immutable public logs with privacy and encryption; risk that on-chain logs expose sensitive moderation signals or methods. (FTI Consulting)
OECD AI Principles	High-level international principles (transparency, accountability, traceability) — blockchain maps neatly to OECD traceability/accountability recommendations. (OECD, OECD AI Policy Observatory)	Not legally binding; lacks operational standards for on-chain proofs or schemas to demonstrate traceability.	Converting principles into standard on-chain metadata and proving those meet auditors’ expectations. (OECD)

ISO/IEC JTC 1/SC 42 (AI standards)	Active international standards work (AI management systems, risk, transparency) — ISO standards can provide the schema/format that makes on-chain evidence interoperable and auditable. (ISO, jtc1info.org)	Many ISO AI standards are still in development; specific guidance for integrating distributed ledgers into ISO artefacts is not mature.	Timing & adoption risk: standards evolve slowly while tech moves quickly — may be gaps between ISO releases and on-chain design decisions. (ISO)
Singapore — AI Verify (IMDA / AI Verify Foundation)	Practical testing & reporting framework (toolkit + tests) for verifying AI against ethics principles — blockchain can store signed test reports and tamper-evident results. (Infocomm Media Development Authority, AI Verify Foundation)	Guidance on decentralised publication of test evidence (what to put on-chain vs off-chain) and how to protect proprietary test data.	Confidentiality of test data vs public auditability; standardising the report hashes/anchors so verifiers globally can validate. (AI Verify Foundation, World Economic Forum)
Canada — Directive on Automated Decision-Making (Treasury Board)	Requires algorithmic impact assessments, transparency and recourse for government systems — good precedent for signed AIA artefacts anchored on-chain for public accountability. (Canada, Statistics Canada)	Guidance on how immutable public AIA records interact with privacy and update/rectification requirements.	Ensuring AIA anchors don't leak sensitive procurement or national-security info; managing updates to assessments as models evolve. (Canada)
US Executive Order (Safe, Secure & Trustworthy AI — Oct 2023)	US policy push for safety, testing, red-teaming and adoption of standards (supports AI RMF alignment) — on-chain storage of red-team/test evidence supports these aims. (The White House, Federal Register)	No single binding federal standard — implementation guidance for blockchain proofs and evidence formats is fragmented across agencies.	Fragmented agency expectations; evidence accepted by one agency may not be suitable for another; nation-level security concerns about publishing test artefacts on public chains. (The White House)
Software as a Medical Device (SaMD) — UK	UK's MHRA regulates software (including AI) as medical devices under UK MDR 2002. SaMD, especially AIaMD, requires classification, post-market surveillance (PMS), cybersecurity risk management, and incident reporting. Blockchain can provide tamper-evident logging of software updates, traceability of model versions, and secure incident logs. (GOV.UK, Penningtons Law, Cyber Alchemy)	Specific guidance on integrating blockchain-based logs with SaMD PMS systems or how immutable on-chain records align with regulatory expectations is lacking. There's little clarity on how on-chain evidence supports cybersecurity obligations or incident timelines.	Ensuring on-chain audit logs meet MHRA standards for PMS and cybersecurity; protecting sensitive medical data while preserving transparency; aligning immutable on-chain records with dynamic clinical updates and corrections.