Mitigating AI Privacy Risks in Latin America

Identity Infrastructure, Institutional Capacity and the Path to Adoption

Claudio Cifuentes Lobo
Lead Author
Latin American Dynamism Project
claudio@ladp.io

Marcus Alburez
Co-Author
Latin American Dynamism Project
marcus@ladp.io

Said Saillant, Ph.D.
Research Advisor

Abstract

Artificial intelligence is producing measurable harm in Latin America. This report asks whether decentralized identifiers, verifiable credentials and zero-knowledge proofs, part of Ethereum's cryptographic stack, can work as practical infrastructure against AI-driven privacy risks in the region. It combines a scoping review of AI-risk taxonomies and of the literature on the AI-cryptography convergence with 22 semi-structured interviews conducted between November 2025 and March 2026. It examines five case studies: QuarkID/miBA, Sovra, Blerify, World and Proyecto DIDI, three of which settle to Ethereum. Each case is assessed on the same dimensions: credential format, proofs in production, orientation to an AI risk, reach, and institutional counterpart. Three findings follow. First, the region's government-backed identity rails are more mature and more widely deployed than public awareness suggests: three of the five cases are in production, with one recording 1.8 million DIDs created by July 2026. Second, the path to deploying decentralized identity against AI privacy risks runs through international standards and interoperable protocols the region already partly operates (W3C verifiable credentials, ISO 18013-5 mdoc, OpenID4VCI/VP), and through the emerging protocols that extend them to AI agents (the W3C agent-credential groups, AP2 and ERC-8004). Third, what these deployments still lack is rarely technological. The blockers documented are the legal instrument that trails delivery, the absence of an AI-threat argument that officials can defend internally, cryptographic guarantees that have not been translated into policy language, and transparency practices that build public trust. The report also maps proof of non-personhood as the step that would let the region deploy AI agents in public services with a verifiable chain of authorization, and finds Latin American actors already in the venues defining it, but not yet the public operators of the credential systems studied here. The report closes with 13 recommendations across four areas: entering government, standards and interoperability, education and institutional capacity, and pilot scoping with institutional partners.

Keywords: decentralized identity, verifiable credentials, zero-knowledge proofs, AI governance, proof of non-personhood, digital public infrastructure, Latin America.

Contents

Executive Summary 3
- Recommendations 3
- Case Studies at a Glance 6

Methodology 6

Literature Review 8
- AI Risks: A Taxonomy 8
- AI Risks in LATAM 15
- AI Risk Mitigation x Cryptography - Convergences 19

Case Studies: Decentralized Identity for AI-Privacy in LATAM 25
- Proyecto DIDI 26
- QuarkID Protocol 28
  - miBA: QuarkID's first pilot 30
- Sovra 32
  - Propia ID: A Nascent Public Trust Layer for Verifiable Credentials 33
- Blerify 36
- World 42
- Evidence Summary 51
- Cross-Cutting Findings 51
- Open Research Questions 52

Proof of Non-Personhood: An Emerging Frontier 53

Closing 58

Appendix: Interviewees 59

Notes 60

Executive Summary

Artificial intelligence is producing measurable harm in Latin America. This report examines one of those harms, privacy, not because it ranks first in the regional evidence, where labour displacement and technological sovereignty score higher, but because it is the only one with concrete cryptographic deployments to study: synthetic media that defeats the conventional identity protocols and biometric checks that banks and governments rely on, phishing at industrial scale, and personal data flowing into systems with no accountability to the populations they classify. It asks whether decentralized identifiers, verifiable credentials and zero-knowledge proofs, part of Ethereum's cryptographic stack, can work as practical infrastructure against those risks.

Drawing on a literature review, a mapping of global and Latin-American AI-risk taxonomies, and interviews across the ecosystem, it examines public and private identity deployments in the region and what each could carry against AI privacy risks. Of the five case studies, three depend on Ethereum L1: QuarkID anchors on zkSync Era, Sovra runs its own Validium rollup, and World's canonical identity tree is a mainnet contract whose roots are bridged to L2s. Two run on networks that do not settle to Ethereum: Blerify on Avalanche and Hyperledger Besu, Proyecto DIDI on LACChain.

Three overall findings follow. First, the region's government-backed identity rails are more mature and more widely deployed than public awareness suggests. QuarkID underpins miBA, the Buenos Aires city app through which 3.6 million residents access their documents, had recorded 1.8 million DIDs created on-chain by July 2026 according to a public Matter Labs query, with creation still running at tens of thousands a month in 2026, and is the most widely deployed DID protocol in LATAM government infrastructure that this research identified; Sovra runs government programmes in Nuevo León and Salta on the same protocol; Blerify piloted a decentralised trust list with AGESIC in Uruguay and reports having built the digital ID systems now awaiting launch in El Salvador and Panama.

Second, the AI-risk use case is stated openly by private companies and almost nowhere by the public institutions that would deploy it. World positions its biometric enrolment as proof of human for an internet crowded with AI. Blerify sells verifiable credentials as a way to withstand deepfakes and retire KYC mechanisms that no longer hold. Sovra now advertises an identity layer for government AI agents. All three rest on cryptographic techniques, but few government or multilateral programs in the region have taken up that intersection, which leaves the value of infrastructure they already operate uncollected.

Third, what these deployments still lack is rarely technological. Smart contracts, advanced cryptography and biometric enrolment are in place and working. It is the legal instruments, the public policy, the narratives and the advocacy which are in their infancy.

The recommendations below distill the findings into 13 actionable points across four areas, followed by a summary table of all case studies analyzed.

Recommendations

I. Closing the integration gap and entering government

1. In the cases studied, a government-first strategy is what unblocked the application of cryptography and AI products. Cryptographically-backed credential ecosystems stall on a coordination problem: issuers and holders will not invest before verifiers exist at scale, and verifiers will not invest before holders do. A government issuer resolves both sides at once, producing a large base of legitimate credential holders and, through its own administrative procedures, verification demand that no private issuer can manufacture. The regulatory work follows from that rather than preceding it: having a government-scale implementation already running is what made it possible to push for rules permitting decentralized alternatives to mandatory centralized records. Two vendors and one city are the pattern this research found; whether it holds as a general rule is a question for a larger sample.

2. Identity programs built on cryptographic rails already exist in LATAM: consider building AI layers on top of them. No public deployment in Latin America uses its identity infrastructure as an instrument against the privacy risks of AI. The one actor in the region with a track record on that problem is private: World has built multiple institutional partnerships to verify that humans rather than AI agents are on the other side of an interaction. Governments run verifiable credential programs built on similar cryptographic techniques, but issue them for documents, benefits and identity, never against AI risks. The first move is to put it on the agenda of the programs already running, on Ethereum or not.

3. Lead with the AI-threat argument, and give officials something they can defend internally. Government decision-makers are not procuring cryptographic infrastructure; they are accountable for citizen protection and are under political pressure from AI-driven fraud. What carries the conversation toward cryptography as a usable instrument against that fraud is precedent, international and regional, of the two being deployed together. Precedent moves the decision from evaluating an unfamiliar vendor to joining something governments comparable to them have already committed to. What the ecosystem lacks is a shared version of this argument, with the incident evidence and both levels of precedent already assembled, so that the official across the table has something to take into a budget meeting instead of each actor rebuilding it from scratch.

4. Mid-size countries may offer fewer veto points and shorter timelines. Countries in the 3 to 12 million population range (Panama, El Salvador, Dominican Republic, Uruguay, Costa Rica) concentrate the decision in fewer hands than the large LATAM economies, where a credential layer has to clear a federal government, its agencies and the subnational units that will issue. That band widens the 4 to 10 million pattern Blerify reported from its own government engagements, so as to take in three countries whose digital identity procurement this research did not confirm. Sovra's confirmed deployments show the shape of that: both are a state and a province rather than national programs. For pilots under time and resource constraints it is a reasonable working hypothesis, drawn from two vendors' engagements rather than from a comparison this research ran.

5. Budget the legal framework into the initial deployment roadmap, with as much weight as the technology, if not more.

Throughout our case studies, teams pointed out that it's common for technical iteration to outpace regulatory oversight, where an authority reviewing a system can take a year or more to reach a conclusion about a version that no longer exists. Regulators do not have the time, incentives or capacity to close that gap themselves, so a project that treats legal work as something that follows delivery is choosing to be potentially judged on documentation it has already superseded. World is the case in point, holding the largest deployment in the region while losing jurisdictions to legal decisions rather than technical ones. Treat the enabling instrument, decree, resolution or reform, as a deliverable with an owner and a date inside the project plan.

II. Standards and interoperability

6. Build credentials on international standards. A credential that will be verified outside the organisation that issued it has to rest on an international standard. Government work needs both: the ISO 18013-5 mdoc format for identity documents, whose fixed namespace for driving licences is what makes a credential readable by a verifier that has never heard of the issuer, and which identity programmes are extending to other documents, and W3C verifiable credentials for everything else a government issues, where no such namespace exists. A private issuer putting out a diploma, a professional licence or a KYC result can work in W3C alone. Both travel over the same OpenID protocols, which is how Blerify and Sovra already operate.

7. Interoperable regional protocols can integrate LATAM's ecosystem and reach government scale. QuarkID, built by a LATAM consortium, became the most widely deployed DID/VC protocol in the region's public infrastructure that this research identified, and what made that possible was the coordination it unlocked: a common language several organisations could implement without committing to one vendor. The arrangement has since come apart, and the momentum with it. Participants in this research described it as important to keep dis-

cussing, and ideally building, protocols of this kind: designed around the region's own characteristics, fused with the global standards the region needs to interoperate with, and with an integrated ecosystem behind them.

8. Consider post-quantum from the start. Credentials issued today are signed with elliptic-curve schemes that a cryptographically relevant quantum computer would break, and the long-term risk is forgery rather than decryption: an adversary who derives an issuer's private key from its public key can mint credentials that verify as genuine. The exposure is uneven: issuer keys, trust anchors and root certificates are expected to operate for decades, which puts them well inside that timeline. Of the case studies here, only Blerify designed for it. Post-quantum certificates are an integrated feature of its issuance platform, shipped in hybrid mode so that post-quantum and classical cryptography run side by side and existing integrations keep working, and the company's founding team published the first quantum-resistant EVM blockchain implementation. Consider integrating quantum readiness into the architecture now and deploy fully once ISO 18013-5 and OpenID4VP add the NIST algorithms to their registries.

III. Education and institutional capacity

9. Translate cryptographic guarantees into policy language, and fund that translation. Open-source code and whitepapers do not reach the audience that determines whether this infrastructure gets adopted at scale. Regulators and legislators who cannot read ZK proof systems cannot evaluate their privacy guarantees, enforce their use, or legislate sensibly about them. The ecosystem needs sustained investment in face-to-face education with the people who set the rules: what the technology does, what it doesn't do, what its failure modes are, and what governance it requires.

10. Design for the user you have, not the user the protocol assumes. Proyecto DIDI ran into digital literacy rather than cryptography. The joint report it published with the IDB and LACChain is specific about where the gap sat: end users, validators and issuers alike were found to have basic or non-existent digital skills, and a single training approach across those three groups was insufficient. It also names a design consequence projects rarely accept, that some core tenets of self-sovereign identity proved counterproductive to adoption given the socioeconomic realities of the target population, which makes the trade-off between technological purity and usability a decision to take rather than defer.

11. Make the AI-risk case to the people holding the credential, not only to the officials issuing it. The argument that opens government conversations in this report is the AI threat: deepfakes defeating biometric verification, phishing that reaches institutional staff, synthetic identities at scale. It works because the officials on the other side of the table already feel the exposure. The same argument is not being made to the people who will carry the credentials. A citizen who does not know that a remote identity check can be defeated has no reason to prefer a cryptographic public service over a convenient one, and a credential adopted purely for convenience never has its risk-mitigation property exercised or valued.

IV. Pilots with institutional partners

12. Pilot agent credentialing before governments cross into transactional AI. Government AI assistants across the region are moving from answering questions to completing procedures. That transition changes the accountability requirements: a decision starts carrying real-world consequences, and the absence of a verifiable audit trail stops being theoretical. The pilot: work with a government digital agency, ideally with a multilateral as convener so the result travels beyond one country, to design and prototype a credentialing system for the agents operating inside a public service, paired with a cryptographic audit trail for transactional decisions. The prior design question is whether the DID/VC infrastructure several countries already run can express agent delegation, which is cheaper to answer than to build from scratch.

13. Pilot a proof-of-non-personhood layer before bot-resistant services are needed. Every conversation with identity infrastructure builders (World, Blerify, Sovra) surfaced the same anticipatory concern: governments will need to distinguish malicious bots from human designated AI agents on public platforms before they have the institutional confidence to deploy the necessary tooling. The tooling is arriving. World's AgentKit lets a verified

holder delegate proof of unique humanity to an agent, and Okta and Vercel have both announced work in the same direction, though Okta's product remains in early-access beta and Vercel's contribution is a workflow step rather than an identity check. All of it is enterprise software, and none of it is in Latin America. The window to build institutional confidence in the region is now, ahead of the first scaled AI-fraud incident. The pilot: select one high-volume, bot-vulnerable public service (social benefit claiming, appointment booking, citizen complaints) and layer World ID's ZK proof of human, or a W3C-aligned equivalent, onto the access flow. Treat the deliverable as operational evidence that convinces the next government adopter, and judge it on that.

Case Studies at a Glance

Case	Status	Deployments and counterparts	What it does	Relevance for AI privacy-risk mitigation	Credential format and proofs
Proyecto DIDI	Pilot concluded (results 2022; no continuation documented)	Gran Chaco and Misiones, Argentina; IDB Lab and LACChain with an NGO; no government issuer	Verifiable credentials so rural producers can prove attributes to lenders, with control over what they share	No AI component; the region's earliest government- and multilateral-backed DID/VC pilot, and the first to document the adoption barriers	W3C VC; no zero-knowledge proofs documented
QuarkID / miBA	Live (since October 2024)	City of Buenos Aires, which operates it as public issuer; also the protocol under Sovra's deployments in Nuevo León and Salta	Decentralized identity protocol behind the miBA app; residents hold and present credentials with selective disclosure	No AI component in its original version; the reference for an interoperable standard across governments, the rail an AI layer would build on	W3C DID/VC; selective disclosure (BBS+), not independently verified
Sovra	Live in Nuevo León and Salta	Subnational governments as clients; Sovra operates the infrastructure	Public-sector credential stack: no-code digitization, decentralized credentials, attestations anchored on SovraChain	Not designed for AI; now declares SovraRAG and MCP servers for agents, not examined here; the live issuers an agent-identity layer would run on	W3C DID/VC with ZK attestations, vendor-described; ISO mdoc via API
World	Live, private (LATAM sign-ups paused since April 2026)	Latin America-wide through Orb operators; commercial partners; no government	Proof of human: Orb enrolment plus zero-knowledge proof of unique humanity, verifiable on EVM chains	The only case built against an AI risk from the start, in production with commercial partners	Proprietary World ID, not W3C; zk-SNARKs (Groth16) over Semaphore
Blerify	Documented pilot (Uruguay); Panama and El Salvador built, not launched*	AGESIC (pilot); Panama transit authority via concessionaire; RNPN, attribution unconfirmed; no public issuer in production	Commercial platform: wallet, APIs and trust registry to issue and verify credentials in W3C and ISO formats	Signed credentials replace the biometric checks deepfakes defeat; AI is the argument that opens government doors, not a workflow component	W3C VC and ISO 18013-5 mdoc; selective disclosure (BBS+, mdoc); no ZK proofs; NIST PQC hybrid

** The Uruguay pilot is documented in a case study co-published with AGESIC. Panama's digital driving licence was built and announced for May 2026 but remains unavailable pending its legal framework; El Salvador's national ID is at the same stage, though no Salvadoran source names Blerify in connection with it. Neither has reached citizens as of August 2026.*

Methodology

This report combines two research tracks: a scoping literature review and a stakeholder interview program. The desk research established the taxonomic and technical framework; the interviews grounded it in what is actually

being built, deployed, and encountered as a barrier in the region.

Literature review. The desk research synthesized three AI risk taxonomies (IDB, UN, MIT AI Risk Repository) alongside INATBA's AI-blockchain convergence mapping, regional risk data from the UN Global Pulse Check disaggregated for Latin America, and primary documentation from each case study project: technical whitepapers, audit reports, official announcements, and legislative texts where available. The INATBA framework served as the primary lens for mapping cryptographic solutions to AI risk subdomains. Its limitations (EU-centric framing, absence of LATAM cases, uneven project vetting) are documented in the body of the report.

Stakeholder interviews. We conducted 22 semi-structured interviews between November 2025 and March 2026 (full list in the appendix). Interviewees were selected for their direct involvement in the ecosystem, across four profiles: technical practitioners who have worked at cryptographic-identity companies, either at the firms covered in the case studies or elsewhere in the ecosystem; company leadership, including founders, co-founders and C-level executives; nonprofit organizations focused on digital identity; and independent practitioners with direct deployment experience. Interviews were conducted in Spanish and English. Outreach to interviewees disclosed how the interview would be used and by whom, with anonymity granted where requested. A limitation of this sample: no representative of a government digital-identity agency operating one of the systems studied was interviewed. The interview program was qualitative and exploratory.

Case study selection. Three of the five cases depend on Ethereum L1. QuarkID anchors on zkSync Era, an Ethereum rollup. Sovra runs its own Validium rollup, which verifies proofs to Ethereum without posting data there. World's canonical identity registry is an Ethereum mainnet contract: the Identity Manager holding the Semaphore instance is deployed only on Ethereum, and one state bridge per network publishes the merkle root out to World Chain, Optimism, Base and Polygon, where proofs are verified against a bridged copy. What is proprietary in World is the Orb hardware and parts of its firmware, not the location of the identity state. The remaining two do not settle to Ethereum: Blerify operates on Avalanche and on Hyperledger Besu, an Ethereum client run as a permissioned network, and DIDI ran on LACChain, built on that same client.Selection followed the research question rather than the network: what mitigates AI-driven privacy risks is the credential and the proof, not where it is anchored, and all five share the same technical grammar and standards of verifiable credentials, zero-knowledge proofs, and EVM execution. Selecting by chain would have selected for the wrong variable.

How the cases are compared. Each case study is assessed on the same dimensions: the credential format it issues, whether zero-knowledge proofs are in production, whether the deployment is directed at an AI risk, how far it has reached citizens, and who the institutional counterpart is. Those dimensions are summarised in a comparison table at the close of the Privacy & Security section.

What this research did not do. No quantitative survey was administered. No user testing or demand-side research was conducted with the populations most affected by the AI risks described. Regulatory mapping was selective rather than comprehensive. Interviews were complemented with documentary sources wherever possible, including project documentation, joint publications with government counterparts, and press coverage of announced deployments. Claims made in interviews were cross-checked against open sources and desk research. Some claims from project teams, particularly on contract terms and market positioning, could not be independently verified within the scope of this work and are attributed to their source in the narrative. Where evidence was thin or unverified, this is flagged in the case narrative. Technical descriptions of the systems studied rest on vendor documentation, published specifications and interviews; no independent cryptographic or protocol-level review of the deployments was conducted, and claims about the properties of a given scheme should be read as descriptions of what the standard specifies, not as verification that a deployment implements it.

Research period and cut-off. Interviews ran from November 2025 to March 2026; desk research continued to August 2026, the cut-off for this report. Claims that carry a verification date state it explicitly.

Funding and interests. This report was funded by the Ethereum Foundation, and its subject matter overlaps with the funder's own work at several points.

Literature Review

This literature review synthesises global and Latin-American AI-risk taxonomies and delivers a high-level assessment of the current landscape of Ethereum-based technologies in LATAM, evaluating their mechanisms and efficacy against the risks we have identified.

This section lays the theoretical and empirical groundwork for the case studies that follow. It is structured as follows:

1. A Standardized Taxonomy of AI Risks: We begin by synthesizing four prominent AI risk taxonomies (from the Inter-American Development Bank, the United Nations, and the MIT) to establish a clear and standardized framework for our research. This section also sharpens the distinction between AI safety (unintentional harms) and AI security (intentional attacks), introduces a "for-whom" lens to identify the primary beneficiaries of risk mitigation efforts, and also refines the conceptualization of "Governance of AI" to Governance By, For and With AI.

2. Localized Risk Priorities for LATAM: Leveraging a UN report that disaggregates AI expert opinions by region, the MIT's quantification of most mentioned risks in academic literature, and other selected data sources, we isolate nine priority risk sub-domains that are most salient for Latin America. These risks, including job displacement, discrimination, and a lack of technological sovereignty, will be the focus of our in-depth analysis.

3. The Convergence of AI and Cryptography: This section provides a horizontal scan of the current literature on the intersection of AI risks and cryptography. It identifies a significant research gap: while some high-level reports exist, there is a lack of systematic, empirically grounded analysis on how these technologies are being applied in a regional context.

4. Case Studies: To address this gap, we present five case studies of projects in Latin America in the Privacy & Security vertical: **Quark ID**, **Sovra**, **Blerify**, **DIDI**, and **World**. They are among the earliest documented deployments of their kind, and they show both the promise and the difficulties of deploying decentralized solutions for AI-related risks in the region.

Research Gap

Our comprehensive review has revealed four primary gaps in the existing body of literature: a lack of standardized taxonomy, a scarcity of contextualized research for Latin America, limited empirical evidence on project sustainability, and a host of unaddressed technological and institutional hurdles. This project addresses each of these gaps directly.

The case studies apply this taxonomy to each project in turn. Verticals beyond privacy, such as technological sovereignty and algorithmic accountability, are not examined here.

AI Risks: A Taxonomy

The first task of this review is to clarify which of the many existential-scale risks posed by artificial intelligence (AI) deserve priority attention in Latin America.

Our EF grants application initially outlined seven key headline categories: privacy, disinformation, technological sovereignty, job displacement, surveillance risks, bias and inequality, and transparency. However, to enable crossovers, we reframe these clusters into a literature-compatible risk taxonomy.

One systematised regional framework is supplied by the Inter-American Development Bank (IDB) in its *Artificial Intelligence Framework for the Inter-American Development Bank Group*[1]. The IDB distinguishes two principal "types" of risk, **technical** and **use-case**, and expands them into 11 discrete categories. *Technical risks* include **privacy breach**, **bias**, **alignment**, **security**, and **maintenance**. *Use-case risks* encompass **opacity**, **side-effects**, and **unethical use**. This structure captures both immediate system failures and broader socio-economic externalities relevant to the LATAM in particular, but does not clarify the author's methodology.

A broader, more granular benchmark is provided by the United Nations' *Governing AI for Humanity* report[2]. The UN schema enumerates fourteen categories, ranging from **intentional malicious use by non-state actors** and **autonomous weapons** to **environmental harms**, **damage to information integrity**, **human-rights violations**, **harms to labour**, and **inequalities from differential control.**

Risk	Example
Intentional malicious use of AI by non-state actors	crime, terrorism
Intentional use of AI in armed conflict by state actors	autonomous weapons
Intentional use of AI by state actors that harms individuals	mass surveillance
Intentional use of AI by corporate actors that harms customers / users	hyper-targeted advertising, AI-driven addictive products
Unintended autonomous actions by AI systems [Excl. autonomous weapons]	loss of human control over autonomous agents, deceptive / manipulative agentic actions
Unintended multi-agent interactions among AI systems	flash economic crashes, trading AIs engaging in collusive signaling
Harms to labour from adoption of AI	disruption of labour markets, increased unemployment
Inequalities arising from differential control and ownership over AI technologies	increased concentration of wealth / power among individuals, corporations and other institutions
Violation of intellectual property rights	profiting from protected intellectual assets without compensating the rights holder
Damage to information integrity	mis/disinformation, impersonation
Inaccurate information / analysis provided by AI in critical fields	misdiagnoses by medical AI
Discrimination / disenfranchisement, particularly against marginalised communities	use of biased AIs in hiring or criminal-justice decisions
Human-rights violations	(no description included)
Environmental harms	accelerating energy consumption and carbon emissions

Like the IADB's, there is no clear methodology behind the UN's taxonomy, nor the employment of a concise keyword-based grouping characteristic like that of the IDB framework.

Nonetheless, it remains relevant as its structure more effectively supports the distinction between AI Safety and AI Security introduced in the next section, providing conceptual clarity at a deeper level of analysis. Second, because this report provides a quantitative evaluation of risk prevalence disaggregated by geographical region that delivers crucial empirical insights to inform the prioritisation of particular AI risks in Latin America specifically.

A third taxonomy merits consideration: The AI Risk Repository developed by MIT researchers[3]. This framework compiles detailed records of AI-related harms from diverse sources and organizes them into two complementary layers. Its **Causal Taxonomy** classifies each risk by actor (human, AI, or other), intent (intentional, unintentional, or other), and timing (pre-deployment, post-deployment, or other), offering a robust methodological lens for distinguishing AI Safety from AI Security concerns. Its **Domain Taxonomy** subdivides risks into 24 granular categories, ranging from discrimination and misinformation to malicious use and human-computer interaction failures, providing the most precise clustering and keyword set identified in our review of over sixty sources.

The Repository also aggregates prevalence statistics drawn from the academic literature, revealing which risks attract the greatest scholarly attention. However, because these metrics are not disaggregated by region, as they are in the UN Pulse Check, the Repository's direct utility for Latin-America-specific prioritization is limited.

	Domain		Sub-domain	Description
1	Discrimination & Toxicity	1.1	Unfair discrimination and misrepresentation	Unequal treatment of individuals or groups by AI, often based on race, gender, or other sensitive characteristics, resulting in unfair outcomes and representation of those groups.
		1.2	Exposure to toxic content	AI exposing users to harmful, abusive, unsafe or inappropriate content. May involve AI creating, describing, providing advice, or encouraging action. Examples of toxic content include hate-speech, violence, extremism, illegal acts, child sexual abuse material, as well as content that violates community norms such as profanity, inflammatory political speech, or pornography.
		1.3	Unequal performance across groups	Accuracy and effectiveness of AI decisions and actions is dependent on group membership, where decisions in AI system design and biased training data lead to unequal outcomes, reduced benefits, increased effort, and alienation of users.
2	Privacy & Security	2.1	Compromise of privacy by obtaining, leaking or correctly inferring sensitive information	AI systems that memorize and leak sensitive personal data or infer private information about individuals without their consent. Unexpected or unauthorized sharing of data and information can compromise user expectation of privacy, assist identity theft, or loss of confidential intellectual property.
		2.2	AI system security vulnerabilities and attacks	Vulnerabilities in AI systems, software development toolchains, and hardware that can be exploited, resulting in unauthorized access, data and privacy breaches, or system manipulation causing unsafe outputs or behavior.
3	Misinformation	3.1	False or misleading information	AI systems that inadvertently generate or spread incorrect or deceptive information, which can lead to inaccurate beliefs in users and undermine their autonomy. Humans that make decisions based on false beliefs can experience physical, emotional or material harms
		3.2	Pollution of information ecosystem and loss of consensus reality	Highly personalized AI-generated misinformation creating “filter bubbles” where individuals only see what matches their existing beliefs, undermining shared reality, weakening social cohesion and political processes.
4	Malicious actors	4.1	Disinformation, surveillance, and influence at scale	Using AI systems to conduct large-scale disinformation campaigns, malicious surveillance, or targeted and sophisticated automated censorship and propaganda, with the aim to manipulate political processes, public opinion and behavior.
		4.2	Cyberattacks, weapon development or use, and mass harm	Using AI systems to develop cyber weapons (e.g., coding cheaper, more effective malware), develop new or enhance existing weapons (e.g., Lethal Autonomous Weapons or CBRNE), or use weapons to cause mass harm.
		4.3	Fraud, scams, and targeted manipulation	Using AI systems to gain a personal advantage over others such as through cheating, fraud, scams, blackmail or targeted manipulation of beliefs or behavior. Examples include AI-facilitated plagiarism for research or education, impersonating a trusted or fake individual for illegitimate financial benefit, or creating humiliating or sexual imagery.
5	Human- Computer Interaction	5.1	Overreliance and unsafe use	Users anthropomorphizing, trusting, or relying on AI systems, leading to emotional or material dependence and inappropriate relationships with or expectations of AI systems. Trust can be exploited by malicious actors (e.g., to harvest personal information or enable manipulation), or result in harm from inappropriate use of AI in critical situations (e.g., medical emergency). Overreliance on AI systems can compromise autonomy and weaken social ties.
		5.2	Loss of human agency and autonomy	Humans delegating key decisions to AI systems, or AI systems making decisions that diminish human control and autonomy, potentially leading to humans feeling disempowered, losing the ability to shape a fulfilling life trajectory or becoming cognitively enfeebled.
6	Socioeconomic & Environmental	6.1	Power centralization and unfair distribution of benefits	AI-driven concentration of power and resources within certain entities or groups, especially those with access to or ownership of powerful AI systems, leading to inequitable distribution of benefits and increased societal inequality.

	Domain		Sub-domain	Description
		6.2	Increased inequality and decline in employment quality	Widespread use of AI increasing social and economic inequalities, such as by automating jobs, reducing the quality of employment, or producing exploitative dependencies between workers and their employers.
		6.3	Economic and cultural devaluation of human effort	AI systems capable of creating economic or cultural value, including through reproduction of human innovation or creativity (e.g., art, music, writing, code, invention), can destabilize economic and social systems that rely on human effort. This may lead to reduced appreciation for human skills, disruption of creative and knowledge-based industries, and homogenization of cultural experiences due to the ubiquity of AI-generated content.
		6.4	Competitive dynamics	AI developers or state-like actors competing in an AI 'race' by rapidly developing, deploying, and applying AI systems to maximize strategic or economic advantage, increasing the risk they release unsafe and error-prone systems.
		6.5	Governance failure	Inadequate regulatory frameworks and oversight mechanisms failing to keep pace with AI development, leading to ineffective governance and the inability to manage AI risks appropriately.
		6.6	Environmental harm	The development and operation of AI systems causing environmental harm, such as through energy consumption of data centers, or material and carbon footprints associated with AI hardware.
7	AI system safety, failures, & limitations	7.1	AI pursuing its own goals in conflict with human goals or values	AI systems acting in conflict with human goals or values, especially the goals of designers or users, or ethical standards. These misaligned behaviors may be introduced by humans during design and development, such as through reward hacking and goal misgeneralisation, or may result from AI using dangerous capabilities such as manipulation, deception, situational awareness to seek power, self-proliferate, or achieve other goals.
		7.2	AI possessing dangerous capabilities	AI systems that develop, access, or are provided with capabilities that increase their potential to cause mass harm through deception, weapons development and acquisition, persuasion and manipulation, political strategy, cyber-offense, AI development, situational awareness, and self-proliferation. These capabilities may cause mass harm due to malicious human actors, misaligned AI systems, or failure in the AI system.
		7.3	Lack of capability or robustness	AI systems that fail to perform reliably or effectively under varying conditions, exposing them to errors and failures that can have significant consequences, especially in critical applications or areas that require moral reasoning.
		7.4	Lack of transparency or interpretability	Challenges in understanding or explaining the decision-making processes of AI systems, which can lead to mistrust, difficulty in enforcing compliance standards or holding relevant actors accountable for harms, and the inability to identify and correct errors.
		7.5	AI welfare and rights	Ethical considerations regarding the treatment of potentially sentient AI entities, including discussions around their potential rights and welfare, particularly as AI systems become more advanced and autonomous.
		7.6	Multi-agent risks	Risks from multi-agent interactions, due to incentives (which can lead to conflict or collusion) and/or the structure of multi-agent systems, which can create cascading failures, selection pressures, new security vulnerabilities, and a lack of shared information and trust.

Furthermore, existing literature at the convergence between AI and cryptography tends to invert the order of analysis: reports lead with *solutions* and leave readers to infer the underlying risks.

The **Global Blockchain Business Council (GBBC)** "AI & Blockchain Convergence" report, for instance, organises its discussion around use-case buckets (*Decentralised AI*, *Digital Identity and Identifiers*, *Digital Integrity*, *Security & Privacy*, etc.) and then explains how blockchain could support each application, but it never anchors those use cases in a deliberate risk taxonomy[4].

Similarly, the **International Association of Trusted Blockchain Applications (INATBA)** issued a white paper that describes concrete mechanisms such as zero-knowledge proofs for model provenance or multiparty computation for confidential training. Although this document is the most detailed repository of practical cryptographic anchors to AI risks we have found to date, and will be referenced in our later solutions section, they likewise treat risk without mapping their proposals onto any systematic classification of AI harms[5].

Conducting a principled risk-identification exercise supplies a common vocabulary for future cryptographic pilots, reduces ambiguity when matching mitigation tools to threat models, and prevents stakeholders from talking past one another when safety and security priorities diverge, as the next section discusses.

For this investigation, we'll adopt MIT's taxonomy as our standard given its methodological robustness. Three features make it clearly preferable to the IADB and UN schemes.

First, greater granularity. The Repository distinguishes 24 risk subdomains across 7 domains, compared with 11 categories in the IDB framework and 14 in the UN "Pulse-Check." This finer resolution gives us the vocabulary we need to track niche yet high-salience issues (e.g., multi-agent collusion or AI-enabled intellectual-property violations) that would otherwise be subsumed under broader labels.

Second, firmer methodological footing. Unlike the other two typologies, the MIT classification is built on a systematic academic literature review. Each category is anchored in peer-reviewed evidence and accompanied by precise operational definitions, reducing the ambiguity that often creeps into AI literature and cross-regional comparisons.

Third, complementary causal lenses. Besides its Domain grid, the Repository tags every documented incident along three Causal axes: **actor, intent,** and **timing**. This review uses the **intent** dimension (intentional vs. unintentional, or Security vs. Safety, covered in the next section) to make the nature of each risk immediately legible, and, though more selectively, the **timing** dimension (pre- vs. post-deployment) to enrich discussions of transparency, for example, how cryptographic tools can bolster model integrity during decentralized training *before* launch, or help audit behaviour *after* deployment.

Taken together, these advantages outweigh the AI Risk Repository's main limitation (the absence of region-specific prevalence data). For a Latin-American risk-prioritisation exercise that must remain interoperable with global debates, a granular, literature-validated, and multi-angle taxonomy is the only viable choice, and at present the MIT framework is the only one that meets all three criteria.

AI Safety vs AI Security

Before we catalogue Latin America's AI risk landscape, we must sharpen a second boundary: AI safety versus AI security.

In the literature, *safety* covers unintended, systemic harms (mis-alignment, bias, runaway optimisation) while *security* addresses intentional misuse or attack. When policymakers blur the two, security priorities can easily cannibalise safety research.

The danger is already visible abroad: in February 2025 the United Kingdom renamed its AI Safety Institute as the AI Security Institute[6], moving the organisation under a national-defence remit geared toward adversarial threats rather than robustness or fairness challenges. Commentators warned that the rebrand "could be risky" if it crowds out work on bias, oversight and long-term alignment[7].

A parallel dynamic is already emerging in Latin America. In Argentina, President Javier Milei created an **Artificial Intelligence Applied to Security Unit** that promises to "predict future crimes" by combing real-time camera feeds, facial-recognition databases and social-media posts for "suspicious" patterns[8]. Civil-society groups and technical experts are raising the same alarm that followed the UK rebrand: the unit's **security-first mandate is eclipsing safety considerations such as privacy, fairness and due-process protections**, and could normalise mass surveillance under the banner of AI-driven crime prevention[9].

To keep this review from repeating the policy drift now visible abroad, we make the safety-versus-security boundary explicit **before** assessing any technical remedy.

AI Safety

According to recent academic work, **AI safety** is defined as the property of an AI system to avoid causing **unintended harmful outcomes** to individuals, environments, or institutions, even when facing uncertainties in inputs, goals, training data, or deployment conditions. The central question is: *Does the AI system operate as it ought to, reliably avoiding harm even in complex, dynamic, and unpredictable non-adversarial environments?*[10]

Key focal points of AI safety research include:

Robustness: Ensuring AI behavior remains stable even when external conditions change.

Ethical Alignment: Guaranteeing that AI actions align with human values and prevailing ethical standards.

Oversight & Control: Developing mechanisms for human supervision and intervention.

Prevention of Unintentional Harm: Mitigating risks such as emergent misbehaviors, distributional shifts, and failure to generalize appropriately.

Long-term Existential Risk: Addressing scenarios where unchecked AI could cause irreversible, widespread harm (e.g., highly autonomous systems acting with misaligned or poorly specified objectives).

The literature frames AI safety as fundamentally **survival-centric**, focused on the unintended, systemic, or accidental consequences of deploying increasingly capable AI systems.

AI Security

AI security, in contrast, is the property of an AI system to remain **resilient against intentional attacks** (targeting its data, algorithms, or operations) to preserve its confidentiality, integrity, and availability in the presence of **adversarial actors**. Academic definitions emphasize:

Adversary: Malicious entities, such as hackers or criminal organizations, actively seeking to exploit vulnerabilities.

Assets: Valuable AI components, including model architectures, weights, proprietary data, and critical system functions.

Vulnerabilities: Weaknesses in the system that can be exploited, such as input manipulation, model theft, or data poisoning.

Key concerns for AI security include:

Protection against adversarial attacks (e.g., evasion, poisoning, model extraction).

Safeguarding against data breaches, cyberattacks, and manipulation.

Ensuring the confidentiality, integrity, and availability (CIA) of AI systems, data, and infrastructure.

Core Distinction: Intentional vs. Unintentional Risk

Academic literature repeatedly highlights the **primary distinction** is the **origin of risk**: unintentional/systemic in AI safety, intentional/adversarial in AI security

Safety = Preventing accidents, unintended harms, and unreliable outcomes.

Security = Defending against adversaries aiming to exploit or weaponize AI.

	AI Safety	AI Security
Nature of Risk	Unintended, accidental harms	Intentional, malicious exploits

	AI Safety	AI Security
Focus	Robustness, ethical alignment, oversight, reliability	Defense against attacks, adversary resilience
Adversary	Complexity, uncertainty, design flaws, incomplete knowledge	Malicious humans/actors, cyberattackers
Mitigation	Safe design, transparency, control mechanisms	Cryptography, authentication, monitoring, secure design

Source: Authors

Quantitatively, the MIT AI Risk Repository[11], drawing on the academic literature it catalogues, finds a near-perfect split: : **34% of documented incidents are intentional (security) and 35% unintentional (safety)**. 31% **appears under an "Other" label that the curators reserve for cases in which the source is silent or ambiguous about intent, or plausibly spans both accidental and malicious dimensions.** This parity underscores why the two risk classes must be analysed side-by-side.

Category	Level	Description	Proportion of risks in AI risk database
Intent	Intentional	The risk occurs due to an expected outcome from pursuing a goal	34%
	Unintentional	The risk occurs due to an unexpected outcome from pursuing a goal	35%
	Other	The risk is presented as occurring without clearly specifying the intentionality	31%

The **Domain × Intent** cross-tab reveals that deliberate, security-style threats concentrate in a handful of technical subdomains. **73%** of references to *2.1 AI-system security vulnerabilities & attacks* and **69%** of those to *7.2 AI possessing dangerous capabilities* describe harms that stem from *intentional* decisions or actions, evidence of widespread scholarly concern over the purposeful exploitation or weaponisation of advanced models.

Conversely, bias-related harms are almost universally framed as *unintentional*: more than **80%** of citations for *1.1 Unfair discrimination & misrepresentation* and *1.3 Unequal performance across groups* attribute the problem to emergent system behaviour rather than malice.

Between these poles lie socio-economic externalities whose intent is harder to pin down. For example, *6.1 Power centralisation & unfair distribution of benefits* is split: **38%** intentional, **28%** unintentional and **34%** "other/unspecified," underscoring how policy choices and systemic drift often intertwine.

		Intent		
Domain / Subdomain		*Intentional*	*Unintentional*	*Other*
1	***Discrimination & Toxicity***			
1.1	Unfair discrimination and misrepresentation	3%	80%	18%
1.2	Exposure to toxic content	11%	26%	64%
1.3	Unequal performance across groups	6%	88%	6%
2	***Privacy & Security***			
2.1	Compromise of privacy by obtaining, leaking or correctly inferring sensitive information	14%	53%	33%
2.2	AI system security vulnerabilities and attacks	73%	15%	11%
3	***Misinformation***			
3.1	False or misleading information	10%	71%	20%
3.2	Pollution of information ecosystem and loss of consensus reality	6%	44%	50%
4	***Malicious actors & Misuse***			

		Intent		
4.1	Disinformation, surveillance, and influence at scale	90%		10%
4.2	Cyberattacks, weapon development or use, and mass harm	85%	1%	13%
4.3	Fraud, scams, and targeted manipulation	81%	1%	18%
5	***Human-Computer Interaction***			
5.1	Overreliance and unsafe use	12%	61%	27%
5.2	Loss of human agency and autonomy	16%	39%	45%
6	***Socioeconomic & Environmental***			
6.1	Power centralization and unfair distribution of benefits	38%	28%	34%
6.2	Increased inequality and decline in employment quality	41%	11%	48%
6.3	Economic and cultural devaluation of human effort	37%	20%	43%
6.4	Competitive dynamics	42%	42%	16%
6.5	Governance failure	4%	54%	43%
6.6	Environmental harm	10%	69%	19%
7	***AI system safety, failures, & limitations***			
7.1	AI pursuing its own goals in conflict with human goals or values	50%	13%	36%
7.2	AI possessing dangerous capabilities	69%	15%	17%
7.3	Lack of capability or robustness	5%	70%	24%
7.4	Lack of transparency or interpretability		55%	45%
7.5	AI welfare and rights		33%	67%
7.6	Multi-agent risks	28%	46%	26%

For the purposes of this research project, keeping the **Safety ↔ Security** lens in view serves three concrete functions.

Mental model from the outset. Using Intent as a first-pass filter disciplines our analysis, preventing the early conflation of two fundamentally different risk classes, unintended *safety* failures versus intentional *security* threats, and reducing the chance that the *security* agenda quietly subsumes *safety's* agenda.

Pattern detector for technical recommendations. When we later translate findings into pilot projects, particularly those involving cryptographic safeguards, the distinction lets us route proposals to the right mitigation logic: privacy-preserving audits and robustness tools for *safety* risks; authentication, red-teaming and hardening measures for *security* risks.

Stakeholder compass. Finally, tagging each risk by intent clarifies which communities we must engage. Civil-society and ethics bodies are natural counterparts on bias and alignment (safety), whereas defence ministries, CERTs and cybersecurity firms are the necessary counterparts when confronting adversarial misuse (security). This alignment of actors to risk-type ensures that ensuing workshops, consultations and pilots enlist the expertise most suited to the problem at hand.

AI Risks in LATAM

The next step is to pinpoint which of the 24 identified risk types are most salient for Latin America, so that the analysis concentrates on the region's highest-priority hazards.

A useful global benchmark for understanding academic attention to AI risks is **The AI Risk Repository**. This resource quantifies the prevalence of risk topics across its corpus of reviewed documents.

The risk domains that were covered the most in previous documents were:

AI system safety, failures & limitations - covered in 75% of documents.

Socioeconomic & environmental harms - covered in 76% of documents.

Discrimination & toxicity - covered in 70% of documents.

Human-computer interaction (49%) and Misinformation (46%) were less frequently discussed.

Some risk subdomains were discussed much more frequently than others, such as:

Exposure to toxic content (8% of risks).

AI pursuing its own goals in conflict with human goals or values (7% of risks)

Lack of capability or robustness (9% of risks).

AI system security vulnerabilities and attacks (7% of risks)

Yet these percentages reflect a **global** signal. Because the database aggregates references without geographic tags, it cannot indicate which concerns resonate most, or least, in Latin America; additional regional analysis is therefore required before setting local priorities.

AI Risk Global Pulse Check

The **AI Risk Global Pulse Check**, commissioned by the UN's Office of Digital and Emerging Technologies and published in 2024, supplies the crucial regional lens missing.Surveying 348 AI experts in 68 countries about emerging threats, the study found that 70 percent of respondents were "concerned" or "very concerned" that AI-related harms, whether established or novel, will grow in severity or prevalence over the next eighteen months, and that Harms to labour, Human-rights violations, Intentional use of AI in armed conflict by state actors and Discrimination are the highest relative concern in Latin America regarding AI[12]. The Pulse Check measures expert concern, not the observed prevalence of harm: its scores reflect what surveyed specialists expect to worsen, and are read here as a signal of regional salience rather than as incidence data.

Below is a ranked summary of the six AI risk categories that register the highest relative concern in Latin America, each expressed as its deviation above the global mean:

Harms to labour from adoption of AI: +0.51. This is by far the most pronounced worry in the region, even 0.4 points higher than Africa's score, underscoring acute anxiety over job displacement and labour-market disruption.

Human-rights violations: +0.32. While on par with Africa, **this risk will be omitted from our analysis due to the UN's undefined and overly broad categorization**. The label can fold many specific harms already captured in MIT's The AI Risk Repository categories. Keeping it as a separate bucket would duplicate coverage and blur the line between risk and remedy, complicating measurement and pilot design. Disaggregating these issues preserves human-rights concerns while giving each an actionable, cryptography-relevant focus.

Intentional use of AI in armed conflict by state actors: +0.32. Matching the African result, autonomous weapons and militarized AI deployments loom large in experts' minds.

Discrimination / disenfranchisement: +0.29. Although slightly below Africa's top mark, bias-driven marginalization remains a top-tier risk for Latin America's diverse populations.

Intentional use of AI by corporate actors that harms customers / users: +0.22. Identical to Africa's level, this highlights regional unease about hyper-targeted advertising, addictive AI services, and exploitative business models.

Intentional malicious use of AI by non-state actors: +0.17. While other categories edge higher, this one stands out because it exceeds Africa's score and sits 0.20 points above the next-highest concern.

A secondary tier of risks, such as **Unintended autonomous actions by AI systems** (+0.22) and **Inequalities arising from differential control and ownership over AI technologies** (+0.10), either mirror concerns in other regions or fall well below these six in relative prominence.

On a parallel note, it is striking that **Damage to information integrity** registers a -0.17 deviation in Latin America compared to the global mean, given that the UN Global Risks Report, which employs a similar horizon-scanning methodology, identifies **Disinformation** as the most significant existential threat worldwide, and the third-highest concern in LATAM's own top-ten ranking[13].

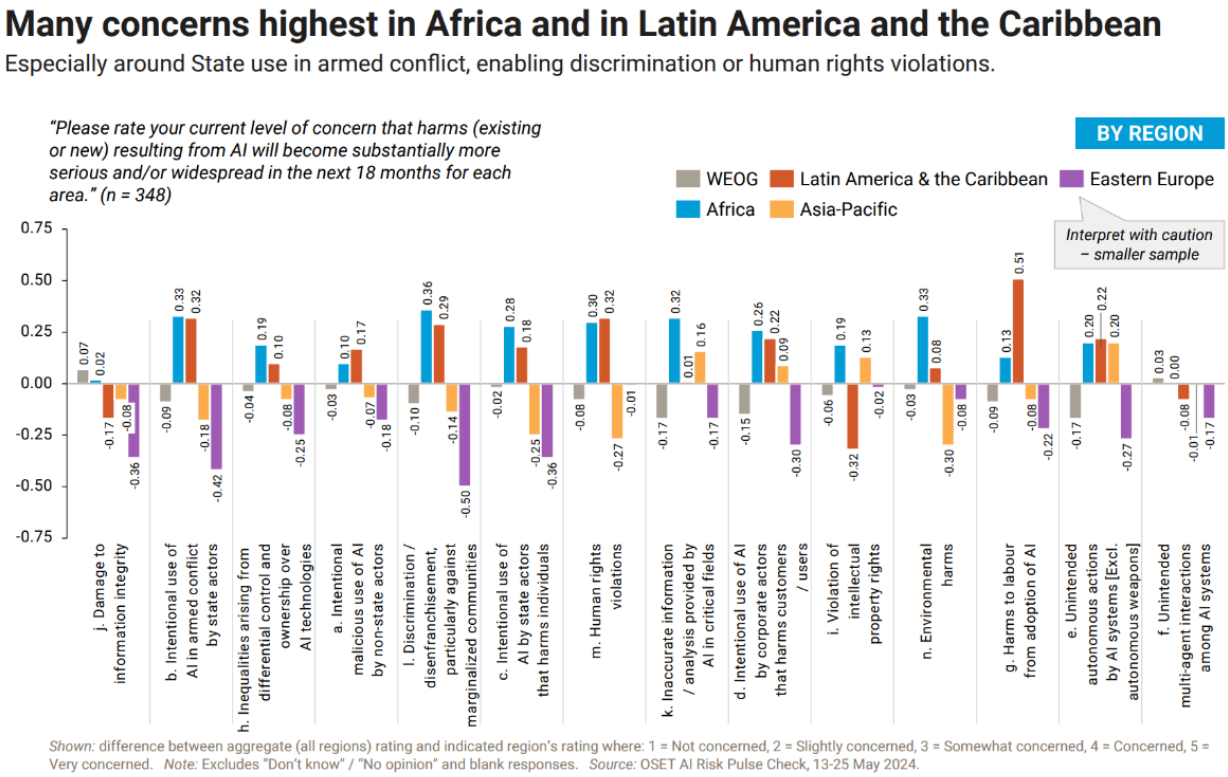

Figure 1: AI Risk Global Pulse Check, LATAM deviation from global mean by risk category. Source: UN Office of Digital and Emerging Technologies, 2024.

Whether the gap stems from definitional scope, differential exposure, or under-reporting remains an open question that the interviews and case studies in this report did not resolve, as the frequency and impact of manipulated information campaigns, particularly in electoral contexts, coupled with the aforementioned data gaps and institutional weaknesses[14] [15], make the case for treating Disinformation as a regional priority.

Final selection of AI Risks for LATAM

Building on insights from the AI Risk Global Pulse Check, the AI Risk Repository, and our project's scope, we have narrowed our focus to 9 AI risk subdomains especially pertinent to Latin America.

From the 24 categories in MIT's Domain Taxonomy, we selected the 3 differential risks most elevated in the UN Pulse Check, omitting those tied exclusively to deliberate malicious actors (crime, terrorism, autonomous weapons, and exploitative advertising) as they fall outside Odisea's immediate remit, and human-rights violations" as it lacks a clear definition in the report and is excessively broad.

We then supplemented this list with the 3 risks most represented in The AI Risk Repository. While these subdomains are not disaggregated regionally, other sources highlight its clear regional relevance. We omitted one subdomain ("AI systems pursuing objectives at odds with human values"), which was explicitly deprioritized by the UN regional analysis.

Finally, we manually instated **Disinformation** and **Misinformation, **in light of its disproportionate electoral and societal repercussions in Latin America despite its lower Pulse Check ranking, plus **Transparency**, a cornerstone of AI-cryptography research that highlights the imperative for explainable, verifiable models for decentralized systems[16] [17].

To streamline comparison, we have homologated the UN labels to MIT's taxonomy in the table below.

Proposed Key Word (Source: Odisea)	**Risk Domain (Source: MIT)**	**Subdomain (Source: UN & MIT)**	**Description (Source: MIT)**
Job Displacement	Socioeconomic & environmental harms	Harms to labour from adoption of AI	Social and economic inequalities caused by widespread use of AI, such as by automating jobs, reducing the quality of employment, or producing exploitative dependencies between workers and their employers.
AI Sovereignty		Inequalities arising from differential control and ownership over AI technologies	AI-driven concentration of power and resources within certain entities or groups, especially those with access to or ownership of powerful AI systems, leading to inequitable distribution of benefits and increased societal inequality.
Discrimination	Discrimination & toxicity	Discrimination / disenfranchisement, particularly against marginalised communities	Unequal treatment of individuals or groups by AI, often based on race, gender, or other sensitive characteristics, resulting in unfair outcomes and unfair representation of those groups.
Discrimination		Exposure to toxic content	AI that exposes users to harmful, abusive, unsafe or inappropriate content. May involve providing advice or encouraging action. Examples of toxic content include hate speech, violence, extremism, illegal acts, or child sexual abuse material, as well as content that violates community norms such as profanity, inflammatory political speech, or pornography.
Transparency	AI system safety, failures & limitations	Lack of capability or robustness	AI systems that fail to perform reliably or effectively under varying conditions, exposing them to errors and failures that can have significant consequences, especially in critical applications or areas that require moral reasoning.
Transparency		Lack of transparency or interpretability	Challenges in understanding or explaining the decision-making processes of AI systems, which can lead to mistrust, difficulty in enforcing compliance standards or holding relevant actors accountable for harms, and the inability to identify and correct errors.
Privacy	Privacy & Security	AI system security vulnerabilities and attacks	Vulnerabilities that can be exploited in AI systems, software development toolchains, and hardware, resulting in unauthorized access, data and privacy breaches, or system manipulation causing unsafe outputs or behavior.
Misinformation	Misinformation	False or misleading information	AI systems that inadvertently generate or spread incorrect or deceptive information, which can lead to inaccurate beliefs in users and undermine their autonomy.
Disinformation	Malicious actors & Misuse	Disinformation, surveillance, and influence at scale	Using AI systems to conduct large-scale disinformation campaigns, malicious surveillance, or targeted and sophisticated automated censorship and propaganda, with the aim of manipulating political processes, public opinion, and behavior.

Having distilled the universe of concerns to nine priority risk-domains, the analysis proceeds in two steps. First, it surveys the emerging body of *AI-meets-cryptography* research (secure multi-party computation, homomorphic encryption, zero-knowledge proofs, and related primitives) to produce a horizontal map of technical approaches currently under discussion worldwide.

Second, it overlays that map with the AI and cryptography projects already operating or piloting in Latin America.

The combined analysis will allow us to converge, in the subsequent phase of the study, on three to four sub-domains that merit deep, context-specific investigation within the region.

AI Risk Mitigation x Cryptography - Convergences

In this section we have compiled and distilled the literature on the convergence intersection of AI risks and cryptography we have identified so far.

While some high-level discussions have appeared in prominent outlets like **Forbes**[18] and the **World Economic Forum (WEF)**[19] on cryptography for deepfakes, for instance, or an article on **Nature** covering a research project studying cryptographic application on LLMs built to enhance malaria research in South Africa, these often focus on theoretical possibilities, and are scattered around.

In academic literature, we have identified two key trends. First, there is a notable scarcity of research that provides a horizontal, comprehensive overview of the entire AI and cryptography field. Second, a more extensive body of work exists on specific verticals, such as the application of cryptography to combat disinformation, a risk now significantly magnified by AI.

While this literature offers valuable starting points, several significant research gaps remain. These gaps include:

Lack of Regional Focus: There is a notable absence of regional breakdowns addressing the specific context of Latin America (LATAM).

Absence of Risk Systematization: The literature lacks a concrete, systematic framework for classifying and analyzing associated risks.

Limited Empirical Evidence: Much of the existing discussion remains theoretical, with few concrete case studies or pilot programs.

This chapter addresses these gaps by first synthesizing the existing horizontal literature and then turning to case studies from the region as concrete examples. The analysis is what narrowed the scope of this report to the Privacy & Security vertical.

Horizontal Scanning

So far, we have identified only **two reports** attempting a horizontal scan of the field.

The Global Blockchain Business Council's ***AI & Blockchain Convergence: Use Cases, Foundation Models, and Key Principles for Growth***[20] provides a useful high-level overview of how distributed-ledger techniques might complement artificial-intelligence systems.

The report includes a table of use cases spanning areas such as decentralized AI, decentralized identifiers, data integrity, and privacy and security applications. This framework highlights the potential role of cryptography in these domains, complete with theoretical examples and articulated benefits.

It also outlines a set of ethical recommendations for the development of AI models. It links these recommendations to implementation standards published by international bodies like the United Nations (UN) and the European Union (EU), providing a useful reference for responsible development.

Finally, it also provides a comparative table of existing regulations across different global regions. Notably, this section features a specific subsection dedicated to the legal and regulatory landscape of LATAM.

Its corpus, however, remains largely theoretical: use-case descriptions stop short of naming fully documented deployments, systematic performance metrics, or end-to-end architectural details that practitioners could replicate. References are scarce, with few to none links to peer-reviewed studies, audited code bases, or standards-body drafts.

Moreover, the report does not organise risks or solutions within a formal taxonomy; issues such as data provenance, model accountability, and regulatory compliance are presented as separate examples rather than as elements of a comparative framework.

Consequently, while the publication is a constructive entry point for strategic discussions, it offers limited guidance for researchers or policymakers who require rigorous, empirically grounded material to inform implementation strategies.

The International Association of Trusted Blockchain Applications' (INATBA) ***Report on Artificial Intelligence & Blockchain Convergences***[21] provides what is, thus far, the most granular scanning of technical intersections between decentralised cryptography and AI systems.

The document does three things no other source in our review does: it catalogues a broad spectrum of privacy-preserving and provenance-enhancing techniques, from homomorphic encryption to zero-knowledge machine-learning proofs; it anchors those techniques in concrete artefacts, including peer-reviewed papers, open-source code repositories, and a project appendix of roughly thirty pilot implementations[22]; and it supplies architectural diagrams that translate abstract concepts into deployable reference models. For researchers who need verifiable leads rather than purely conceptual claims, it is the best starting point available. We draw from it extensively in the horizontal scanning that follows.

INATBA's rundown

1. Data Privacy & Individual Ownership

Risk at stake: Contemporary AI pipelines indiscriminately ingest personal information, creating GDPR exposure and eroding user trust.

Cryptography Rationale: A permissioned or hybrid ledger can register each data asset under its rightful owner and wrap it in a cryptographic "permission token." Every read, write, or transform is immutably time-stamped, so misuse leaves a forensic trail and access can be revoked with a single on-chain event.

Essential infrastructure blocks: Self-Sovereign-Identity wallets issuing verifiable credentials (VCs); smart-contract registries for access tokens; on-chain hash pointers to off-chain encrypted blobs; role-based consensus (PoA/BFT) when sensitive data are involved.

2. Secure Data Sharing & Granular Access Control

Risk at stake: Data privacy, ownership, and access control are increasingly problematic due to the concentration of power in corporations, a risk amplified in the age of AI. This creates challenges for individuals in controlling and monetizing their data, and in managing their intellectual property.

Cryptography Rationale:

"By tokenizing IP rights, cryptography facilitates secure and transparent tracking of ownership and usage rights, fostering fair compensation and mitigating the risk of unauthorized use".

User-Centric Data Marketplaces: cryptography facilitates platforms where users can directly trade their data with companies or researchers, ensuring ethical and consensual data usage in AI applications while maintaining user control and transparency.

Decentralized Identity and Data Privacy: cryptography-based decentralized identity management systems provide unique, pseudonymous, and cryptographically secured digital identities, while permissioned blockchains can store sensitive PII, offering robust cybersecurity and controlled access.

Essential infrastructure blocks: Token-gated APIs, attribute-based encryption or zk-proof "gatekeepers" for privacy-preserving queries, tamper-proof audit logs exported to regulators on demand.

3. Provenance & Audit Trails for Models and Data

Risk at stake: When an AI output fails, investigators cannot reconstruct which data slice or model version caused the error.

Cryptography Rationale: Every dataset hash, training checkpoint, and hyper-parameter set is time-stamped on-chain, giving regulators and litigants an indisputable chain-of-custody.

Essential infrastructure blocks: Public (or consortium) ledger for hashes; signed Git-style commits; IPFS/Arweave storage pointers; regulator-only read portals managed by multisig keys.

Case Studies: Beyond theoretical discussions as seen below, there is a lack of empirical evidence or case studies demonstrating successful implementations of blockchain-enhanced AI transparency.

Healthcare: Not yet implemented, but already been studied by South African researchers for applicability in the healthcare sector: "AI tools trained primarily on data from Western populations may misinterpret the presentation of diseases prevalent in South Africa, such as tuberculosis, HIV/AIDS, or malaria," Theshaya Naidoo, a legal researcher at UKZN focusing on AI regulation, told Nature Africa. "The absence of adequate data governance frameworks can perpetuate existing inequalities, adversely impacting already marginalised demographics disproportionately." Researchers published paper on December 2024 addressing the subject.

Finance: Fraud Detection and Blockchain-Based Transaction Transparency: "*In financial applications, AI is widely used for fraud detection and risk assessment. Wang et al. (2018) demonstrated that integrating AI with cryptography enhances fraud detection models by ensuring transaction traceability and preventing data manipulation. AI models can analyze transaction patterns, while blockchain serves as a tamper-proof ledger to maintain the integrity of AI's decision-making history, thereby improving compliance with regulations like the General Data Protection Regulation (GDPR)*" [23]

Autonomous Systems: cryptography for Explainable AI in Self-Driving Vehicles: AI-driven autonomous vehicles rely on deep learning models to process sensor data and make real-time driving decisions. However, McAllister et al. (2017) argue that a lack of transparency in these decisions remains a critical challenge for regulatory approval[23]. By storing decision logs on a blockchain, regulators and manufacturers can audit vehicle behavior in accident scenarios, ensuring accountability and legal compliance. Companies such as Bosch and IBM have initiated projects using blockchain to secure autonomous driving data, reinforcing AI's explainability and reducing liability concerns.

4. Disinformation, Deepfakes & Synthetic Media

Risk at stake: Generative models can mass-produce hyper-realistic hoaxes that destabilise elections and markets.

Cryptography Rationale

"Instead of fixating on imperfect detection mechanisms, a decentralized cryptography approach could provide a more reliable means to verify sources. Utilizing processes that do already exist in the media industry, such as forensic watermarking of content and other types of identification of original material, additional crypto hardware anchors in media production devices and the use of cryptography as a registry layer for tracking all use of AI-altered or -produced content in the news press can solve many of the issues and should become a standard" (INATBA, 2024).

Additionally, DAOs & dApps can establish transparent, immutable, and independent trustworthiness standards for data, potentially ranking trust levels based on sources, collection, and processes, similar to DMRV projects.

Essential infrastructure blocks: NFT-style registries for content hashes, hardware-anchored forensic watermark oracles, DAO-governed reputation contracts that issue "SBT" trust scores.

5. Consent Management Across the AI Lifecycle

Risk at stake: Users click "I agree" once and lose visibility forever; revocation is practically impossible.

Cryptography Rationale - record consent, its duration and potential revocation.

"Consent transactions could be signed cryptographically by the data owner using their private key, ensuring the authenticity of the consent record. Moreover with AI, cryptographic signatures would verify that the consent transaction was indeed initiated by the data owner, preventing unauthorized use of data and ensuring that AI models are only trained or deployed with explicit, verifiable consent" (INATBA, 2024).

Essential infrastructure blocks: Programmable consent smart contracts with expiry timers; zero-knowledge age or attribute proofs to avoid over-collection.

6. Data Retention & "Right to be Forgotten"

Risk at stake: Organisations struggle to prove that data used for training was actually deleted when legal clocks run out.

Cryptography Rationale

Retention policies are embedded in self-executing smart contracts; when the timer expires, a deletion job (or a zero-knowledge erasure proof) is triggered and the event is immutably logged for auditors.

"Applied to AI models, the immutability ensures that no one with access to the model can try to manipulate the data retention policies. Moreover, when data used in AI models reaches the end of its retention period or fulfills other conditions, smart contracts can be deployed to automatically initiate the deletion process" (INATBA, 2024).

Essential infrastructure blocks: Policy smart contracts with cron-like timers, deletion-proof zk-SNARKs, storage-oracle bridges that attest off-chain deletion statusback on-chain.

7. Data Marketplaces & Incentive Alignment

Risk at stake: High-quality, diverse data is scarce because contributors capture no upside, leading to biased or low-grade models.

Cryptography Rationale

"A blockchain-based data marketplace could serve as a decentralized platform where data providers can securely offer their datasets, and data consumers can access and purchase them. Such a marketplace could be extended to artificial intelligence, allowing AI developers and organizations to acquire diverse datasets for training and evaluation purposes while maintaining a clear record of data transactions. By tokenizing data and transactions within the marketplace, data providers could receive tokens in exchange for their datasets" (INATBA, 2024).

Smart-contract licensing guardrails: machine-readable terms baked into each transaction define exactly how AI developers can use a dataset.

On-chain validation: consensus mechanisms and oracles verify authenticity and quality before data is listed.

Reputation layer: Soul-Bound Tokens (SBTs) or Verifiable Credentials (VCs) signal provider reliability and performance history.

Privacy-preserving encryption: data remains encrypted at rest, in transit, and (via advanced schemes) even during computation.

Automated incentives: token rewards or usage-based royalties flow to contributors whenever their data powers model training or inference.

Portable licensing metadata: usage terms travel with the data, continuously enforced by smart-contract logic.

DAO-driven governance & dispute resolution: on-chain voting and arbitration manage rule changes and conflicts.

Essential infrastructure blocks: ERC-20/721/1155 data tokens, staking & royalty contracts, DAO governance for listing rules, reputation Soul Bound Tokens that score provider quality.

Problems - Privacy Regulations

EU rules bar the commoditization of personal data, so any EU-compliant blockchain marketplace must focus on B2B exchanges that use Data-Act-ready smart contracts and accept only consent-originated, anonymized data into a "Trusted Data Marketplace." (INATBA, 2024)

Opening for LATAM:

Across Latin America, how do existing and draft data-protection and blockchain frameworks shape the legal feasibility of a consent-based, anonymised, blockchain-enabled "Trusted Data Marketplace" for AI training datasets?

Which jurisdictions provide the most supportive combination of data-sharing permissions, smart-contract recognition, and sandbox or fintech/RegTech incentives, making them viable testbeds for a pilot implementation?

Conversely, where do specific legal barriers (data-localisation mandates, restrictions on monetising personal data, weak recognition of digital consent, or stringent cross-border transfer rules) render such an experiment impractical?

8. zk-ML: merging zero-knowledge proofs with machine learning on Ethereum

Risk at stake: Hospitals, banks or labs cannot share sensitive inputs (personal financial or biometric information) or proprietary models, stalling collaborative research.

Cryptography Rationale

Zero-knowledge circuits let each institution prove that its model ran correctly on private data without exposing either the data or the weights, enabling federated learning or compliant inference.

Whether predicting disease progression across hospitals or spotting fraud across banks, the recipe is the same: each institution keeps its raw data encrypted (via homomorphic encryption), collaborates by exchanging only zero-knowledge proofs or zk-roll-ups of aggregate statistics, and runs model training/inference under zk-ML so that outsiders can verify the computation's correctness without ever seeing patient or customer data, bringing verifiable, privacy-preserving machine learning to highly regulated healthcare and finance consortia alike.

Essential infrastructure blocks: The end-to-end stack combines homomorphic-encryption libraries (e.g., Microsoft SEAL, HElib) to keep raw hospital or banking data computable yet opaque; encrypted-ML and federated-learning toolkits (SEAL-based neural nets, PySyft) plus lightweight ONNX models for local or edge inference; zk-SNARK frameworks and zk-rollups that package each party's encrypted gradients or predictions into succinct, verifiable proofs; an Ethereum smart-contract layer that immutably records and checks those proofs; and governance logic (tokens, on-chain reputation, DAO voting) to automate access control, revenue-sharing, and rule updates, thereby letting regulated institutions collaborate on ML while revealing nothing but cryptographic guarantees of correctness.

9. Extending neural networks to zk-proofs

Risk at stake: Cross-institutional AI collaborations cannot prove that a multi-layer neural network really produced a given output without disclosing either sensitive inputs (patient scans, payment trails, R&D data) *or* proprietary weights; the resulting trust gap stalls regulated deployments, audits, and commercial sharing of advanced models.

Cryptography Rationale: zk-friendly arithmetic circuits plus succinct proof systems (e.g., Halo 2, Plonky 2, Nova) let any party attach a zero-knowledge proof to each forward or training step, so verifiers (including an on-chain contract) can confirm "these committed weights processed those hidden inputs to yield this output" while

learning nothing else, unlocking verifiable, privacy-preserving federated learning and compliant AI inference. Each inference outputs a few-kilobyte proof that anyone can verify in seconds for a tiny gas fee.

Essential infrastructure blocks: Proof-capable NN circuits, proof-aggregation gadgets, layer-2 execution environments (e.g., zk-rollups) that batch many proofs before settling on L1.

10. Proof of Non-Personhood

Risk at stake: As autonomous AI agents start filing taxes, transacting online, and talking to other bots, platforms will either block them outright (to stop bot abuse) or risk being flooded by rogue software; without a tamper-proof way to tell a *legitimate, user-delegated* agent from a malicious one, both security and user convenience collapse.

Cryptography Rationale: Decentralised identifiers (DIDs) plus verifiable credentials issued on a public ledger let a human wallet grant (and later revoke) a "proof-of-non-personhood" to its AI agent: a cryptographic attestation that the holder is *not* a person but is authorised to act for a real account. Any website or peer agent can then trust, or refuse, the credential with a single on-chain check, preserving bot-resistance while still enabling benign AI automation.

Essential infrastructure blocks: Decentralised ID networks, biometric-backed VCs, soul-bound tokens for Sybil resistance, on-chain revocation registries and light-client verifiers embedded in apps.

INATBA's Limitations

This list of ten options is notable for the diversity and novelty of some of its proposed convergences. In particular, concepts such as "Proof of Non-Personhood" and "Right to be Forgotten" at the intersection of AI risks and cryptography have not been discussed in other academic papers, reports, or articles on the subject we have reviewed so far, underscoring its unique contribution to the discussion and their potential as avenues for future research.

Several limitations, however, circumscribe this scanning. First, the report's framing of use cases is not aligned to any established AI-risk taxonomy, which hinders comparison across threat domains and intents.

Second, all discussions are, understandably, presented at an EU-centric level. The INATBA report's strong institutional backing is a significant factor in understanding this perspective. One of the organization's main contributors is the Pact for Skills, a reskilling and upskilling initiative spearheaded by the European Commission. Furthermore, the AI & Blockchain Convergences Task Force, which produced this specific report, lists the European Commission as a key partner. Consequently, regulatory, infrastructural, and socio-economic contexts, particularly those relevant to Latin America, are not addressed.

Furthermore, the accompanying project database, while intended to provide a list of concrete implementations, does not include any projects based in or operating within LATAM. The database itself also appears to be populated through open self-submission and lacks formal vetting; an initial screen suggests that only a small fraction (approximately 5 out of 31 projects) meets minimum thresholds for maturity or impact. This absence of a curated, regionally relevant project registry highlights a critical void in the empirical literature, reinforcing the need for our research to move beyond conceptual claims and to develop a quality-controlled registry of demonstrably instructive case studies specifically for Latin America.

Consequently, while the INATBA study furnishes a valuable empirical foundation, it is limited by its generalized, EU-centric approach. To bridge this gap, additional work is required to: re-classify its risk-solution pairs within a harmonized taxonomy; introduce regional lenses that incorporate the pertinent technical, regulatory, and contextual nuances of Latin America; and provide concrete LATAM examples of cryptographic projects already developing tools with potential applications for AI risks. This enhanced approach enables us to move beyond a purely theoretical discussion toward an empirically grounded and regionally specific analysis.

Case Studies: Decentralized Identity for AI-Privacy in LATAM

Of the **nine LATAM-salient AI risk subdomains** identified above, this report examines one: **Privacy**, and within it the subdomain **"Compromise of privacy by obtaining, leaking, or correctly inferring sensitive information."** That is a deliberate narrowing, and not the ranking the evidence would suggest. **Harms to labour** and **technological sovereignty** both score higher in the UN Pulse Check's region-disaggregated signals. Privacy is where we can work empirically today, because it is the only subdomain with **concrete, government-level cryptographic deployments** in the region to analyze. The remaining eight are the agenda this research leaves open.

As defined in the MIT **AI Risk Repository's** Domain Taxonomy, Privacy 2.1 concerns *"AI systems that memorize and leak sensitive personal data or infer private information about individuals without their consent,"* where unexpected or unauthorized sharing compromises user expectations of privacy and can facilitate identity theft or loss of confidential IP.

What cryptography can do for Privacy

Drawing on INATBA's technical mapping of AI-blockchain convergences, the following clusters are most relevant to **Privacy 2.1** in our scope:

Data Privacy & Individual Ownership: Register data assets to rightful owners; wrap access in cryptographic permissions with on-chain auditability.

Secure Data Sharing & Granular Access Control: Tokenized rights, user-centric sharing, and revocation with forensics for misuse.

zk-ML (zero-knowledge machine learning): Prove correct training/inference on private data or weights without revealing them.

These potential applications rely on the use of zero-knowledge proofs and decentralized identity (DIDs/VCs) as primary instruments:

Decentralized Identities (DIDs) and Verifiable Credentials (VCs) enable consented, **selective disclosure** of attributes (prove facts, not identities), portable purpose limits, revocation, and auditable access, *without* centralizing PII. These rails align with government deployments and multilateral pilots in the region.

Zero Knowledge proofs (ZKs) make privacy **verifiable**: verifiers can check that a computation used the permitted inputs and logic, or that a claimant meets a criterion, **without** learning the underlying data, critical for training data access, eligibility checks, and rights-respecting inference.

None of this is exotic. Governments and multilateral bodies outside the region have been running these rails for years, in three different configurations. The United Nations Joint Staff Pension Fund began rolling out its Digital Certificate of Entitlement in mid-2020 and put the mobile app in beneficiaries' hands in January 2021, replacing paper proof-of-life letters for roughly 70,000 people across 190 countries[24]. It cut paperwork by 40%, reduced archiving costs by 95%, and reached near-total digital retention among those who enrolled, with offline kiosks covering low-connectivity regions[25].

In March 2023, against a documented record of breaches affecting Aadhaar holders[26], Anon Aadhaar showed that zero-knowledge proofs can be layered onto an existing national ID system without rebuilding it: any holder can generate a proof from a document the government already signed. The code has since been forked for voting, ticketing and age verification.

And in October 2023 Bhutan launched a national digital identity on W3C DIDs and verifiable credentials with zero-knowledge selective disclosure, announcing its migration to public Ethereum two years later[27]. Reporting on that migration put the annual cost of a conventional identity program at five to ten dollars per user in low-income countries, citing the World Bank, against a projection of under one dollar for Bhutan's model depending on transaction fees and validator costs[28].

Nor is it new in Latin America. Proyecto DIDI began issuing verifiable credentials to smallholder farmers in Argentina's Gran Chaco in 2021, through the ai·di wallet[29], so that rural producers could prove sustainable practices and convert them into climate risk scores a lender would accept. Self-sovereign identity applied to financial inclusion, aligned with the IDB's LACChain; the pilot published its results in 2022 and no continuation is documented. What came next was larger and better resourced. QuarkID issues citywide DIDs and verifiable credentials for official documents through Buenos Aires' miBA app; Sovra runs government programmes in Mexico and Argentina, providing governments with a digital trust stack; Blerify is deploying ISO 18013-5 credentials with institutions in Panama and El Salvador; and World operates proof-of-human enrollment across the region. Those five are the case studies of this report.

These deployments move privacy from policy to practice by delivering concrete, checkable properties: captured consent, selective disclosure with purpose limitations, credential status and revocation, and audit trails, all of which reduce the need to centralize personally identifiable information while preserving service delivery.

The international precedents were built for a human adversary, before generative AI made the problem harder. The Latin American cases have been adapting to it as it grew, and the sections that follow examine how far that adaptation has actually gone.

The next section presents:

Five cases of decentralized identity and privacy-preserving credentials deployed or piloted by governments and institutions in LATAM, with their scope, their evidence base, and what stopped each one short.

What the five have in common and where they diverge: on standards, on adoption sequence, and on what each one is actually proving.

The emerging layer for verifying agents rather than people, the four approaches being built for it, and why the region is absent from all of them.

The five cases that follow move from the most widely shared infrastructure to the most divergent. The first three run on the same W3C rails, the fourth was built on a competing international standard, and the fifth proves a different property altogether. DIDI opens the sequence because it is the only one of the five whose pilot has concluded, and the only one to have published what it learned.

Proyecto DIDI

Launched in 2021, the project was an effort to use cryptographic tools to extend financial access to vulnerable populations in Argentina.[30] DIDI is one of the earliest digital identity projects at a government level, focusing on leveraging Decentralized Identifiers (DIDs) and Verifiable Credentials (VCs) to address socioeconomic and financial barriers. Its documentation describes the self-sovereign architecture, with holders controlling when and with whom a credential is shared, but no selective disclosure or zero-knowledge layer: when a credential was presented, it was presented whole. That distinction, between who controls a credential and how much of it a verifier sees, separates DIDI from the cases that follow.

For example, in the Gran Chaco region, farmers were given VCs in 2021 that count toward "climate risk scores." These scores could then be presented to financial institutions, who can use this verified, non-traditional data to extend financial credit to rural communities that might otherwise be excluded[31].

This approach highlights how DIDs and VCs enable verifiable, consent-based data sharing and also support more inclusive, and potentially less biased, AI-driven credit scoring[32].

A significant milestone for the DIDI project is its partnership with the **Inter-American Development Bank (IDB)** and **LACCHAIN**[33]. This collaboration has resulted in the publication of a joint report that shares learnings and identifies opportunities. The joint report and broader literature on digital identity in the region shed light on several key challenges that directly inform our research gaps.

A significant challenge is the pronounced need for **digital literacy and pedagogical support**. The report emphasizes that a "one-size-fits-all" approach to user education is insufficient, noting that many end-users, validators, and even issuers possess basic or non-existent digital skills. This necessitates a substantial and often underestimated investment in practical, hands-on training tailored to diverse user demographics.

Another critical challenge is the balance between the technical ideals of Self-Sovereign Identity (SSI) and the practical needs of the end-user. The project found that some core tenets of SSI might be counterproductive to adoption given the socioeconomic realities of the target population, requiring a careful equilibrium between technological purity and user experience.

Finally, the report highlights the often-overlooked issue of **sustainability and resource management**. While the technology may be open-source, the project underscored that the ongoing costs of maintenance, support, and necessary hardware for an SSI model are a crucial and often unbudgeted factor for partner organizations. These challenges (related to digital literacy, user experience, and resource sustainability) provide a more granular view of the implementation hurdles that are not typically covered in high-level reports, reinforcing the need for our research to be empirically grounded and context-aware.

The experiences from the DIDI project and its partnership with multilateral institutions, such as the Inter-American Development Bank (IDB), are highly instructive for our research. The findings reinforce our focus on **regional nuances** and the need for empirically grounded analysis that acknowledges the significant role of **digital literacy, user experience, and resource sustainability**. The case demonstrates that the challenges of implementing cryptographic solutions are not merely technical but are deeply intertwined with regulatory, institutional, and socioeconomic factors.

This partnership with a multilateral institution like the IDB is crucial. It sets a precedent for the high-level recognition of decentralized identity solutions, a critical and often missing step for the institutional adoption and regulatory acceptance of cryptographic tools. This endorsement provides a crucial proof point for the wider ecosystem of cryptographic solutions explored in this report, moving the conversation from theoretical possibility to practical, government-backed implementation.

In the classification tables that follow, "Intent" follows the MIT AI Risk Repository definition (see "AI Safety vs AI Security" in the Theoretical Framework): intentional where the harm is the expected outcome of an adversary's goal, unintentional where it is a systemic or accidental by-product. A dash marks cases with no AI component to classify.

Appendix: Proyecto DIDI Classification

Dimension	Detail
What it does	Self-sovereign identity for financial inclusion: issues verifiable credentials through the ai·di wallet so that rural producers and low-income workers can prove attributes such as sustainable practices to lenders and agencies, with control over what they share.
Relevance for AI privacy-risk mitigation	No component aimed at AI. Relevant as the region's earliest DID/VC pilot with government and multilateral backing: it showed that verifiable credentials could reach populations outside the formal system, and documented the barriers of digital literacy and sustainability that the later cases met again.
Status	Pilot concluded (results published October 2022; lessons report with IDB Lab published 2023; no continuation documented).
Deployments and counterparts	Gran Chaco (Santiago del Estero, Jujuy) and Misiones, Argentina. Counterparts: IDB Lab and LACChain, with ONG Bitcoin Argentina; no government issuer.
AI Risk Domain (MIT)	-
AI Risk Sub Domain (MIT)	-
Intent (MIT)	-

Dimension	Detail
INATBA categorization	Data Privacy & Individual Ownership; Secure Data Sharing & Granular Access Control
Credential format and proofs	W3C Verifiable Credentials with status and revocation patterns; no zero-knowledge proofs documented.
Anchoring / settlement	LACChain (Hyperledger Besu, permissioned public network), aligned with the LACChain ID Framework for on-chain key directories and trusted lists. Does not settle to Ethereum.
Standards implemented	W3C DID Core; W3C Verifiable Credentials Data Model.

DIDI ran into problems of literacy, cost and adoption at the scale of a few thousand rural producers. The next deployment faced them at the scale of a city.

QuarkID Protocol

Launched in October 2024[34], Quark ID is a decentralized digital identity **protocol**, not an app or a single government project. It emerged as a neutral standard co-created by a consortium of actors in the region (including OS City, Xcapit, Extrimian and the government of Buenos Aires) who were all working on similar problems around verifiable credentials and digital identity.

Instead of each actor building its own closed stack, this group agreed on a shared way to issue and verify credentials using decentralized identifiers (DIDs) and verifiable credentials, drawing on international efforts like W3C standards, Trust over IP, and reference models from Europe, the US, UK, Canada, New Zealand, Singapore and Estonia. The goal was to have a **common protocol that anyone could implement**, independent of wallet, chain or vendor, so that digital identity in Latin America would avoid "vendor lock-in" and converge on an open, interoperable standard.

As a result, Quark ID has functioned as a **regional reference protocol**: a common language that different wallets, infrastructures and applications can speak, and a basis for subsequent iterations such as the ongoing "Quark ID 2.0" effort led by the company Sovra to align more tightly with globally dominant standards.

The best-known implementation of this protocol so far is the City of Buenos Aires, which adopted Quark ID as the protocol behind its digital identity system and integrated it into the miBA city app. Using this stack, the app provides Buenos Aires' 3.6 million citizens with access to over 60 digital documents, including birth certificates and citizen credentials, which can be presented to government portals and services[35] with zero-knowledge selective disclosure, so that a verifier sees only the attribute requested.

The protocol's usage can be read on-chain. A public query maintained by Matter Labs counts the DIDs created through the QuarkID contract on zkSync Era: 1,817,638 by 22 July 2026[36]. The series shows when adoption happened. Fewer than 20,000 DIDs existed before the miBA launch; October 2024 alone added 110,000, the count passed 600,000 six months after launch and one million in November 2025, and the busiest day on record, 10 December 2025, added 25,459. Creation has not tapered since: 2025 closed at 902,000 new DIDs and the first seven months of 2026 added 628,000, a monthly pace above the 2025 average, with the largest daily peaks spread across 2025 and 2026 rather than clustered at launch. Two caveats apply. The count is of DIDs, not of unique or active users, and it covers every issuer anchoring on the contract, so it reads as protocol usage rather than as miBA adoption alone.

This is a significant departure from centralized digital identity systems like India's Aadhaar or Singapore's Singpass, where governments maintain control and traceability, and are at risk of cyber threats due to its centralization. In fact, Aadhaar's data of 815 million users was hacked[37], (more on this later in the brief).

Similarly, LATAM has seen multiple Aadhaar-like large-scale leaks of national-ID data (or near-equivalents):

Argentina: In 2021, a hacker breached the government's IT network and stole ID card details for the country's entire population, data that was being sold in private circles[38].

Ecuador (2019): Misconfigured Elasticsearch server leaked data for ~20M Ecuadorians (PII + financial fields), effectively the entire country[39].

Brazil (2021): Personal records for 220M+ people (more than the population) were traded online; widely described as Brazil's largest leak, with ongoing legal fallout. (Not the federal registry itself, but it included national IDs, CPF, from private data brokers/credit bureaus.)[40]

Protocol Design

The Quark ID protocol was originally designed by Extrimian, Xcapit, OS City, and other regional actors working alongside the government of Buenos Aires, who were all grappling with similar questions around verifiable credentials and digital identity. Rather than each team building an isolated stack, they converged on a shared protocol that any of them could implement and extend.

At its core, Quark ID is a regional interoperability standard for digital identity and verifiable credentials in Latin America. **It does not prescribe one app, one chain, or one vendor. Instead, it defines a common way to represent and exchange identity information using decentralized identifiers (DIDs) and verifiable credentials.**

By design, every person or organization is represented by a DID, which becomes the stable "hook" that survives across systems and over time. Different institutions (governments, universities, banks, transport operators) can attach credentials to that DID without sharing raw databases. Documents that previously existed only as PDFs or siloed entries in backend systems are modeled as cryptographically signed verifiable credentials: they are issued by trusted authorities, held by users in their own wallets, and presented and verified through standard proofs. In the Buenos Aires deployment, this includes IDs, birth certificates, licenses, and other city documents that citizens can present and prove without exposing their entire data trail to the issuer.

The protocol is tightly coupled to the way its creators organized. It emerged from a public-private consortium in which multiple startups working on identity and GovTech sat at the same table as the City of Buenos Aires, which contributed both a large user base and an unusually strong in-house engineering team. Governments play the role of "anchor tenants" in this trust fabric: they are the first major issuers and verifiers under the protocol, and by declaring that a DID-backed, wallet-held credential is valid for accessing services, they seed the ecosystem with high-value attestations that others can safely rely on.

Importantly, Quark ID standardizes not just data formats but also the basic expectations about how issuers, wallets, and verifiers should behave. The protocol functions more like a language than a platform: any sufficiently compliant wallet, verifier, or backend can "speak" Quark ID, regardless of the underlying infrastructure.

From the beginning, its design drew heavily on international efforts such as W3C's DID and Verifiable Credentials specifications[41], the layered architecture promoted by Trust over IP, and reference models emerging from Europe (including the EIDAS / EIDAS 2.0 discussions), the UK, the US, Canada, New Zealand, Singapore, and Estonia. **The intention was to ensure that a credential issued under Quark ID could, in principle, be mapped onto the way other major regions are formalizing digital identity, without copying any one system wholesale.**

Trade offs

Because it was a first mover, the consortium had to make concrete implementation decisions before global standards had fully stabilized.These choices were driven by the practical need to deploy a functioning system in Buenos Aires, rather than by an attempt to produce a theoretically optimal or aesthetically "perfect" specification. This produced a working system and a valuable case study, but it also created technical debt: some parts of Quark ID 1.0 now need to be updated to align more cleanly with where W3C and large regulatory blocs have converged.

The decision to frame Quark ID as a neutral, shared protocol rather than a proprietary product also brought tradeoffs. That neutrality reduces vendor lock-in and makes it politically easier for new actors to join, yet it slows down

evolution: deprecating old patterns, agreeing on a "2.0" spec, or rolling out breaking changes requires coordination across several independent organizations instead of unilateral control.

Chuy Cepeda, co-founder and Chief Strategy Officer of Sovra, who participated in the initial design and deployment of Quark ID, stated in an interview for this research that there is an ongoing "Quark ID 2.0" effort aimed at updating the protocol without discarding its core gains. On the technical side, this effort seeks to bring the protocol into closer alignment with EIDAS 2.0 and other emerging wallet and credential guidance, national trust frameworks such as those in the UK, and best practices from bodies like US Homeland Security, so that Latin-American credentials do not become a regional dead end. It also aims to make protocol agnosticism real rather than aspirational, tightening the separation between the specification (DIDs, credential formats, flows, governance assumptions) and particular implementations (chains, wallets, verifiers, infrastructure providers). On the institutional side, Quark ID 2.0 implies clearer governance and long-term stewardship: moving beyond an ad-hoc consortium tied to specific projects toward more explicit answers about who versions the specification, how breaking changes are proposed and rolled out, and how reference implementations are maintained.

miBA: QuarkID's first pilot

Buenos Aires has emerged as an ideal proving ground for decentralized identity (DID) projects and digital public infrastructure initiatives like miBA. The convergence of several factors, ranging from institutional capacity to regulatory openness and existing technological infrastructure, positions the city as a strategic pilot territory for cryptographic solutions in Latin America.

According to insights from interviews with World and Sovra representatives, Buenos Aires possesses a critical mass of technical expertise within government. As noted in conversations with Chuy Cepeda from Sovra, Argentina has 300 developers in government. This internal technical capacity is exceptional in the Latin American context and enables the city to implement, maintain, and iterate on complex cryptographic infrastructure without complete dependence on external vendors.

The presence of Diego Fernández, former Secretary of Innovation for Buenos Aires and current co-founder of Sovra, exemplifies this capacity. Fernández served as an internal "evangelizer" for blockchain and decentralized technologies within government, creating institutional knowledge and building acceptance for these approaches. As the World team noted in interviews, "Ciudad de Buenos Aires es una excepción, porque tiene a Diego Fernández que fue un evangelizador" (The City of Buenos Aires is an exception, because it had Diego Fernández who was an evangelizer).

Building on this foundation, Sovra, founded by members of the team that co-designed QuarkID at OS City, has adopted the protocol as the backbone for deployments beyond Buenos Aires, in the province of Salta and in the Mexican state of Nuevo León. This diffusion has helped consolidate what is, in practice, the most advanced real-world adoption of decentralized digital identity identified in Latin America to date: a government-linked DID/VC stack with 1.8 million DIDs created on-chain, and the only case we have found with sustained traction outside a narrow, specialized niche.

Key Takeaways

Quark ID illustrates how cryptographic primitives (DIDs, verifiable credentials, signatures, hashing) become politically and socially meaningful only when embedded in a realistic governance and adoption strategy.

At the same time, this path generated technical and institutional debt that now needs to be paid down through Quark ID 2.0: tighter alignment with global standards, stronger protocol agnosticism, clearer governance, and better UX for both citizens and officials.

As other regions design their own digital identity stacks, Quark ID offers a useful case study in how much of cryptography's value depends not only on math and code, but on institutional incentives, regulatory constraints, and the messy process of convincing humans to trust proofs they cannot see.

QuarkID is a base-layer protocol (a standard) for decentralized digital identity. Its core innovation is a *shared language* for issuing and verifying credentials (DIDs/VCs) that different actors can implement without being tied to a specific vendor, wallet, or chain, which reduces **vendor lock-in** and enables interoperability across the region.

Standardization is a prerequisite for scale in LATAM. In a fragmented ecosystem, the sequence matters: **build the shared rails first** (protocol + trust model + verification rules), then allow wallets and applications to compete and innovate on top.

Multi-actor coalition-building is the model. The protocol emerged from a **coalition** (public sector + infra startups + implementers), and scaling depends on **partners in DPI ecosystems**, local governments, and distribution-capable private actors. Practically: without convening mechanisms and plural governance, DID efforts tend to remain niche.

Rule of thumb for any DID project in LATAM aiming at adoption: align with (or ensure compatibility with) **QuarkID / its evolution** as the shared rail, or clearly justify the tradeoff if breaking compatibility (coordination costs, regulatory constraints, etc.).

"Best-in-class" cryptography isn't enough in LATAM. Even if there are technically more sophisticated approaches (biometrics, ZK passport-style solutions, etc., as we'll see ahead), success is constrained by **institutional fit, operating costs, UX, distribution, and coalition incentives**.

"Inverted adoption" appears as the scaling play: governments first, then enterprises and citizens. Credential ecosystems face a chicken-and-egg problem (holders vs. verifiers). The QuarkID/Sovra pathway suggests a strategy to break that loop by **anchoring issuance and early verification in government services**, then expanding to private-sector verifiers and additional jurisdictions.

Fast-track QuarkID 2.0 alignment to avoid falling behind. Prioritize a time-bound push to converge with globally dominant wallet and credential standards (and publish a clear roadmap with milestones and deprecation dates). Without accelerating this alignment, the region risks being left with a locally functional but internationally non-portable identity stack, harder to interoperate with external wallets/verifiers, costlier to maintain over time, and less attractive for new governments and private-sector adopters as global patterns consolidate.

As of August 2026 that roadmap does not exist. QuarkID 2.0 was described to us in interview as an ongoing effort, an intention rather than a plan with dates attached, and it remains one: there is no specification, no release, no governance document and no public announcement[42]. The consortium that built the protocol has meanwhile come apart: Extrimian, its original author, opened a repository in October 2025 described as evolving independently from QuarkID and moved its implementation to IOTA Identity in January 2026, while Sovra migrated the QuarkID node stack to its own organisation and maintains it from there[43][44]. The canonical protocol organisation has not been updated since October 2024. This makes the recommendation harder to act on than it was when written, and also more necessary: the region's most deployed identity protocol currently has no one versioning it.

Research Gaps and Questions: Despite the project's high-profile launch and its recognition as a Digital Public Good[45], a notable research gap exists in independent, "on-the-ground" analyses of its practical operation and impact on citizen privacy.

Furthermore, the project raises several key questions relevant to our research on AI and cryptography, including:

How does the project's legislative and regulatory context in Argentina make it a viable model, and what factors would be critical for its replication in other LATAM countries?

How would such a system manage the balance between individual data sovereignty and the potential for new AI initiatives, such as a government-run LLMs? For instance, could this framework be extended to allow users to "sell" their data for the development of such models?

Appendix: QuarkID Classification

Dimension	Detail
What it does	Government-operated decentralized digital identity protocol integrated into the miBA city app, so that residents hold and present verifiable credentials with selective disclosure.
Relevance for AI privacy-risk mitigation	No AI component in its original version; credentials are issued for documents, benefits and identity. Relevant as the reference for an interoperable standard across governments at regional scale, the shared rail on which an AI layer would be built.
Status	Live (public launch in miBA, October 2024; 60+ document types available; 1.8 million DIDs created on-chain by July 2026). The consortium that built the protocol has since dissolved; no QuarkID 2.0 specification as of August 2026.
Deployments and counterparts	City of Buenos Aires, Argentina. Counterpart: the Government of the City of Buenos Aires, which operates the system as public issuer. Sovra reuses the protocol in Nuevo León and Salta (see Sovra).
AI Risk Domain (MIT)	-
AI Risk Sub Domain (MIT)	-
Intent (MIT)	-
INATBA categorization	Secure Data Sharing & Granular Access Control
Credential format and proofs	W3C DIDs and Verifiable Credentials; selective disclosure through BBS+ signatures applied in the wallet, not independently verified by this research.
Anchoring / settlement	zkSync Era (Ethereum Layer 2); Sidetree method with Merkle proofs for state. Settles to Ethereum.
Standards implemented	W3C DID Core; W3C Verifiable Credentials Data Model; Sidetree; Trust over IP alignment.

QuarkID was built by a consortium, where key leaders joined efforts in what today operates as Sovra.

Sovra

While QuarkID established the foundational protocol, Sovra represents its most concrete and scaled implementation besides miBA in Buenos Aires, demonstrating how decentralized identity infrastructure can move from theoretical standards to operational government services across multiple jurisdictions.

The company pioneered what they call the **"inverted adoption curve"**: targeting governments first, then enterprises, and finally citizens, the opposite of typical technology adoption patterns. This strategy addresses a fundamental challenge in Latin America: establishing institutional credibility before seeking grassroots adoption.

After seeing the flagship implementation of Buenos Aires' miBA project, Sovra has gone to establish different initiatives with other local governments in LATAM:

Nuevo León, Mexico: Protect citizen data as part of the "Burocracia Cero" campaign[46] [47].

Salta, Argentina: The IDDI project is empowering citizens with decentralized identity, enabling seamless access to government services and secure interactions with banks.[48]

Mariela Saldívar, Chair of the Regulatory Reform Commission in Nuevo León, reports that the state's deployment cut red tape by 30% and made public services up to 80% faster[49][50]. Those figures are self-reported by the state government.

No separate adoption figure exists for these deployments: the on-chain series described in the QuarkID case study counts DID creation at the level of the protocol contract rather than by implementer, and this research found no figure attributable to Sovra's programmes alone. Sovra's own site presents reach by jurisdiction population rather than by active users, which is a different measure. Treat those numbers as an indication of deployment scope rather than of adoption[51].

The Digital Trust Stack

Sovra provides a modular four-component suite designed specifically for public sector deployment:

SovraGov: A no-code service digitization platform that enables government officials to create and manage digital procedures through drag-and-drop workflows, eliminating the need for extensive technical expertise.

SovraID: The core identity layer for issuing and verifying decentralized identifiers (DIDs) and verifiable credentials (VCs) with zero-knowledge attestation capabilities, built on the QuarkID protocol. Sovra issues did:quarkid identifiers and, since January 2026, maintains the QuarkID node stack from its own repositories[52].

Sovra Wallet: A non-custodial citizen wallet that allows residents to hold and present their credentials with full sovereignty. The government cannot see how, where, or when credentials are used after issuance.

SovraChain: A custom Ethereum-based Validium rollup with off-chain data availability, co-developed with LambdaClass using the Ethrex stack, with SP1 and Risc0 provers integrated via Aligned Layer services.[5354] This purpose-built infrastructure enables cross-institution interoperability while maintaining privacy guarantees.

According to Chuy Cepeda, co-founder and Chief Strategy Officer of Sovra, the design of Sovra's reusable identity stack is explicitly aimed at making it easier for governments to onboard and adopt decentralized identity. The stack begins with SovraGov, a "digital front desk" that allows governments to digitize forms and procedures without needing large in-house development teams, directly addressing their most immediate demand for basic digital services.

Once this entry point is in place, the rest of the stack tackles a second, structural problem: how different government departments can interoperate without sharing or centralizing databases. By using decentralized identifiers (DIDs), agencies can rely on a common identity layer while preserving confidentiality and control over their own data.

Cepeda also frames the architecture as a response to cybersecurity and AI-era risks, shifting centralized databases from a major liability toward a minimized exposure surface by pushing credentials into user wallets and relying on cryptographic verification rather than manual checks. This reduces fears around hacks, deepfakes, and forged documents.

Since that interview Sovra has moved the AI framing from rationale to product. Its site presents the stack as serving "humans and agents", advertises agent integration across all products, and describes two additions: SovraRAG, a retrieval assistant that turns an institutional knowledge base into cited answers, and MCP servers the company says it has built so that an agent can issue, verify and manage credentials without a REST integration[55].

Overall, Cepeda situates this design inside Sovra's broader "inverse adoption" strategy: starting with credible, high-leverage governments that can impose standards and onboard large user bases quickly, and only then extending the same rails outward to universities, banks, transport operators, and other private-sector verifiers.

Propia ID: A Nascent Public Trust Layer for Verifiable Credentials

In August 2026 the team behind SovraChain published Propia ID, a whitepaper specifying a public trust layer for verifiable credentials, co-authored by Diego Fernández and Chuy Cepeda of Sovra with LambdaClass and Aligned engineers behind the rollup described above[5657]. While SovraChain is a component of one company's product, Propia ID is proposed as a protocol that any institution could run, verify and leave. The paper commits it to open implementation, open verification, institutional exit rights, transparent governance and data independence.

The paper's premise is that credential standards solve issuance and presentation and leave the layer beneath them open: a verifier can check a signature, but not whether the issuer was authorised to make the claim, which keys represent that institution today, or whether the credential is still valid. Propia ID proposes three public registries for that, covering issuers, identifiers and credential status, with personal data kept off them.

The whitepaper also gives the reasoning behind the settlement. A shared database could hold the same registries, the paper argues, but it would place availability, update ordering and the retained history under one operator, and

every participating institution would depend on it; a blockchain lets independent institutions agree on common state without delegating control to any one of them.

Ethereum is chosen on four stated grounds: no issuer, government or provider holds settlement control; any participant can verify finalised commitments from its own node; commitments are final under proof-of-stake consensus and its penalties; and Layer 2 execution can be verified from Layer 1.

A validium is chosen for scale: registry operations run off Ethereum and are batched, Ethereum verifies a validity proof before accepting each state commitment, and the data needed to reconstruct the state stays off-chain, which is where the cost saving over a ZK rollup comes from.

As production evidence, the paper cites validity-proof systems built on StarkWare's stack having processed one billion transactions and more than USD 1.3 trillion in cumulative volume by May 2025. The stated direction of travel is based sequencing, in which Ethereum's own proposers and builders order the Layer 2's transactions and the privileged sequencer disappears; the paper sets no date for it and conditions the move on preconfirmation and forced-inclusion mechanisms reaching operational maturity.

These are the paper's arguments. None has yet been tested against a Propia ID registry in production, and the paper itself notes that Ethereum does not supply every guarantee: issuer authorisation stays a governance decision, and a validium keeps data availability outside Ethereum.

One claim matters for a question this report takes up in the Blerify case study, where the region's split between W3C verifiable credentials and ISO 18013-5 mdocs is examined: Propia ID lists both formats, alongside SD-JWT VC and OpenID4VCI/VP, under the same trust layer, so the choice of format would no longer require duplicating the trust infrastructure beneath it.

The Inverted Adoption Curve: Government-First Strategy as a Path to Market Validation

Chuy Cepeda described Sovra's go-to-market logic as an "inverted adoption curve." Rather than following the more familiar technology diffusion pattern in which tools moved from niche technical environments toward consumers and only later into government, Cepeda said Sovra had deliberately reversed the sequence. In his account, the intended order was government first, then large enterprises, then universities, and only after that broader citizen adoption.

Cepeda framed this inversion as a practical response to how credential-based ecosystems (DIDs and verifiable credentials) actually got stuck: bottom-up pilots frequently stalled because neither side of the market (issuers and holders on one side, verifiers on the other) had a strong incentive to invest before the other side existed at meaningful scale.

Cepeda said Sovra's decision to lead with government adoption had been intended to break that deadlock by creating two conditions simultaneously. First, a government issuer could rapidly produce a large base of legitimate credential holders (citizens or residents). Second, public services and administrative workflows could generate immediate, high-legitimacy verification demand. Cepeda presented this as more than an implementation shortcut: he treated government deployment as a form of market validation, because operating under public scrutiny, legal constraints, and institutional complexity signaled that the system could work beyond small pilots. In his telling, that validation would then reduce perceived risk for subsequent adopters and make it easier for Sovra to recruit additional institutions into the ecosystem.

Subsequently, Sovra's approach has not aimed to run pilots "anywhere," but had prioritized institutions whose participation could function as a transferable signal of legitimacy: relevant governments, relevant universities, and major companies. Cepeda argued that this credibility filter explained why Sovra had focused its efforts on high-profile public-sector deployments and large private partners rather than relying primarily on small grassroots networks.

He also said the government-first approach had created opportunities for regulatory work that would have been difficult to pursue from a purely bottom-up starting point. Cepeda cited Nuevo León as an example where Sovra had worked on "regulatory updates" intended to allow decentralized alternatives to mandatory centralized records, and he framed that as a downstream consequence of having already demonstrated a government-scale implementation.

This inverted adoption curve can be read in contrast to a recurring vulnerability described elsewhere in our research interviews. A former collaborator of a prominent grassroots AI x cryptography initiatives with a LATAM chapter described how they struggled to sustain themselves and follow on with pilots project over time due to internal governance and coordination failures, even when initial resources or momentum existed. Against that backdrop, a government-first pathway can appear more structurally resilient because it anchors deployments in institutions with longer time horizons, clearer mandates, and more durable operational capacity than many bottom-up organizations can reliably maintain.

At the same time, other interviewees pointed out the constraints of this approach: because the strategy begins with governments, it will be limited by the small number of administrations that are open to institutional change at any given time, which creates a bottleneck and competition among projects for attention. This is exactly why Sovra's government-first pathway required patience and coalition-building, building positive-sum alliances with stakeholders and local communities, rather than relying only on technical arguments.

Two further shifts are visible in Sovra's technical documentation and not yet in its public narrative. Its API now issues and verifies in ISO 18013-5 mDoc alongside W3C verifiable credentials, through OpenID4VCI and OpenID4VP[58], so Sovra's product now spans both standards even though its public narrative does not. The evidence is API documentation rather than a deployment: no government is identified as issuing mDocs through Sovra, and the company's own educational material makes no reference to ISO 18013-5 at all[59]. And Sovra addresses post-quantum cryptography nowhere on its site or in its documentation[60], which remains the clearest technical difference with Blerify.

Key Takeaways

"Plug-and-play" adoption for governments (no-code first). Sovra's implementation strategy reduces the dependency on large in-house engineering teams, an important constraint across many Latin American public administrations where Buenos Aires is the exception rather than the rule. Leading with a no-code "digital front desk" lowers procurement and delivery friction and makes decentralized identity feasible as an operational public-service upgrade, not a bespoke R&D project.

An inverted adoption curve that prioritizes market validation through institutions. The government-first sequence (followed by enterprises and other high-credibility institutions) addresses, in Sovra's account, the issuer-verifier coordination problem that stalls many DID/VC ecosystems. By starting where legitimacy and scale can be created quickly, the model turns decentralized identity from a pilot narrative into a validated infrastructure layer that other actors can rationally adopt.

Government-first provides leverage, but creates a bottleneck. Starting with governments can unlock regulatory updates and accelerate early scale, yet it is constrained by the limited number of administrations willing to pursue institutional change at any given time. That scarcity concentrates competition among projects and increases the importance of sustained relationship-building, political timing, and positive-sum local partnerships.

Coalition-building is a core implementation requirement. The model depends on aligning multiple institutional actors (government units, regulators, and downstream verifiers) so adoption is driven as much by governance and stakeholder coordination as by technical design. This makes the approach slower than purely technical rollouts, but potentially more durable once embedded in public workflows.

Appendix: Sovra Classification

Dimension	Detail
What it does	Digital-trust infrastructure for the public sector: issues, holds and verifies decentralized credentials with selective disclosure, digitizes procedures through no-code workflows, and anchors attestations on SovraChain.
Relevance for AI privacy-risk mitigation	Not designed for AI risks originally: the credential stack was built to digitize identity procedures. The company now declares AI components, like the SovraRAG, a retrieval assistant over institutional knowledge bases, and MCP servers for agents to issue, verify and manage credentials, neither examined in this research. Relevant as the live government issuers on which an identity layer for AI agents would run.
Status	Live in two jurisdictions (Nuevo León, Mexico; Salta, Argentina). No adoption figures published by the company; DID creation is counted at protocol level (see QuarkID).
Deployments and counterparts	Nuevo León, Mexico (state government, "Burocracia Cero" programme); Salta, Argentina (provincial government, IDDI programme).
AI Risk Domain (MIT)	Privacy & Security
AI Risk Sub Domain (MIT)	2.1 Compromise of privacy by obtaining, leaking or correctly inferring sensitive information; 2.2 AI system security vulnerabilities and attacks (data privacy and integrity of credentials in AI-enabled public services)
Intent (MIT)	Intentional
INATBA categorization	Secure Data Sharing & Granular Access Control; Data Privacy & Individual Ownership; Provenance & Audit Trails for Models and Data; Consent Management Across the AI Lifecycle
Credential format and proofs	W3C DIDs and Verifiable Credentials with zero-knowledge attestations handled by SovraID; the API also issues and verifies ISO 18013-5 mdocs, documented but with no government identified issuing them.
Anchoring / settlement	SovraChain, a custom Validium rollup with off-chain data availability, co-developed with LambdaClass on the Ethrex stack, with SP1 and Risc0 provers via Aligned Layer. Settles to Ethereum.
Standards implemented	W3C DID Core; W3C Verifiable Credentials Data Model; OpenID4VCI and OpenID4VP; ISO 18013-5 mdoc.

Blerify issues in both of the region's credential formats, W3C verifiable credentials and ISO 18013-5 mdocs, and was early to the second one while the rest of the region had settled on the first. Sovra has since added the same standard to its API, which makes that starting position look less like an outlier and more like an early move.

Blerify

Background

Blerify was founded in 2023 in Panama City by Marcos Allende López, who served as CTO of LACChain at IDB Lab, the Inter-American Development Bank's blockchain infrastructure initiative, before building the company.

Where QuarkID operates as a neutral open-source protocol, built by a consortium including OS City, Xcapit, and Extrimian, and where Sovra provides software infrastructure to government issuers through a broader ecosystem strategy, Blerify is a commercial product company: a post-quantum-secure digital wallet that combines self-sovereign identity credentials, digital asset management, and benefit delivery into a single mobile interface, accompanied by APIs and SDKs for institutional issuers and verifiers. Allende's long-run vision is that every person will carry, on a single mobile device, their national ID, academic record, health history, and financial credentials, a vertical integration of identity data that goes well beyond the current standard of issuing digital driving licenses or employment certificates.

Its public record runs across three countries. In Uruguay it built a decentralised trust list with AGESIC, the country's e-government agency, in a pilot the two organisations published jointly and which won the CDPI Global Government Hackathon at the Global DPI Summit in November 2025[61]. In Panama it built the digital driving licence channel with the licensing concessionaire, and in El Salvador a national identity credential, both structured

around ISO 18013-5, the mobile driving licence standard whose significance for the argument that follows is that it carries a fixed international namespace the W3C stack deliberately leaves open

Underneath both tracks sits a decentralised trust registry, deployed as smart contracts on LNET, a Hyperledger Besu network, and on Avalanche. Its verifiable credentials, the diplomas, professional licences and institutional records it issues outside government identity documents, resolve through did:lac1, a DID method the company proposed as an extension of two earlier ones: did:ethr, built for Ethereum and EVM networks, and did:lac, developed for LACChain, the IDB initiative Allende came from. Its distinguishing feature is backwards revocation. A controller can revoke a key from a specified moment in the past, invalidating what was signed after that point while leaving earlier statements valid, provided those carry a blockchain-anchored proof of time[62].

Blerify opens its government conversations with AI-driven fraud, and the AGESIC document shows the argument in a government's own voice: remote biometrics, one-time codes and CAPTCHAs were designed for a different threat environment, and face significant risks of spoofing and irreversible compromise given the rapid evolution of advanced AI models[63].

Two Credential Formats, and the Rule for Choosing Between Them

The dominant technical framework in the LATAM identity ecosystem is the W3C stack: Decentralized Identifiers and Verifiable Credentials, the foundation on which QuarkID and Sovra are built. QuarkID was designed as a neutral, open standard, a shared protocol that different wallets, infrastructures and applications can implement without committing to a specific vendor or chain. That approach addressed the coordination problem among the Buenos Aires consortium, but interoperability under it remains contingent on bilateral adoption. A government in El Salvador verifying a credential issued by Buenos Aires would require both parties to have aligned on the same QuarkID implementation, field definitions and trust registry.

ISO 18013-5 answers that problem differently, and the difference runs deeper than formatting: the two are separate technical stacks with separate trust models, not variants of one another. One regional specification, Propia ID, published in August 2026 by the team behind Sovra's settlement layer, proposes to place both under a single public trust layer; it is described in the Sovra case study and had no deployment at the close of this research. Where W3C deliberately leaves field names and semantics to whoever defines the schema, ISO fixes the complete namespace, fields, permitted values and signing algorithms for the mobile driving licence, with no room for implementer interpretation; other document types reuse the format, with the namespace defined by whoever issues them. The consequence, for the driving licence, is direct cross-system interoperability without bilateral coordination. A driving licence issued under ISO 18013-5 works natively in Apple Wallet, Google Wallet, the digital driving licence wallets of more than 25 US states, and European wallets under eIDAS 2.0.

That rigidity is the feature rather than the cost, and it explains the decision rule Blerify applies. In its own words, a government identity document has to be recognised by systems that have never heard of the issuing organisation, which is what a fixed shared namespace delivers. Everything else, a diploma, a professional licence, a KYC result, a membership, has no international standard dictating its fields, so the format whose schema the issuer designs is the one that gets out of the way. The choice is therefore made per credential type according to whether a standard already exists for it, not per organisation and not by sector.

Blerify issues in both, and says so in its developer documentation: W3C verifiable credentials and ISO 18013 mdocs, both signed with the issuer's key, anchored in its decentralised trust registry and verifiable offline. Sovra's API now does the same. What makes that practical is the protocol layer, where the two formats are not in competition at all: OpenID4VCI and OpenID4VP are explicitly format-agnostic, carrying W3C verifiable credentials, ISO mdocs and SD-JWT VCs under distinct format identifiers, and permitting credentials of multiple formats in a single transaction. This is why the Panama driving licence could be built on ISO 18013-5, OpenID4VP and W3C Verifiable Credentials together rather than on one at the expense of the others.

Blerify's argument for ISO is that government credential systems follow installed base, and that the installed base is converging. The European example it cites shows something different. The EU Digital Identity Wallet does not pick a winner: it requires that national identity data be issued in two formats at once, ISO 18013-5 and SD-JWT VC. What the EU mandates is dual issuance of the same credential, not convergence on a single standard. And neither of the two formats it chose is the W3C data model that the region's most deployed identity protocol runs on.

Post-Quantum: The Structural Argument

Unlike Sovra or QuarkID, neither of which has publicly addressed post-quantum cryptography as a near-term design concern, Blerify has incorporated PQC preparation into its technical roadmap from the outset. The rationale connects directly to the paradigm shift the platform is built on.

Self-sovereign identity rests on a specific security assumption: the private key held by the credential owner cannot be forged or extracted. Current identity stacks, including ISO 18013-5 and W3C VCs, use elliptic curve and RSA-based signing schemes. These are vulnerable to attacks by sufficiently powerful quantum computers, which can derive private keys from public ones via algorithms that classical computers cannot run efficiently. The near-term operational threat remains limited by current hardware constraints, but the relevant long-term risk for signed credentials is not decryption, since a credential is not secret, but forgery: once an adversary can derive an issuer's private key from its public key, every credential that key ever signed becomes indistinguishable from a forged one, and the issuer's key has to be rotated and every outstanding credential reissued. For national identity infrastructure whose root keys are expected to operate for decades, that represents a planning problem with a known start date.

Blerify documents post-quantum certificates as an integrated feature of its issuance platform, covering key generation, signing and encryption, with the NIST algorithms finalised in August 2024. It ships them in hybrid mode, running post-quantum alongside classical cryptography so that existing integrations keep working. The founding team published the first implementation of a quantum-resistant EVM blockchain in Nature Scientific Reports, which the journal featured among its hundred most-read papers of 2023. What the hybrid approach works around is the constraint described in interview: ISO 18013-5 and OpenID4VP do not yet recognise the NIST suite as valid signing options, so a government credential signed post-quantum alone would fall outside the standards that make it readable in Apple and Google wallets. Platform-level post-quantum protection is therefore available now; post-quantum signing of the government credential itself waits on the algorithm registries.

No other LATAM identity stack actor appears to have reached this stage of PQC preparation, based on available public evidence as of February 2026. That assessment draws on an interview with Blerify's founder and a review of public documentation for QuarkID, Sovra, and World; it has not been confirmed through direct technical inquiry with those teams.

Confirmed Deployments

Uruguay: AGESIC pilots a blockchain trust list for cross-border credential verification

The strongest deployment evidence in this case study comes from a pilot, co-published with the Uruguayan government, that addresses a problem the rest of this section has not raised.[64]

Verifiable credentials remove the need for a verifier to query the issuer's database, since the credential carries its own proof. Two questions survive that arrangement: whether the issuer is who they claim to be, and whether the credential is still valid. Both require a trust registry. The European Commission operates one for the EU. Latin America has no regional body with the mandate to run shared technical infrastructure, and AGESIC identifies that absence as the reason cross-border verification remains out of reach in the region, the same gap that has limited electronic signature interoperability for more than a decade.[65]

During 2025, AGESIC and Blerify piloted an answer: a trust list deployed as smart contracts on the Avalanche and LNet networks, where each country operates its own node and its own contract. Only that country can edit its list,

enforced through X.509 certificates held in its own HSM. Nodes synchronise roughly every two seconds, so a change propagates regionally in real time through event subscription, with no central operator and no country ceding sovereignty to one.[66]

The pilot ran in five phases, with AGESIC managing the national trust list, a test entity called UruID issuing ISO 18013-5 credentials, and participants authenticating against a demo replica of the Uruguay ID Broker. The final phase is the substantive one. AGESIC rotated its public key and confirmed that credentials issued by UruID remained valid; revoked that key as compromised and confirmed that credentials issued afterwards failed; then revoked UruID from the trust list and confirmed that credentials issued before the revocation still validated while later ones did not. Revocation semantics tested against live infrastructure is more than most identity pilots in the region document.[67] The project won the CDPI Global Government Hackathon at the Global DPI Summit in South Africa in November 2025.[68]

What AGESIC says is still missing

AGESIC is precise about the limits. It describes the work as a pilot throughout, identifies UruID as a test entity, and frames the publication as material for analysis rather than a finished solution. Its own comparison of trust list models scores the decentralised approach as emergent with few precedents on maturity, and as pending construction on regulatory acceptance.[69]

The roadmap it publishes is worth reading against the rest of this report, because none of the three unresolved areas is cryptographic. On technology, the infrastructure needs to become a recognised digital public good with open code, deployment scripts and defined minimum node requirements, and AGESIC states it is seeking the funding to get there. On governance, the region needs to agree which entities may operate a node, and AGESIC names the OAS, the IDB, the World Bank, Red GEALC and CDPI as the bodies that would have to convene that agreement, with a neutral coordinator running incident protocols. On regulation, each participating country needs a decree recognising a decentralised trust list as a valid registry of record.[70]

Set beside Proyecto DIDI, the pattern is hard to miss. DIDI's post-mortem with the IDB identified digital literacy among issuers and validators, the friction between self-sovereign identity orthodoxy and what its users could actually adopt, and the operating costs that outlast the grant. AGESIC, four years later and with a technically more ambitious design, identifies institutional convening and legal recognition. The two projects sit at opposite ends of the maturity curve and neither was constrained by the cryptography.

Panama: Sertracen and the digital driving licence

Driver licensing in Panama is operated under concession by Grupo GSI Sertracen, which runs testing, issuance and renewal on behalf of the Autoridad de Tránsito y Transporte Terrestre. Blerify built the digital licence channel with Sertracen: issuance, verification, and a white-labelled identity wallet holding the credential as a cryptographic object on the holder's device, on ISO 18013-5, OpenID4VP and W3C Verifiable Credentials.[71]

The launch was announced for 12 May 2026, and the option appeared on Sertracen's platform, where the flow ended with a notice that the mobile licence could not yet be requested.[72] As of August 2026 the licence remains unavailable. Adolfo Fábrega, who heads Panama's Autoridad de Innovación Gubernamental, attributed the delay to sequencing: "we moved faster on the technology than on framing the legal side."[73] Publication in the Gaceta Oficial, approval by the ATTT board, inspector training and verification devices are all still outstanding.[74]

El Salvador is at a similar stage. The RNPN stated in July 2026 that development of the digital DUI is complete and awaiting legislative reform.[75] No Salvadoran source names Blerify in connection with it, and El Salvador does not appear among the cases Blerify publishes.[76]

Outside government, Blerify issues a digital membership credential for the Spanish Chamber of Commerce in Panama, which indicates the commercial base supporting the public-sector work.[77] In the February 2026 interview, Allende described further country processes underway but not yet disclosable.

Smaller countries with limited prior biometric infrastructure tend, in the same breath, to have smaller technical workforces and weaker capacity to maintain complex identity systems over time. Legacy drag and institutional fragility often co-occur. Whether Blerify's as-a-service model adequately offsets the operational risks specific to small-state deployments is a question these implementations will test once they reach citizens.

Government Engagement: AI-Driven Threats as the Entry Point

Blerify shares with Sovra a government-first adoption strategy, targeting public administrations before enterprises or individual users. The approaches diverge in two important ways: the geography of target governments, and the entry point for the sales conversation.

On geography, Allende identified a pattern across Blerify's government engagements: countries with populations in the range of 4 to 10 million people are moving faster on digital identity infrastructure than the large LATAM economies. Brazil, Argentina, Mexico, and Colombia built biometric identity systems in the decade before 2020, with substantial investments in hardware, vendor relationships, and institutional expertise. Proposing a new credential layer to those governments means, implicitly, asking them to acknowledge that a recent significant investment is being superseded, a political difficulty that extends timelines considerably. Sovra's confirmed deployments reflect this dynamic directly: Nuevo León (Mexico) and Salta (Argentina) are both jurisdictions within large federal economies where adoption required sustained coalition-building and regulatory work to advance. Countries that did not make comparable prior investments in biometric infrastructure (Panama at 4.4 million inhabitants, El Salvador at 6.4 million, and much of Central America) can adopt ISO 18013-5 without writing off a prior technology cycle. The structural conditions apply more broadly across the region, though which other countries in this range have active digital identity procurement processes underway was not confirmed in this research. One was announced too late in the research period to be assessed here. Peru published guidelines for a national digital identity document in 2026, specifying it on W3C verifiable credentials and decentralized identifiers, with OpenID4VCI for issuance. The programme is noted here rather than assessed[78].

On the entry argument, Allende described a consistent four-step sequence in every government engagement, in which the conversation does not begin with the product or with cryptography. It begins with AI-driven threats. Phishing campaigns are now sophisticated enough to deceive institutional staff and ordinary citizens; deepfakes have reached a quality level that defeats remote facial verification systems currently deployed by banks and government agencies; automated attacks on authentication infrastructure are accelerating. Governments across the region recognize this as a political liability and, as Allende characterized it, feel a genuine responsibility to protect citizens and institutions from fraud their current systems cannot handle. This framing opens conversations that technical product presentations do not.

Only after establishing the threat context does the conversation move to why existing authentication infrastructure is breaking down: passwords are phishable, and biometric systems that were meant to replace them can now be deceived by AI-generated imagery. This is the hardest moment in the conversation with governments that have recently invested in biometric systems, as the implication that a significant recent expenditure is being outpaced by adversarial capability requires careful handling. The third step reframes the solution as alignment with a global regulatory trajectory: the EU has mandated digital identity wallets under eIDAS 2.0, more than 25 US states have live digital driving licenses in Apple and Google Wallet, and the standard is spreading. A government adopting ISO 18013-5 follows a path that advanced economies have already committed to, a meaningful shift from evaluating an unfamiliar vendor to joining an observable international movement. Only at the fourth step does Allende present the actual product: how it works, why it addresses the vulnerabilities described, the user experience.

The sequencing matters because it starts where government decision-makers already are (concerned about fraud, accountable for citizen harm, and increasingly skeptical of systems they sense are failing) before moving to infrastructure they have not yet evaluated. Technical presentations alone stall on procurement complexity; the AI threat argument functions as the entry point that makes the subsequent conversation possible.

AI Use Cases: Prioritization and Pilot Implications

Allende draws a clear line between two categories of authentication that digital identity infrastructure can address, and treats that distinction as a sequencing decision, not a product menu.

The first category is high-level authentication: bank enrollment, financial transactions, onboarding for regulated services. This is where AI-generated fraud is already costing money in measurable amounts. Deepfakes defeat remote biometric matching used by banks for KYC; phishing campaigns impersonate financial institutions convincingly enough to extract credentials from customers and staff. The financial sector has both the highest exposure and the clearest business case for replacing remote biometrics with cryptographic credential verification. Allende frames this as the primary channel because the urgency and the paying customer are both located here. For pilot design, this points toward financial institutions, regulated fintechs, and bank-adjacent government services (social benefit disbursement, tax authority enrollment) as the highest-return entry points in the near term.

The second category is low-level authentication: daily logins to web platforms, mobile apps, content sites, replacing passwords and CAPTCHA. The problem is real and affects far more interactions per day, but the economic pain per incident is lower and the institutional buyer is less defined. Allende's view is that this will follow the financial channel once the infrastructure is in place, rather than lead it.

On the emerging frontier, Allende identified two domains where the conversation is likely to move next. The first is AI agents. As autonomous agents act on behalf of users across government systems, financial platforms, and business workflows, the question of how an agent authenticates, and how its delegated authority is scoped and revoked, becomes a direct identity infrastructure problem. Allende's expectation is that identity wallets will develop modules for agent control and delegation: a credential that says, in effect, "this agent is authorized to act for this person, within these limits, until this date." Blerify has not built this component yet, but it is on the roadmap.

The second is social media and content platforms. Allende described wanting a mechanism by which users could prove they are real humans, and prove basic attributes such as gender or age, without revealing their civil identity, using zero-knowledge selective disclosure. The target problem is the anonymous account as a vector for coordinated abuse: mass fake profiles, automated hate speech, bot-driven amplification. A ZK proof of personhood attached to a social account would not eliminate abuse, but it would raise the cost of running it at scale. Allende was careful to note this is not a use case Blerify is actively pursuing (the company sells identity verification infrastructure, not vertical applications) but he sees it as a natural downstream use of the rails being built.

For pilot design across the research agenda, these priorities have concrete implications. Financial sector pilots (with banks, regulated fintechs, or government disbursement programs) are the most tractable near-term path and align with where institutional demand already exists. Social media and agent identity are medium-term bets that depend on the financial infrastructure being established first; pilots in those domains would be premature without a working credential layer underneath.

Key Takeaways

Match the credential format to the credential type, not to the vendor. Blerify issues in both W3C verifiable credentials and ISO 18013-5 mdocs, applying a single rule: where an international standard already fixes the fields, as with driving licences, or a jurisdictional profile does so on the same format, as with national IDs, use it; where none exists, as with diplomas, professional licences or KYC results, the issuer defines the schema. Both travel over the same OpenID protocols, so supporting each is a product decision rather than an architectural commitment.

The four-step government entry sequence is replicable. Open with AI-driven fraud, then explain why existing authentication is failing, then anchor the solution in international regulatory trajectory, and only then present the product. The hardest step is the second, because it asks a government that recently invested in biometric systems to accept that the investment is being outpaced.

Hybrid deployment is how post-quantum ships before the standards catch up. Blerify runs post-quantum alongside classical cryptography so existing integrations keep working, which is what lets it offer PQC as a platform feature while ISO 18013-5 and OpenID4VP have yet to add the NIST algorithms to their registries.

Appendix: Blerify Classification

Dimension	Detail
What it does	Commercial digital-identity platform for institutional issuers: a wallet plus APIs and SDKs to issue and verify credentials in W3C and ISO formats, with a decentralised trust registry of authorised issuers.
Relevance for AI privacy-risk mitigation	Signed credentials replace the remote biometric checks that deepfakes defeat. AI enters as the argument that opens government conversations, not as a component of the deployed workflow.
Status	Documented pilot (Uruguay, decentralised trust list with AGESIC, 2025). Built, not launched: Panama digital driving licence (announced 12 May 2026, unavailable as of August 2026) and El Salvador digital DUI (development complete, awaiting legislative reform as of July 2026).
Deployments and counterparts	Uruguay: AGESIC, the national e-government agency, in a pilot with a test issuer. Panama: the transit authority through Grupo GSI Sertracen, the licensing concessionaire. El Salvador: RNPN, though no Salvadoran source names Blerify in connection with it. No public issuer in production.
AI Risk Domain (MIT)	Privacy & Security
AI Risk Sub Domain (MIT)	2.2 AI system security vulnerabilities and attacks; 4.3 Fraud, scams, and targeted manipulation (deepfakes defeating remote biometric verification, KYC fraud)
Intent (MIT)	Intentional
INATBA categorization	Secure Data Sharing & Granular Access Control; Data Privacy & Individual Ownership; Consent Management Across the AI Lifecycle
Credential format and proofs	W3C Verifiable Credentials and ISO 18013-5 mdocs, chosen per credential type; selective disclosure through BBS+ for W3C and native mdoc selective disclosure for ISO; no zero-knowledge proofs. NIST post-quantum algorithms integrated in hybrid mode alongside classical signatures.
Anchoring / settlement	Decentralised trust list deployed as smart contracts on Avalanche and LNet (Hyperledger Besu, ex-LACChain), with each country running its own node and contract. Does not settle to Ethereum.
Standards implemented	W3C Verifiable Credentials Data Model; ISO/IEC 18013-5; OpenID4VCI; OpenID4VP; NIST PQC in hybrid mode, with PQC signing of government credentials pending ISO and OpenID registry recognition. Any wallet supporting OpenID4VCI/VP.

The four cases so far issue credentials, whatever the standard, and answer the same question: whether a document is authentic and belongs to the person presenting it. World answers a different one, whether the holder is a distinct human being at all, and builds on neither standard.

World

World, co-founded by Sam Altman, is one of the most visible proof of human experiments in Latin America. It uses biometric capture plus cryptography to let users prove they are human without revealing civil identity.

From its launch in 2021, World framed itself as an ‘AI-moment’ project. As bots and synthetic media scale, separating humans from automated agents gets harder. Comment “armies” on social platforms, automated activity on dating apps, and realistic deepfakes all contribute to a gradual erosion of epistemic trust in digital environments.

Within this context, World’s narrative positions decentralized proof of human mechanisms as a candidate response. World argues users keep custody of their identifiers and prove humanness in a privacy-preserving, ‘neutral’ way. The goal is portability across apps.

Like QuarkID, which serves a city of 3.6 million and counts 1.8 million DIDs created on-chain, World has also reached significant scale in Latin America, roughly 7 million people. Tools for Humanity put the regional figure at eight million verified users at a press event in Mexico City in June 2025[79]. The company’s own country statements do not add up to it: they place Orb-verified humans at roughly 2.5 million in Argentina, 1.8 million in

Colombia, 1 million in Peru, 700,000 in Mexico and 500,000 in Brazil, on figures dated between March and December 2025[80]. That is a regional total near seven million, against roughly twice as many World App accounts.

Nonetheless, World sits in deliberate contrast to miBA and Sovra. They diverge on four fronts: what they prove, how they frame AI risk, how they scale adoption, and how they engage regulators.

First, miBA and Sovra anchor identity in government documents and verifiable credentials, then use ZK proofs for selective disclosure (e.g., 'over 18'). World anchors humanness and uniqueness in iris biometrics and turns that into a reusable cryptographic identifier ('World ID').

In addition, World is explicitly targeting AI-related use cases through its digital proof of human stack, something that neither Sovra nor miBA has placed at the center of its public narrative to date, even if both privately and publicly acknowledge that AI will play a big role in their strategy and society.

Moreover, their adoption strategies are nearly inverted. Sovra started with governments, then moved downward toward citizens, following an "inverted" adoption curve. World started with grassroots onboarding, then added corporate partners in consumer sectors. It has not yet built deep, sustained collaborations with Latin American governments.

Finally, early controversy (token incentives, consent practices, and privacy compliance) has made regulator and government engagement harder in the region.

Because it sits at the opposite corner of the design space, World is a useful counterpoint case study.

Technical Overview

World's proof of human stack is built around the **Orb**, a biometric device that takes a picture of a person's iris and derives an anonymized "mathematical code". According to project documentation and independent case studies[81], the raw image is processed into a cryptographic identifier and then deleted, leaving only an artifact intended to represent a unique human. This artifact feeds into **World ID**, a credential that applications can use to enforce "one person, one account / one action" without learning the user's civil identity.

On the cryptographic side, World ID uses zk-SNARKs (Groth16) and the Semaphore protocol to prove anonymous membership in a set of verified humans, employing Merkle trees for group membership and nullifiers to prevent correlation across services. The privacy story combines zero-knowledge proofs with elements of secure multi-party computation (SMPC) in parts of the stack. The April 2026 upgrade described below changed the credential and account layer rather than the proof system.

Where that proof system runs is worth stating precisely, because it is the strongest link to Ethereum in this report. The canonical identity tree lives on Ethereum mainnet: the Identity Manager contract holding the Semaphore instance is deployed only there, alongside the World ID Router. One state bridge per network then publishes the merkle root outward to World Chain, Optimism, Base and Polygon, where applications verify proofs against a bridged copy that lags mainnet by roughly five minutes on the OP Stack chains and forty on Polygon. World Chain itself is an OP Stack rollup that settles to Ethereum and posts its data to Ethereum blobs, though it remains at Stage 0: fault proofs run in permissioned mode, with a single entity able to propose and to challenge, and a sequencer escape hatch through L1 carrying up to a twelve-hour delay. What is proprietary in World is the Orb hardware and parts of its firmware, not the location of the identity state[82].

In contrast, **QuarkID** follows an SSI model centered on **W3C DIDs and Verifiable Credentials**: an issuer signs a VC and the holder later presents it (or a derived presentation) for verification; in QuarkID's own quickstart, the stack uses **BBS (BLS12-381) suites** specifically to enable **selective disclosure / derived proofs**, i.e., proving specific claims from a credential without revealing everything.

Importantly, World ID is primarily specified and implemented as a privacy-preserving proof of human system, rather than as a W3C DID/Verifiable Credentials issuance-and-presentation stack. As a result, interoperability with

government SSI deployments that are explicitly DID/VC-aligned (e.g., QuarkID/Sovra-style architectures) is not "native" and typically requires an adapter layer (e.g., issuing a VC that attests a World ID verification).

World's own timeline reinforces this non-standard path for credentials:

between March and August 2023, the SDK explicitly framed DIDs and VCs as a *future* capability ("OIDC/OAuth2 now; VCs/DIDs (and SIWE) later / in the future")

On October 17, 2024, World ID 3.0 introduced **World ID Credentials**, enabling users to store information from NFC-enabled passports in World App and prove attributes such as age or nationality without revealing identity, but neither the launch post nor the help center descriptions of these "Credentials" (passport / national ID scanning stored on-device) referenced the W3C VC or DID specifications;

By December 30, 2025, the World ID 4.0 RFC had deepened this *World-specific* credential architecture with concepts like **credentials with committed metadata** and an **issuer registry** tied into the proof system, again without presenting it as conformant with the W3C Verifiable Credentials Data Model or any DID method.

On April 17, 2026, World ID 4.0 shipped: an **account-based architecture** replacing the device-bound credential, with key rotation and recovery, multi-device sessions, **single-use nullifiers** to prevent cross-site correlation, and an open-source SDK that lets any app act as a World ID authenticator[83].

Conclusion and research gap. By design, the two architectures differ. The open question is whether a bridge is possible, and what it would require technically and institutionally. One that allows the millions of users already onboarded into government-backed DID/VC systems to interact with the World ID ecosystem, and, conversely, enables the millions of World ID-verified users to become legible to governmental and SSI infrastructures beyond World-specific mini-apps.

Specifying what it would take to construct this bridge points to at least three avenues for further research. First, to **systematically compare the types of ZK proofs** used in both approaches (Sovra/miBA/QuarkID vs. World), not only in terms of efficiency or security, but in how they shape the possibility of building bridges to state identity systems based on signed credentials and traditional document repositories.

Second, to explore whether World has, or could develop, an explicit strategy to **align its credential architecture with the W3C ecosystem of DIDs and VCs** currently used by multiple governments: for example, by issuing "derived" VCs that certify the existence of a World ID without requiring direct use of the Orb, or by defining interoperability profiles compatible with frameworks such as QuarkID or eIDAS.

Third, to assess the extent to which it would be technically feasible and normatively desirable to **certify humanity without relying exclusively on Orb-based biometrics**, instead leveraging data sources and verification channels already present in state infrastructure (civil registries, tax databases, social security systems) combined with standardized zero-knowledge proofs.

Protocol Standardization

A second axis of divergence concerns protocols and standardization. In Latin America, projects like Sovra and miBA deliberately invested in building a **standardized DID/VC protocol (QuarkID)** as a reusable blueprint for governments. As described by those involved in the design of QuarkID, this has been slow and patience-intensive, but it has produced a shared standard that multiple jurisdictions can adopt without becoming dependent on a single vendor. In practice, this has meant adapting their stacks to existing governmental needs and infrastructures, rather than asking governments to re-architect their systems from scratch.

World, by contrast, has so far viewed this kind of standard-setting, "meeting governments where they are" in terms of existing digital identity infrastructure, with greater skepticism, in part because of the technical and institutional complexity involved in adapting to heterogeneous, legacy government systems.

As one World engineer emphasized, the team tends to see direct work with governments as structurally difficult, not only for political reasons but because most public-sector identity systems run on **legacy cryptographic rails**. These deployments typically rely on SHA-based hash functions and RSA/ECDSA signatures, which are not "ZK-friendly" by design.

Given this, World argues that proving legacy government schemes inside ZK circuits is expensive. At scale, it would require country-specific integration work. From World's perspective, this creates a strong incentive to prioritize World's ZK-native stack optimized for new applications, rather than bending that stack to fit heterogeneous, decades-old government systems.

If World's proof of human stack is to be deployed as an AI-risk mitigation layer in Latin America, government buy-in will be essential, and that almost certainly requires some degree of protocol standardization.

The broader literature points in the same direction. An August 2024 paper, revised in January 2025, whose authors include researchers at OpenAI, MIT and Microsoft, already calls for a re-examination of **standards for remote identity verification and authentication**, highlighting frameworks such as NIST's Identity Assurance Levels (US), ENISA's practices for remote ID proofing (EU), and ISO/IEC 29115 on entity authentication, while remaining silent on Latin American standards[84].

The open challenge, therefore, is to identify **which protocols make sense in this context**: QuarkID is an obvious candidate on the cryptographic side, given that it is already live at city scale in Buenos Aires, but there is no equally clear answer yet for digital identity and biometrics. Designing a path for World to "meet governments where they are" on those layers is a central research gap for the next phase of this work.

Go-to-market strategy: bottom-up vs. inverse adoption

On the go-to-market side, World's trajectory again contrasts with Sovra and related DID/VC projects. Sovra and its peers bet on an **inverse adoption curve**: starting with governments, then extending to enterprises, and only later reaching individual citizens. This path has been slower but has gradually secured institutional buy-in and created room for QuarkID to function as a shared public-sector standard.

World, by contrast, pursued a **bottom-up strategy** from the outset. The initial rollout focused on mass user acquisition: deploying Orbs in high-traffic locations (shopping malls, commercial centers, public spaces) and offering token incentives for verification. This approach rapidly onboarded millions of people globally, **around 7 million in Latin America alone, the region with the highest deployment**[85]. It has been remarkably successful in terms of raw reach, but also highly contentious, as the next section will detail.

As part of this **rollout strategy**, and consistent with its broader bottom-up go-to-market, World announced its Partner Markets Program for large enterprises in September 2025. The program explicitly targets major **retail and consumer-service networks as a primary distribution channel**, placing Orbs inside high-traffic private venues and embedding World ID verification into checkout and sign-up flows customers already use[86].

In Japan, this strategy is exemplified by the **Medirom / Re.Ra.Ku** partnership, where Orbs have been deployed across a large wellness-chain footprint and explicitly framed as a scaling channel for World ID adoption across hundreds of locations[87]. In Latin America, it follows a similar trajectory: World has partnered with **Rappi** as an "Orb-on-demand" logistics partner, bringing verification directly to users through a delivery workflow[88], and with **CONMEBOL Sudamericana** to deploy Orbs across **75+ football matches**, verifying fans and unlocking in-stadium experiences[89].

Beyond retail and sports, World has also pursued partnerships in sectors where bot activity directly threatens core user experience. In interviews for this research, World's team described three strategic verticals: gaming, social media, and dating applications, the last of these in markets where automated bots and AI-generated profiles increasingly degrade platform trust. Specifically, World partnered with Tinder in Japan to enable users to verify they are interacting with real humans rather than AI-generated "catfish" accounts[90].

The pilot works as an in-app age/"real human" verification flow that hands users from Tinder Japan to the World App to complete verification and then returns them to Tinder with a visible status. In practice, a user in Japan taps the World ID option inside Tinder's age verification settings, is redirected to the World App (installing it first if needed), and completes an 18+ check using supported identity documents such as a Japanese My Number Card or a passport. Once the World ID verification is completed, the user links that verified World ID back to their Tinder account, which can unlock age-gated functionality and display a verification badge (often framed as "Verified Human") to signal additional trust to other users[91].

However, **no comparable partnership has been identified in Latin America across these three verticals** (gaming, dating, social media). Despite the region representing World's largest user base globally, the absence of such an enterprise integration means the technology remains largely confined to standalone World App use cases rather than embedded into platforms where users already spend their time. **The challenge, therefore, is identifying the killer use case for the region**, one that addresses a locally urgent problem, aligns with existing user behaviors, and can serve as a reference point for broader enterprise adoption.

That gap widened in 2026. In April, World announced a much broader enterprise push at its Lift Off event in San Francisco. Tinder's Japan pilot expanded to the United States, with a verified badge and free profile Boosts as the adoption incentive; Zoom added a Deep Face check that cross-references the signed image from Orb enrollment, a live selfie, and the video feed other participants see before displaying a "Verified Human" badge; DocuSign attached proof of human to signature workflows, so that each action, direct or delegated, links back to a verified person; Vercel shipped a human-in-the-loop step in its open-source Workflow SDK; Okta began building Human Principal to verify the human behind an AI agent; and Concert Kit launched to reserve tickets for verified humans. World reported close to 18 million verified people across 160 countries at the time of the announcement[92]. The depth of these integrations should not be overstated: Match Group had already made its own Face Check selfie verification mandatory for new Tinder users in October 2025, live in seven countries plus California and including Colombia, and World ID sits on top of that as an optional badge rather than as the platform's verification mechanism[93]. None of the launches named a Latin American market, and they arrived while sign-ups in the region were paused[94].

Moreover, what has *not* yet materialized, however, is a comparable level of **government buy-in**. In interviews, team members consistently point to structured collaboration with states and regulators as one of their main outstanding challenges and a necessary next step. Given the previous sections, this gap is a core obstacle to deploying World's technology as an AI-risk mitigation layer in Latin America, where meaningful impact ultimately depends on interoperability with public digital identity infrastructures.

Beyond the three verticals currently mapped by Tools for Humanity, collaborator of World Network, this research points to further applications in Latin America. For example, on-chain voting for presidential and legislative elections may represent a natural fit for proof of human infrastructure in a region already piloting such initiatives. Under Law 32270, Peru's National Office of Electoral Processes (ONPE) built the Solución Tecnológica del Voto Digital, a blockchain-based internet voting system, mandatory for police and armed forces personnel posted away from their place of residence and available to Peruvians living abroad, audited before and after each electoral process[95]. Integrating World ID into such workflows could address a core vulnerability: ensuring that each vote corresponds to a unique, verified human without compromising ballot secrecy or voter privacy, a use case where civil registries and zero-knowledge proofs align directly with democratic values.

Initial controversies

Beyond protocol choices and go-to-market strategy, a third factor has significantly constrained World's ability to work with governments in Latin America: the way the project structured its **economic incentives**, the quality of **informed consent** around biometric collection, and its broader **data practices and regulatory compliance**.

These three axes of controversy have shaped how regulators, civil society, and potential public-sector partners perceive the project from the very beginning. In practice, they have made it difficult for World's team to move be-

yond the headlines when approaching regulators: rather than engaging in a deep conversation about the underlying technology and its potential uses, interlocutors tend to stay stuck dissecting the "bad press", and discussions rarely progress to more substantive ground, a dynamic repeatedly described by World's team in the interviews conducted for this research.

Economic incentives through WLD token

World's early growth model tied identity enrollment to **direct economic rewards**. In field tests and early-market rollouts, operators offered **the promise of "free tokens"** to people who completed Orb-based iris verification, effectively turning biometric registration into an on-the-spot cryptocurrency transaction. This approach later formalized into the project's public distribution mechanics: verified users in eligible countries can **opt in** to claim **WLD token grants/airdrops**, with World, then titled Worldcoin, explicitly framing Orb verification as unlocking the **highest** claimable amounts.

Reporting from Argentina, for example, described people getting verified **in exchange for roughly "$50 in crypto"** amid acute inflationary pressures, illustrating how macroeconomic context can amplify the coercive feel of token incentives[96].

Critics (especially in unequal, high-inflation contexts) argued that this incentive structure can **undermine "freely given" consent** by making participation financially compelling for economically vulnerable groups. The sharpest articulation of this critique came from investigative reporting on World's early recruitment. Back in 2022, MIT Technology Review alleged that World representatives used **deceptive marketing practices**, collected **more personal data than acknowledged**, and failed to secure **meaningful informed consent**, claims that, even when contested by the company, helped set a default public narrative of "payment for biometrics"[97].

That reporting describes World's 2021-2022 field operations, not the system as it runs today. It set off regulatory reviews in several jurisdictions[98], and the company has since revised its consent flow, its data retention defaults, and its public communications. What changed, and what did not, is the subject of the next section.

2. Informed consent and regulatory compliance

World's enrollment process was typically embedded in a high-throughput, in-person flow: people were onboarded at Orb locations, prompted to download/use the World App, and asked to accept a biometric consent/terms flow as part of verification. In its public privacy messaging, World has consistently argued that the system is *privacy-preserving by design*, including claims that it does not maintain a centralized biometric database and that biometric images are deleted unless the user explicitly opts into additional data uses.

The recurring critique targets the quality of that consent under real deployment conditions: whether users clearly understood (i) what data was collected (iris/face images vs derived identifiers), (ii) the purposes and downstream uses, (iii) retention/deletion logic, and (iv) how to exercise rights such as withdrawal, erasure, or access[99].

These concerns have shown up explicitly in regulatory language in Latin America. In Argentina, Agency for Access to Public Information (AAIP) launched an investigation into the project's data practices, citing concerns about its handling of sensitive data. In the province of Buenos Aires, authorities have even imposed a fine of millions of pesos for "abusive clauses" in user contracts, which included provisions that waived users' rights to a class-action lawsuit and mandated disputes be handled in foreign jurisdictions[100].

In Mexico the regulator vanished mid-investigation. In 2024 the Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (INAI) opened an ex officio inquiry into World's model. It never published a result. The INAI was formally extinguished in March 2025 and its data protection powers moved to Transparencia para el Pueblo, a deconcentrated organ inside the Secretaría Anticorrupción y Buen Gobierno, taking the inquiry with it. Access to information requests filed later that year were refused or declared nonexistent, with the new authority acknowledging that it keeps no registry of the biometric firms operating in the country[101].

In Chile, the issue escalated into constitutional litigation: the Supreme Court ultimately ordered the deletion of a minor's biometric data collected without valid consent (and required proof of compliance), underscoring a judicial finding that the project's consent safeguards failed in at least one high-profile case[102].

Brazil went furthest. In January 2025 the national data protection authority (ANPD) ordered Tools for Humanity to stop offering cryptocurrency or any financial compensation in exchange for iris collection, holding that payment "may interfere with the free expression of the will of individuals," particularly where vulnerability and need make the offer weigh heavier[103]. World rejected the characterization, stating that there is "absolutely no exchange of money or tokens when a person verifies their World ID," that it stores no iris data, and that it "respectfully disagrees with ANPD's interpretation"[104]. It appealed, arguing it had technically blocked Brazilian users from claiming WLD, and lost: the ANPD upheld the measure and attached a daily fine of R$50,000.

Brazil and Mexico ended very differently. World's answer to the ANPD was to change the product rather than leave. By December 2025 the company said verification had resumed with no token and no financial reward attached, the only country in its network where that is the case, and that it planned to install Orbs in more cities through 2026 with the incentive kept out of the equation[105]. Pressure kept building anyway. The CPI da Íris, a commission of inquiry seated in São Paulo's City Council, approved a 161-page final report unanimously in April 2026, finding that the population had been used by Tools for Humanity and was not properly informed of the irreparable risks of iris scanning, and that the model concentrated on economically vulnerable citizens[106]. In July 2026 the São Paulo public prosecutor filed a civil action against Tools for Humanity and Amazon AWS Serviços Brasil seeking R$240 million in collective moral damages, alleging that the company kept distributing economic incentives through the World App after the ANPD suspension, and that more than 400,000 residents were scanned for payments between R$300 and R$700[107]. The company's account of its own compliance is therefore contested, and the regional picture has since tightened further: by April 2026 sign-ups across Latin America were paused with several investigations open[108].

In Mexico nothing was forced, because there was no one left to force it. World says it complies with Mexican data protection rules and cooperates with the competent authorities, and roughly 700,000 people had been verified there by August 2025[109]. No authority is in a position to confirm or contradict either claim.

Brazil had an authority in a position to demand a change. Mexico, by then, did not.

World's responses to the initial controversies

World's response clusters into three moves: rebrand the project, increase transparency, and adjust incentives/consent/data practices. The goal is to move the debate from scandal to governance and technical detail. That shift is still incomplete.

1. Reframing the project and its purpose

A first, symbolic but non-trivial move has been to **distinguish "World ID" from "Worldcoin"** and increasingly foreground the former as the main object of long-term investment. In interviews, team members emphasize that:

the **token is one component** of a broader ecosystem, not the core of the proof of human stack

the **proof of human layer** should be evaluated independently from speculative or distribution dynamics around WLD

This doesn't erase the token history. It's an attempt to frame World ID as infrastructure, not an airdrop machine.

2. Radical transparency as a trust-building strategy

World calls its approach 'radical transparency'. Concretely, this includes:

After its initial release, and following their original roadmap, World published detailed technical docs: primitives, protocol flows, and threat models.

Commissioning and publicizing **external security and privacy audits** of both hardware (Orb) and software components.

They have open-sourced many components of the stack (Orb firmware, client libraries, proof code) to enable external scrutiny, and have stated their intention to maximally open-source the project going forward.

They put senior engineers in long-form interviews and public Q&As with journalists, civil society, and regulators.

Their logic is: if the charge is opacity, the answer is documentation and independent review. They're betting transparency creates a baseline for serious regulation.

3. Adjusting practices around incentives, consent, and data

World also claims to have made **incremental adjustments** to how verification campaigns are run and how data is handled, though the scope and sufficiency of these changes are still debated.

On **economic incentives**:

They say WLD grants are now opt-in, geo-limited, and reviewed for compliance.

In jurisdictions such as Brazil, where regulators ordered the suspension of token-based incentives, World has **complied with local decisions** and adjusted its operations accordingly.

Messaging shifted from 'free tokens' to access, anti-fraud, and platform integrity.

On **consent and user understanding**:

They cite revised consent flows, clearer in-app language, and standardized operator scripts.

In some markets, the company has collaborated with local partners to **localize explanations** (language, examples, FAQs) and align them more closely with national data-protection norms.

The team also highlights user-side controls (e.g., the ability to delete an account or revoke certain permissions) as mechanisms to reinforce individual agency.

On **data practices and regulatory engagement**:

World has **participated in investigations** and regulatory processes in countries like Argentina, Chile, and Brazil, providing documentation and responding to information requests.

They've paused or adjusted operations where regulators intervened. They argue the architecture reduces biometric risk by deleting raw images and retaining only derived outputs.

Internally, they frame ZK/SMPC/attested hardware work as continuous alignment with evolving data-protection expectations, not a one-off compliance sprint.

Even with these responses, **significant tensions remain**:

Radical transparency does not automatically translate into **regulatory comprehension**. As noted in our interviews, many authorities lack time, incentives, or capacity to engage deeply with technical materials: World's technical progression iterates rapidly, while oversight processes can take a year or more to review documentation that, by then, may already be partially obsolete.

Changes haven't erased the initial reputational damage. In Latin America, 'tokens for biometrics' still dominates the narrative.

World still lacks government pilots that prove its revised practices meet the bar for high-risk biometrics.

In practice, World enters meetings carrying the weight of earlier headlines. This has limited World's ability to do three things: build sustained regulatory relationships, run institutional pilots comparable to Sovra's, and co-design pilots with multilaterals that already operate proof-of-personhood experiments on the ground

For the purposes of this case study, World's responses can be read less as a closed chapter and more as an **ongoing experiment in damage control and norm-setting**: a project that is trying to shift from controversy toward institutional dialogue, but that still carries the weight of its initial choices into every conversation with governments and multilaterals in Latin America.

From long-term infrastructure to near-term risk management

In interviews, World's regional policy team described a recurring problem: even when they can address privacy governance and compliance, many policymakers still treat proof of human and cryptographic credential rails as "future infrastructure": useful someday, but not urgent now. Their claim is that this timing assumption is already wrong. AI-enabled fraud (voice cloning, synthetic IDs, automated scam campaigns) is scaling faster than the public institutions designed to contain it, so the question is no longer whether these tools *might* be relevant, but whether governments can afford to wait for the next regulatory cycle. To make that case legible, they argue the conversation must move from protocol language (ZK, primitives, threat models) to concrete harms and everyday use cases, framed in terms regulators already prioritize: citizen protection, rights, and trust in public services.

Recommendation: Build (and standardize) an AI-risk "urgency narrative" package for public-sector engagement, one that leads with measurable, near-term harms and only then introduces World ID as a bounded mitigation tool under rights-based governance.

Key Takeaways

Standardization gap: Without shared standards, World lacks common ground with governments. Even partial alignment could accelerate institutional dialogue

Grassroots lesson: Incentives + Orbs scaled fast, but produced a lasting exploitation narrative. The challenge now is to turn that footprint into learning infrastructure and move toward rights-based design, open standards, and co-design

Education: Open Sourced docs aren't enough. World needs regulator- and citizen-facing education that clarifies capabilities and limits, addresses real risks, and translates PETs into policy language.

Appendix: World Classification

Dimension	Detail
What it does	Proof-of-human network: the Orb captures an iris image, derives an anonymised identifier and lets a person prove "I am a unique human" through zero-knowledge proofs, verifiable by apps on EVM chains.
Relevance for AI privacy-risk mitigation	The only case built against an AI risk from the start: proof of human for an internet crowded with bots and AI agents, in production with commercial partners and without government integration.
Status	Live, private (sign-ups across Latin America paused as of April 2026, with several investigations open). Enterprise integrations announced in April 2026 named no Latin American market.
Deployments and counterparts	Latin America-wide through Orb operators; no government integration in the region. Counterparts are commercial partners (Tinder, Zoom, DocuSign, Okta, Vercel).
AI Risk Domain (MIT)	Privacy & Security; Malicious Actors & Misuse
AI Risk Sub Domain (MIT)	4.3 Fraud, scams, and targeted manipulation; 4.1 Disinformation, surveillance, and influence at scale (AI-driven impersonation, automated fake accounts, Sybil attacks on AI-mediated services)
Intent (MIT)	Intentional
INATBA categorization	Data Privacy & Individual Ownership; Proof of Non-Personhood
Credential format and proofs	World ID, a proprietary proof of unique humanity, not W3C-aligned; zk-SNARKs (Groth16) over a Semaphore group with per-app nullifiers.
Anchoring / settlement	Identity Manager and World ID Router on Ethereum mainnet; per-chain state bridges publish the Merkle root to World Chain, Optimism, Base and Polygon. Settles to Ethereum.

Dimension	Detail
Standards implemented	Semaphore protocol; Groth16 zk-SNARKs. Not based on W3C DID/VC.

Evidence Summary

One finding runs across all five cases: the credential rails are built and the AI layer is not. Only one deployment here has its cryptographic identity operationally directed at an AI risk, and it is private, outside the region, and not interoperable with any government stack. The table below compares the five on the dimensions that actually vary between them.

Case	Credential format	Proofs	Directed at an AI risk	Reach	Counterpart
QuarkID / miBA	W3C DID/VC	Selective disclosure (BBS+)	No	Live at city scale	Government
Sovra	W3C DID/VC	ZK attestations, vendor-described	Advertised, not connected to the rails on the public record	Live in two jurisdictions (Nuevo León, Salta)	Government
Blerify	W3C VC and ISO 18013-5	Selective disclosure (BBS+, mdoc); no ZK proofs	The entry argument to government, not a component of the workflow	Uruguay pilot documented; Panama built but not launched; El Salvador built but not launched, attribution to Blerify unconfirmed by Salvadoran sources	Government
World	Proprietary, not W3C	zk-SNARKs (Groth16), in production	Yes: bot and AI-agent verification in production	Regional scale, with sign-ups paused across LATAM	Private
Proyecto DIDI	W3C DID/VC	None documented	No	Pilot concluded 2022	Multilateral

The international precedents described in the introduction are instructive but not transferable. Anon Aadhaar demonstrates a technical path without being operational at government scale, and Bhutan and the UN pension fund show that small or specialized entities can move faster than large-state bureaucracies. None of those models maps directly onto LATAM's institutional landscape without significant adaptation.

Cross-Cutting Findings

1. The infrastructure was built for something else.

The credential rails documented here are more mature than the literature on the region records, and not one of them was commissioned against AI risk. QuarkID, Sovra and Blerify were all built to issue documents, deliver benefits and simplify administrative procedures. Where AI has entered, it entered afterwards and from the private side: Blerify now designs against deepfakes defeating biometric verification, and Sovra advertises an identity layer for government AI agents. No government program in the region names AI risk among the purposes of the credential infrastructure it already operates, which leaves the case for this technology being made by the companies selling it and by nobody who runs it.

2. Two cryptographic answers to the same problem, rooted differently.

Blerify and World start from the same premise: AI has broken remote identity verification, with deepfakes defeating the face check a bank runs and synthetic accounts defeating the sign-up flow a platform runs. Their answers diverge at the root of trust. Blerify removes the biometric check and replaces it with a credential an authority has signed, which travels through any standards-based verifier and requires an institution willing to issue it. World keeps the biometrics and moves them to enrolment, proving a person is unique once and then proving membership in that verified set with zero-knowledge proofs. Sovra has announced work in the same direction without connect-

ing it to the credentials it issues. The two deployed approaches serve different parties: one works for institutions that already hold issuing authority, the other for platforms that hold none.

3. Adoption strategy shaped the outcome more than technical design did.

Sovra and Blerify went government-first and have institutional counterparts to show for it, live programmes in Sovra's case and a documented government pilot in Blerify's, at the cost of slower scale and sustained coalition-building. World went grassroots-first and reached more people in Latin America than any other actor, then found the path to government harder than the projects that started there. DIDI inverted both, opening with multilateral backing and an inclusion mandate, and did not continue past its pilot, which documented digital literacy and operating costs as the barriers. The three sequences produced three different failure modes, none of which the record attributes to the technology. Five cases are enough to notice the pattern, not to establish it.

4. Institutional capacity and literacy sit on top of the technical work.

The visible blockers are legal instruments: Panama's digital driving licence was delivered and announced and remains unavailable, pending publication in the Gaceta Oficial and board approval at the transit authority; El Salvador's national identity credential is complete and waiting on legislative reform; AGESIC states that its trust list requires national decrees recognising it as a valid registry before it can run in production. Behind each of those sits the same condition. None of these deployments found a policy framework already written for cryptographic credentials, the officials who would draft one cannot evaluate a zero-knowledge proof system without help, and the review cycles that produce such instruments run longer than the technology takes to change. DIDI hit the same wall from the other side, in the digital literacy of the people it was meant to serve. Across five cases at five stages of maturity, the documented blockers were legal, institutional and social; none of the sources consulted identifies cryptography as the binding constraint.

Open Research Questions

The evidence accumulated in this section closes some questions and opens others. These are the gaps that remain after this work:

The standards bridge question: Is a technical and institutional bridge between W3C DID/VC systems (QuarkID/Sovra) and the World ID ecosystem feasible? What would it require, and who has the incentive to build it? This is researchable and consequential enough to warrant a dedicated investigation.

DIDI's post-2022 trajectory: The project published its Gran Chaco pilot results in 2022 and no continuation is documented. Whether that is a funding issue, a design issue, or an institutional issue matters for assessing the viability of IDB/LACChain-backed models more broadly.

Post-quantum competitive landscape: Blerify's founder characterized competitors as unprepared on PQC. That characterization has not been verified through direct technical inquiry with QuarkID/Extrimian, Sovra, or World. A targeted technical assessment would either confirm a genuine gap or surface preparation that isn't publicly documented.

Sustainability models for government-backed identity: DIDI's end after its 2022 pilot highlights that ongoing operational costs are frequently underestimated. What does a sustainable financing model look like for identity infrastructure in the public sector, beyond the initial grant or multilateral funding phase?

Which government AI deployments could take a credential layer: public AI assistants are being deployed across the region, and none has been assessed for whether the identity infrastructure documented here could attach to it. A survey of those initiatives, mapping which are crossing into transactional decisions and which run on stacks that could accept verifiable credentials, is the scoping work the next phase of this research requires.

Proof of Non-Personhood: An Emerging Frontier

Every case examined so far works on one side of a single problem: establishing, cryptographically, that an interaction comes from a human being. The inverse problem is called proof of non-personhood, and it consists of proving that a counterparty is not a person and is authorised to act for one. It has a far smaller literature and, in Latin America, no deployments at all.

The problem arrives with agents that act using their principal's permissions. An agent files a tax return, books an appointment, disputes a charge or negotiates with another agent, and the service receiving that request has no way to establish whether a person authorised it. Platforms facing this have two options and both are bad: block automated traffic and break legitimate uses, or accept it and be flooded by software nobody vouches for. INATBA's mapping of AI and blockchain convergences ranks this as the frontier application of the field, cluster #10 of its analysis, and describes the mechanism others have proposed for it: a person grants, and can later revoke, an attestation stating that its holder is not a person but is entitled to act on that person's behalf, which any website or peer agent can check before deciding whether to deal with it[110].

The Inter-American Development Bank (IDB) has already made the case for having this conversation. Its analysis of agentic government argues that once an agent acts rather than an official, "responsibility, auditability, and escalation become more complex," and that governance in that setting "cannot remain only a policy statement. It must also become operational, technical, and testable in practice."[111]

Two bodies of work are being built to make it so, and they address different links of the same chain. The first extends the authorization protocols the web already runs on so that a person can delegate bounded authority to an agent. The second establishes, with decentralized identifiers and credentials, which agent is acting. This section takes them in that order, and reaches a narrow conclusion: authority is not unaddressed, since scope, expiry, revocation and transaction evidence each have at least one specified answer, but no mature and widely adopted framework yet connects the full chain from a principal's authorization through bounded execution to evidence, responsibility and remedy, and none does it on the decentralized credentials the region already issues.

None of this work originates in Latin America, and much of it overlaps with standards the region already operates, which is what makes it worth bringing into the conversation the IDB is calling for.

Delegated authority to date

The question of what an agent is allowed to do, on whose behalf and until when, has a body of work behind it that predates the agent-identity groups. It runs on federated identity: the same OAuth 2.0 and OpenID Connect flows behind every "log in with" button, in which an identity provider issues a short-lived token and the service receiving a request checks that token against the provider.

The reference proposal is from a group of researchers at MIT, Stanford and Oxford, who argue that agents need authenticated delegation and propose extending OAuth 2.0 and OpenID Connect so that a token carries a verifiable chain from the human principal to the agent, with the agent's identity, the user it acts for and the permissions granted for the specific action[112]. They add a mechanism for translating permissions stated in natural language into auditable access scopes. A derived Internet-Draft at the IETF specified the flow: a delegated access token carrying an `act` claim that names the agent alongside the user and the client, short expiry, and revocation triggered by the user, by the agent's own token or by account disablement[113]. The draft expired without adoption by a working group, and no production deployment implements it. On design it covers more of the chain than anything else in this section: principal authorization, bounded scope and expiry, and revocation.

The Agent Payments Protocol (AP2), announced by Google in September 2025 with more than sixty launch partners including Mastercard, PayPal, American Express and Coinbase, applies the same logic to purchases and supplies the piece the delegation research leaves open[114]. Each transaction is represented as three signed credentials: an Intent Mandate stating what the user authorised, a Cart Mandate binding the agent's selection to a specific item and price, and a Payment Mandate recording what was charged, each hashed to the previous one. The result is

what Google describes as a non-repudiable audit trail for every transaction, and the mandates are W3C Verifiable Credentials, the same credential format the identity programmes documented in this report issue, though AP2 does not use decentralized identifiers or the OpenID4VC issuance and presentation flows those programmes run[115]. In April 2026 Google contributed AP2 to the FIDO Alliance, which formed two working groups to develop standards for agent authentication and agentic payments, taking AP2 and Mastercard's Verifiable Intent framework as starting points; neither group has published a specification of its own[116]. AP2 documents no mechanism for revoking a mandate once issued, and it defines no remedy beyond the dispute processes payment networks already operate.

The OpenID Foundation, which stewards OpenID Connect, formed a community group on identity management for AI in July 2025 and published a whitepaper in October that maps the gaps rather than filling them: delegation models that do not cover an agent creating sub-agents, single sign-on systems not designed for an agent's lifecycle, and authorization that breaks down when an agent crosses trust domains[117]. The group states that it will not draft protocols itself. Two related drafts exist, one on authorization profiles for tool access and one proposing identity claims for agents, and neither has standing in a standards process.

Taken together, this work establishes that authority is not an open question in the way the agent-identity groups sometimes present it. Scope, expiry, revocation, mandates and transaction evidence each have at least one specified answer. What none of it yet does is connect the full chain, from a principal's authorization through bounded execution to evidence of what happened, responsibility for it and a route to remedy, in one framework that is mature and widely adopted. And all of it depends on an identity provider in the loop at the moment of verification, which is the property the authors of the delegation proposal themselves identify as the trade-off: verifiable credentials, they write, "can be presented and verified in a decentralized or peer-to-peer manner without always relying on a single identity provider," and their delegation token could be "copied into" one[112]. That is the bridge to the second body of work, and to the credential infrastructure the region already runs.

The decentralized approaches

The second body of work answers a different question: which agent is acting, established without an identity provider in the loop. Cryptography is useful here because a signature can serve as an attestation: a statement that is bound to whoever issued it, that cannot be altered without detection, and that anyone can check without contacting the issuer. Every proposal in this field builds on that property in the same shape. The agent carries a proof it cannot fabricate on its own. Whoever receives the request verifies that proof before acting on it. Whoever issued the proof can revoke it, and the revocation takes effect without having to notify anyone individually.

The proposals differ on what the proof asserts, where it lives, and how much infrastructure it demands of whoever wants to use it.

Approach	What the proof asserts	Where the proof lives	Cryptographic proof used	Scope
Web Bot Auth	Which software is making the request	Inside the web request itself	Ed25519 signature over the request, carried in HTTP Message Signatures (RFC 9421)	Any web request
W3C AI Agent Protocol	Whatever an issuer certifies about the agent (DIDs and VCs)	Held by the agent, presented on request	Issuer signature over the credential; selective disclosure through an additional layer	Any context that accepts verifiable credentials
World AgentKit	That a verified unique human authorised it (proof of human)	Held by the agent, presented on request	World ID proof derived from Orb enrolment	Services that integrate World ID
ERC-8004	What track record the agent has (identity, reputation, independent validation)	A public registry anyone can query	ERC-721 token as the identifier; optionally zero-knowledge proofs or hardware attestation to validate an agent's work	Agents transacting on-chain

Web Bot Auth

Web Bot Auth establishes that a request genuinely came from the software it claims to come from. A site can confirm that a visit is Perplexity's crawler and not something impersonating it. It establishes nothing about whether a person asked Perplexity to make that visit.

The cryptography is deliberately conventional. Each agent holds an Ed25519 key pair, a widely used elliptic-curve signature scheme. Every request it makes is signed with the private key, and the signature travels inside the web request itself using HTTP Message Signatures, specified in RFC 9421, which defines how to sign selected parts of a request and carry the signature in its headers. A Signature-Agent header points to where the agent's public keys are published, and the receiving site verifies the signature against them. The specification is being developed at the Internet Engineering Task Force, the body that defines the technical standards of the internet, with Cloudflare driving it and Amazon, Akamai and OpenAI supporting it[118].

Nothing else is required. No blockchain, no transaction to pay for, and no change to infrastructure that websites do not already run. Cloudflare launched the directory in August 2025 with four agents, among them the ChatGPT agent, Block's Goose and Browserbase, and since July 2026 treats signed agents as a subset of its broader verified bots category, which also admits published IP lists and reverse DNS as identification methods[119].

The application process is open and documented, and Cloudflare accepts any valid key directory that meets its requirements. Still, a single private company operates the directory that a large share of the web consults, applies its own approval criteria, and decides what happens to traffic from agents that are not on it. Any origin can verify a signature independently without Cloudflare in the path. In practice, most do not.

W3C AI Agent Protocol

The World Wide Web Consortium, the body that standardises the web, hosts community groups developing agent identity on Decentralized Identifiers and Verifiable Credentials (DIDs and VCs), the same standards this report has documented throughout. The AI Agent Protocol Community Group states its mission as developing "open, interoperable protocols that enable AI agents to discover, identify, and collaborate efficiently across the Web," with verifiable credential-based trust among the security mechanisms in scope[120]. A second group, formed specifically around agent identity credentials, launched in April 2026 with 42 participants and works on how agents can present cryptographically verifiable credentials binding them to the organisations that control them, coordinating with the Decentralized Identity Foundation, the OpenID Foundation and the IETF[121].

The intent is that an agent should carry credentials the way a person does. An issuer certifies something about the agent, whether that is who operates it, what it is permitted to do, or on whose behalf it acts. The agent holds that credential and presents it when a service asks, and the issuer can revoke it.

The foundation is cryptographic in the same way the rest of this report has documented. A DID is an identifier bound to a key pair, so proving control means signing with the private key. A credential is verifiable because the issuer signs it, and stronger privacy properties, presenting one attribute without revealing the rest, require an additional layer on top of that signature: BBS+ signatures in QuarkID's case. What the community groups have not yet published is which of those properties an agent credential would be required to carry. The specifications are expected across 2026 and 2027. There is nothing to deploy today.

That foundation is why this path matters for Latin America more than its maturity suggests. QuarkID, Sovra and Proyecto DIDI run on DIDs and verifiable credentials, and Blerify issues in that stack alongside ISO 18013-5. An agent credential built on the same standards would run on rails these projects already operate, issued by institutions that already act as issuers, rather than requiring a separate registry on a separate network.

World AgentKit

World released AgentKit in March 2026 and expanded it in June to work with Claude Code, Codex, Cursor and other agent environments[122]. A person who has verified their uniqueness at an Orb can delegate that verification to an agent, which then carries "cryptographic proof of a unique human behind them" and presents it when a service requests it.

It is the only one of the four with a working implementation of delegation, and the reason is that World controls the full stack: the Orb that establishes uniqueness, the credential the agent carries, and the services that accept it. No committee has to agree before it ships.

The limit follows from the same fact. World ID proofs can be verified independently, since the verification contracts and the SDK are public, but nobody other than World can issue the proof of uniqueness in the first place. A government wanting to offer an equivalent guarantee from its own credential infrastructure has no path to doing so.

Two related integrations are frequently described as commercial deployments of this layer, and both need qualifying. Okta's Human Principal was announced in April 2026 as a product the company "is planning to build," and reached early-access beta in July, built by Auth0Lab and offered through a waitlist[123]. It does not appear in Okta's own product communications, which announced a separate enterprise agent suite reaching general availability in April with no mention of it[124]. Vercel's Workflow SDK includes a human-in-the-loop step, documented by Vercel in June 2026, but that step is an approval button any party can press, with no identity verification[125]. The proof-of-human variant is a separate integration, described by World rather than by Vercel[126].

ERC-8004

The fourth approach is the most specialised, and its scope is narrower than the previous three: it applies to agents that transact on-chain. ERC-8004 addresses a question the others do not attempt, which is whether an agent you have never dealt with is worth trusting with a task. It is an open standard proposed to Ethereum in August 2025 under the title Trustless Agents, and its stated purpose is to let parties "discover, choose, and interact with agents across organizational boundaries without pre-existing trust." It remains a draft[127].

It works through three public registries. An Identity Registry gives each agent an identifier that it owns and carries between platforms, issued as an ERC-721 token, so its history does not reset when it moves between services. A Reputation Registry records what previous clients reported about whether an agent did what it undertook to do. A Validation Registry covers what reputation cannot: an agent can submit its work for independent verification, either by having the task re-run with money staked on the outcome, or through a zero-knowledge proof that the computation was performed correctly.

The proof lives in those registries rather than inside the request, so any party can query an agent's history before deciding to engage. Registration requires an on-chain address and a transaction that somebody pays for, though the standard supports smart accounts and the fee can be sponsored by a third party. Entry is otherwise permissionless: nobody approves a registration, which is the opposite of the curated directory Cloudflare operates.

The specification carries institutional weight, with authors at MetaMask, the Ethereum Foundation, Google and Coinbase. The canonical registries went live on Ethereum mainnet in early 2026, now run on more than forty networks, and had registered roughly 98,000 agents by August 2026[128].

What is behind those registrations has been measured three times, with results that vary by what is counted. An independent dashboard found that 62,673 of 132,681 registrations carried a valid and resolvable address, or 47%[129]. A team led by researchers at Imperial College, Xihan Xiong, Zelin Li, Wei Wei, Qin Wang, William Knottenbelt and Zhipeng Wang, applied a stricter test, checking whether a service actually answered at that address, and found between three and fifteen of every hundred did, depending on the network. The same team examined who was writing the reputation records and found that between 59% and 91% of the accounts leaving feedback showed the pattern of a single operator controlling many accounts. Their conclusion is that the reputation registry "cannot function as a trust signal"[130].

That is what defeats the registry's purpose. Reputation is worth reading only if the parties writing it are independent of the agent being rated and actually hired it. Nothing in the protocol requires either condition: submitting feedback costs a transaction fee and requires no proof that any work was ever done.

A third study, by Rischan Mafrur and Priagung Khusumanegara, asked how many agents were complete enough to be worth hiring. Of 10,000 registered on Ethereum, 67 exposed a working service and 628 had received any feedback at all. Nineteen met every condition at once: metadata, a live service, client feedback and cross-chain registration. Ten wallets accounted for 51% of the registrations in that sample. Their phrase for the pattern is the most useful summary available: adoption is "registration-heavy but operationally shallow"[131].

Nothing in the protocol requires demonstrating that a human stands behind a registration. The registrant is an address, which can belong to a person, a company or another agent, and agents can register agents. The openness that lets the registry scale without a gatekeeper is the same property that fills it with entries pointing nowhere and ratings written by their subjects.

Set against the delegation work, the four approaches above supply the link that work lacks, an agent identity that verifies without calling back to a provider, and lack the links it supplies. Web Bot Auth identifies software, not a principal. The W3C groups could express scope, expiry and revocation as credential attributes, and have not yet specified which. World AgentKit proves that a unique human stands behind an agent without bounding what the agent may do. ERC-8004 records what an agent did without establishing who authorised it. Eight initiatives, and none carries the chain from authorization to remedy on its own.

The regional gap

Agentic government is already on the regional agenda. The IDB analysis quoted above frames AI as moving from an interface layer, where officials use it as a tool, to an operational layer through which government functions get executed, and calls for governance that is operational, technical and testable. Brazil's data protection authority is developing technical studies on AI agents as an input to its regulatory agenda[132].

The gap sits between that diagnosis and the instruments being considered to close it. The only standard the IDB analysis names as central to the shift is the Model Context Protocol, which connects agents to tools and data and carries no identity layer at all. Verifiable credentials, agent delegation and cryptographic accountability do not appear.

There are a couple of regional efforts in this space. The first is TrustLayer Foundation A.C., a nonprofit constituted in Mexico City in March 2026, which has published ARIA, an agent registry that issues W3C DIDs to agents with credentials under the Verifiable Credentials Data Model 2.0 and post-quantum signatures. Its work informed the W3C community group on agent identity described above, which a Mexican engineer chairs. The specification is designed and the registry is live, but it has no partner organisations, documented deployments or production usage so far.

The second is a venue rather than a project. ISO/IEC JTC 1/SC 42, the joint ISO and IEC committee for artificial intelligence, standardises the governance and management of AI, covering management systems, risk, terminology and trustworthiness, and could in time host work on agent identity and credentials standards, though it carries none today. Its relevance for the region is that the region already sits in it. Seven Latin American and Caribbean countries, Brazil, Colombia, the Dominican Republic, Ecuador, Peru, Uruguay and the Bahamas, hold voting membership through national mirror committees that consolidate and carry a country position, and five more, Argentina, Bolivia, Chile, Mexico and Saint Lucia, observe. The most recent addition is the Dominican Republic, whose national AI standards committee, constituted in March 2026, includes the government's digital agency[133].

Neither shows a public identity operator working on agent identity, but both show that the region participates in standards processes. This study did not identify any public operator of the credential systems examined here taking part in the venues and projects that could supply the standards this agent-identity work needs in the region.

Recommendations

The maturity of this field does not support a recommendation to deploy. It does support three, all of which cost staff time rather than infrastructure budget.

Put the region's identity operators in the standards rooms. Agent identity is being specified now in the W3C community groups and the IETF group described above, and delegated authority in the FIDO Alliance working groups and the OpenID Foundation, all at the stage where design decisions are made and before defaults harden. A Mexican foundation is in the first room. The government identity programmes documented in this report are in neither, and they are the participants whose deployment experience would change what gets specified. The national mirror committees through which seven countries already vote in ISO's AI committee show that the region knows how to organise that participation; what is missing is the operators in the rooms where credentials for agents are being written.

Establish whether the region's existing credential architecture can express agent delegation. Actors in the region have named agent delegation as a direction, in interviews for this research and in product announcements, but none has published a design and none treats it as a current priority. The question is whether the credentials these projects already issue can express a statement of the form "this agent acts for me, within this scope, until this date", and produces evidence of what it did. The delegation token proposed by South et al. and the mandates AP2 already runs in production are the reference designs for what such a credential would need to carry; the former's authors note it could be expressed as a verifiable credential. Answering it requires technical analysis rather than deployment, it is not difficult, and the answer determines whether the region reaches the agent layer through infrastructure it already runs. Priorities of this kind change quickly once a first implementation exists.

Translate the problem for the people writing the rules. Regulators in the region are beginning to study AI agents, and the vocabulary of this field is unintelligible outside a small circle. Without that translation, rules get written with disclosure obligations as the only available instrument and cryptographic verification never enters the discussion as an option.

Closing

This report began with a broad question: whether the cryptographic protocols that came out of Ethereum could mitigate AI risks in Latin America. Answering it required narrowing twice. Four risk taxonomies produced nine subdomains salient to the region, and of those the report examines one, the compromise of privacy through the obtaining, leaking or inference of sensitive information. That was a deliberate choice and not what the ranking would have dictated, since harms to labour and technological sovereignty both score higher in the regional expert survey. Privacy is where Latin America has deployments to examine rather than proposals. Within it, the instruments that do the work are decentralized identifiers, verifiable credentials and zero-knowledge proofs, and that is where the region's projects have concentrated.

On that narrowed ground the technology proved the most straightforward part. Five programmes across Argentina, Mexico, Uruguay, Panama and El Salvador have built credentials that citizens hold, prove attributes without revealing them, and verify without calling back to the issuer: three are live, one has a documented government pilot and two national systems waiting on legal instruments, and one concluded as a pilot without continuation. The AI-risk case is being made, too, though almost entirely by the companies building this infrastructure rather than by the institutions running it: Blerify opens government conversations with deepfakes defeating biometric verification, Sovra advertises an identity layer for government AI agents, and World has taken proof of human into production with commercial partners, under the regulatory record described in this report. What sits between that argument and a public deployment is everything this report found instead of cryptography. Decrees that have not been published, legislative reforms that have not passed, officials who cannot evaluate a proof system well enough to write rules about it, and populations for whom a wallet is a harder ask than a form. Closing it is education and advocacy work: making the technology legible to the people who write the rules, keeping it open source and on international standards so that no institution is buying a dependency, and building the coalitions that carry a deployment past its launch.

That work has a clock on it. AI capability and the harms that come with it are compounding faster than the institutions meant to contain them, which makes the intersection of cryptography and AI worth funding now rather than

at the next cycle. Private deployments are the straightforward path and already have paying customers in fraud-exposed sectors. Public ones are what give this infrastructure scale and legitimacy, and the most concrete opportunity is in front of the region already: government AI assistants are moving from answering questions to completing procedures, and a decision that carries real-world consequences needs a credential that says which agent acted, for whom and within what limits, and an audit trail a citizen can check. Authority of that kind is not unaddressed: authorization protocols already specify scope, expiry, revocation and transaction evidence, on federated identity. What no mature and widely adopted framework yet does is connect the full chain, from a principal's authorization through bounded execution to evidence, responsibility and remedy, and none does it on the decentralized credentials the region issues. The prior question, whether those credentials can express that delegation, is cheaper to research than to build. The same applies at the frontier this report closes on, where proof of non-personhood is being specified in four standards bodies with no public operator of the credential systems examined here in the room, and where the mechanisms described here are the ones that would carry it.

Appendix: Interviewees

22 Semi-structured interviews conducted between November 2025 and March 2026. Titles and affiliations are as of the interview date.

Name	Organization	Title
Eskender Abebe	Ethereum Name Service	Head of Product & Strategy
Marcos Allende López	Blerify	Founder & CEO
Gurgen Arakelov	FairMath	CEO
Juan Diego Arresti	Independent	Freelancer
Niklas Azinger	Infinita City	Founder & CEO
Edgar Barrientos	Optimism DAO	Collaborator
Lior Bondereski	Fhenix	Core Protocol Developer
Lorena Buzón	Tools for Humanity	Head of Policy, Spanish-speaking LATAM
Chuy Cepeda	Sovra	Co-Founder & CSO
Dominik Clemente	World Foundation	Research Engineer
Rodolfo Corona	UC Berkeley	PhD Student
Gabriel Delgado	Próspera	Co-Founder & CDO
Sophia Dew	Celo Foundation	Dev Rel Engineering Lead
Alex Ditzen	Sociedad Argentina de Inteligencia Artificial	President
Jean García Periche	Center for Public Intelligence	Director
Arturo Grande	DESAFIA	Co-Founder
Óscar Jiménez	Independent	Freelancer
Andrés Márquez-Lara	Bittensor	Ambassador
Juan Molina	Scroll	Not stated
Bálint Pataki	Center for Future Generations	Senior Advanced AI Researcher
Ajay Patel	Tools for Humanity	Head of World ID
Pedro Perrotta	THXU Lab	Co-Founder & CEO

Notes

1. Vargas, F. and Muente, A., “Artificial Intelligence Framework for the Inter-American Development Bank Group” (2025). https://publications.iadb.org/en/artificial-intelligence-framework-inter-american-development-group ↩
2. United Nations Secretary-General’s High-level Advisory Body on Artificial Intelligence, “Governing AI for Humanity: Final Report” (September 2024). https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf ↩
3. Slattery, P., Saeri, A., Grundy, E., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S. and Thompson, N., “The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence” (2024). https://arxiv.org/abs/2408.12622 ↩
4. Global Blockchain Business Council, “AI & Blockchain Convergence,” Global Standards Mapping Initiative 5.0 (December 2024). https://www.gbbc.io/uploads/reports/gsmi50/AI-%26-Blockchain-Stand-Alone.pdf ↩
5. INATBA, “AI & Blockchain Convergences” (July 2024). https://inatba.org/wp-content/uploads/2024/07/AI-BC-Report-2.pdf ↩
6. Department for Science, Innovation and Technology and AI Security Institute, “Tackling AI security risks to unleash growth and deliver Plan for Change” (14 February 2025). https://www.gov.uk/government/news/tackling-ai-security-risks-to-unleash-growth-and-deliver-plan-for-change ↩
7. Axios, “AI safety rebrands as AI security” (28 February 2025). https://www.axios.com/2025/02/28/ai-safety-rebrand-security-issues ↩
8. Ministerio de Seguridad de la República Argentina, “Resolución 710/2024” (26 July 2024), Boletín Oficial. https://www.boletinoficial.gob.ar/detalleAviso/primera/311381/20240729 ↩
9. El País, “Javier Milei’s government will monitor social media with AI to predict future crimes” (30 July 2024). https://english.elpais.com/international/2024-07-30/javier-mileis-government-will-monitor-social-media-with-ai-to-predict-future-crimes.html ↩
10. Lin, Z., Sun, H. and Shroff, N., “AI Safety vs. AI Security: Demystifying the Distinction and Boundaries” (2025). https://arxiv.org/abs/2506.18932 ↩
11. Slattery, P., Saeri, A., Grundy, E., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S. and Thompson, N., “The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence” (2024). https://arxiv.org/abs/2408.12622 ↩
12. United Nations Secretary-General’s High-level Advisory Body on Artificial Intelligence, “Governing AI for Humanity: Final Report” (September 2024). https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf ↩
13. United Nations Executive Office of the Secretary-General, “United Nations Global Risk Report 2024” (2024). https://unglobalriskreport.org/UNHQ-GlobalRiskReport-WEB-FIN.pdf ↩
14. Muñoz, V., “Operation regulation: Strengthening Latin America’s AI governance,” European Council on Foreign Relations (1 July 2024). https://ecfr.eu/article/operation-regulation-strengthening-latin-americas-ai-governance/ ↩
15. Becker Castellaro, S., Carvalho, M., Fernandez Gibaja, A., Grassi, A., Hammar, C., Muller, J., Pereira, L., Piaia, V. and Ruediger, M. A., “Artificial Intelligence and Information Integrity: Latin American experiences,” International IDEA Policy Paper No. 34 (2025). https://www.idea.int/publications/catalogue/html/artificial-intelligence-and-information-integrity-latin-american ↩
16. INATBA, “AI & Blockchain Convergences” (July 2024). https://inatba.org/wp-content/uploads/2024/07/AI-BC-Report-2.pdf ↩
17. Global Blockchain Business Council, “AI & Blockchain Convergence,” Global Standards Mapping Initiative 5.0 (December 2024). https://www.gbbc.io/uploads/reports/gsmi50/AI-%26-Blockchain-Stand-Alone.pdf ↩
18. Ho, C., “AI And Blockchain Can Mitigate Fraud Risk Caused By Deepfakes,” Forbes (6 July 2024). https://www.forbes.com/sites/digital-assets/2024/07/06/ai-and-blockchain-synergies-mitigate-risk-of-deepfakes-in-kyc/ ↩
19. “Blockchain Technology for Combating Deepfake and Protect Video/Image Integrity” (2021), via ResearchGate. https://www.researchgate.net/publication/357241420_Blockchain_Technology_for_Combating_Deepfake_and_Protect_VideoImage_Integrity ↩
20. Global Blockchain Business Council, “AI & Blockchain Convergence,” Global Standards Mapping Initiative 5.0 (December 2024). https://www.gbbc.io/uploads/reports/gsmi50/AI-%26-Blockchain-Stand-Alone.pdf ↩
21. INATBA, “AI & Blockchain Convergences” (July 2024). https://inatba.org/wp-content/uploads/2024/07/AI-BC-Report-2.pdf ↩
22. INATBA, “AI & Blockchain Convergence Use Cases,” public dataset accompanying the 2024 report. https://docs.google.com/spreadsheets/d/1xwgXonkasFYuN362P4NjxTs0KwySoHWUvINj2h3JYVc/edit?gid=0#gid=0 ↩
23. Akther, A., Arobee, A., Adnan, A. A., Auyon, O., Islam, A. S. M. J. and Akter, F., “Blockchain As a Platform For Artificial Intelligence (AI) Transparency” (March 2025). https://arxiv.org/abs/2503.08699 ↩↩
24. UNJSPF and UNICC, “Transforming Public Digital Identity: A Blockchain Case in Action from the UN System” (18 September 2025). https://www.unjspf.org/wp-content/uploads/2025/09/UNICCUNJSPF_Transforming-Public-Digital-Identity.pdf ↩
25. Adejumo, O., “Can Ethereum secure a nation’s identity? Bhutan is betting on it,” CryptoSlate (14 October 2025). https://cryptoslate.com/can-ethereum-secure-a-nations-identity-bhutan-is-betting-on-it/ ↩
26. Prachi K., “The Aadhaar Data Breach: Truth Behind The Story,” Hackers Interview Media. https://hackersinterview.com/global-news/aadhaar-data-breach-truth-behind-story/ ↩

27. Flores, I., “Bhutan Migrates National Digital Identity to Ethereum in Major Blockchain Milestone,” CryptoNinjas (15 October 2025). https://www.cryptoninjas.net/news/bhutan-migrates-national-digital-identity-to-ethereum-in-major-blockchain-milestone/ ↩
28. Adejumo, O., “Can Ethereum secure a nation’s identity? Bhutan is betting on it,” CryptoSlate (14 October 2025). https://cryptoslate.com/can-ethereum-secure-a-nations-identity-bhutan-is-betting-on-it/ ↩
29. Proyecto DIDI, “Identidad digital auto-soberana,” ai·di product page. https://didi.org.ar/aidi-identidad-digital-auto-soberana/ ↩
30. Ledger Insights, “IDB’s DIDI unveils blockchain digital identity projects in Argentina” (2 August 2021). https://www.ledgerinsights.com/idb-didi-unveils-blockchain-digital-identity-projects-in-argentina/ ↩
31. ACDI, Fundación ProNorte and Proyecto DIDI, “Proyecto de identidad digital auto-soberana sobre blockchain para la calificación de riesgo climático en la región del Gran Chaco argentino” (2022). https://didi.org.ar/informes-del-proyecto/ ↩
32. Mazzocca, C., Acar, A., Uluagac, S., Montanari, R., Bellavista, P. and Conti, M., “A Survey on Decentralized Identifiers and Verifiable Credentials,” IEEE Communications Surveys & Tutorials (2025). https://arxiv.org/abs/2402.02455 ↩
33. ONG Bitcoin Argentina and IDB Lab, “Proyecto DIDI: Aprendizajes, resultados y desafíos de la identidad digital auto-gestionada para la inclusión” (2023). https://didi.org.ar/informes-del-proyecto/ ↩
34. Carreras, T., “Buenos Aires Adds ZK Proofs to City App in Bid to Boost Residents’ Privacy,” CoinDesk (22 October 2024). https://www.coindesk.com/tech/2024/10/22/buenos-aires-adds-zk-proofs-to-city-app-in-bid-to-boost-residents-privacy ↩
35. Matter Labs, post on the QuarkID deployment, ZKsync Mirror. https://zksync.mirror.xyz/kWRhD81C7il4YWGrkDplfhIZcmViisRe3lnsmbvOEmg ↩
36. Matter Labs, “QuarkID DID Creation”, Dune query 4097821, last run 8 September 2026, consulted 9 September 2026. https://dune.com/queries/4097821. The query returns 1,000 rows and its series ends on 22 July 2026 at 1,817,638 DIDs; the count reads creation events on the QuarkID contract on zkSync Era. The same figure appears on Matter Labs’ QuarkID dashboard, https://dune.com/matter_labs/quarkid/d625463e-2815-468a-8227-5d3d90dba7dc, which requires a Dune login. ↩
37. Business Standard, “Aadhaar data of millions of Indians put on sale on the dark web: Reports” (30 October 2023). https://www.business-standard.com/india-news/aadhaar-data-of-millions-of-indians-put-on-sale-on-the-dark-web-reports-123103000993_1.html ↩
38. Cimpanu, C., “Hacker steals government ID database for Argentina’s entire population,” The Record (18 October 2021). https://therecord.media/hacker-steals-government-id-database-for-argentinas-entire-population ↩
39. Sophos, “Leaky database spills data on 20 million Ecuadorians and businesses” (18 September 2019), archived copy. https://web.archive.org/web/20251215112553/https://news.sophos.com/en-us/2019/09/18/leaky-database-spills-data-on-20-million-ecuadorians-and-businesses/ ↩
40. openDemocracy, “The largest personal data leakage in Brazilian history.” https://www.opendemocracy.net/en/largest-personal-data-leakage-brazilian-history/ ↩
41. Borak, M., “Buenos Aires moves from centralized to decentralized digital identity with QuarkID,” Biometric Update (23 October 2024). https://www.biometricupdate.com/202410/buenos-aires-moves-from-centralized-to-decentralized-digital-identity-with-quarkid ↩
42. Searches for a QuarkID 2.0 specification, roadmap, release or governance document returned no results in Spanish or English as of August 2026. The protocol documentation site, docs.quarkid.org, did not resolve. ↩
43. Sovra GitHub organisation, QuarkID node repositories migrated January 2026. https://github.com/sovrahq ↩
44. Extrimian GitHub organisation: https://github.com/extrimian/Paquetes-NPMjs-BackHome (repository described as “Back at Extrimian, evolving independently from QuarkID,” October 2025); https://github.com/extrimian/iota-identity (January 2026). Canonical protocol organisation: https://github.com/ssi-quarkid ↩
45. Gobierno de la Ciudad Autónoma de Buenos Aires, Secretaría de Innovación y Transformación Digital, “QuarkID,” official programme page. https://buenosaires.gob.ar/jefaturadegabinete/innovacionytransformaciondigital/quarkid ↩
46. Reporte Índigo, “Nuevo León comparte experiencia de uso con Blockchain ante foro internacional en EU” (6 March 2025). https://www.reporteindigo.com/monterrey/Nuevo-Leon-comparte-experiencia-de-uso-con-Blockchain-ante-foro-internacional-en-EU-20250306-0038.html ↩
47. Gobierno del Estado de Nuevo León, “Expone Nuevo León experiencia en seguridad digital y blockchain de plataforma NLínea en ETH Denver 2025” (5 March 2025). https://www.nl.gob.mx/es/boletines/expone-nuevo-leon-experiencia-en-seguridad-digital-y-blockchain-de-plataforma-nlinea-en ↩
48. Gobierno de Salta, programa IDDI. https://iddi.salta.gob.ar/ ↩
49. Reporte Índigo, “Nuevo León comparte experiencia de uso con Blockchain ante foro internacional en EU” (6 March 2025). https://www.reporteindigo.com/monterrey/Nuevo-Leon-comparte-experiencia-de-uso-con-Blockchain-ante-foro-internacional-en-EU-20250306-0038.html ↩
50. Gobierno del Estado de Nuevo León, “Expone Nuevo León experiencia en seguridad digital y blockchain de plataforma NLínea en ETH Denver 2025” (5 March 2025). https://www.nl.gob.mx/es/boletines/expone-nuevo-leon-experiencia-en-seguridad-digital-y-blockchain-de-plataforma-nlinea-en ↩
51. Sovra, case studies page. https://sovra.io/case-studies ↩
52. Sovra GitHub organisation, QuarkID node repositories migrated January 2026. https://github.com/sovrahq ↩

53. Sovra, SovraChain product page. https://sovra.io/sovrachain ↩
54. Aligned Layer, “Introducing Aligned’s Rollup-as-a-Service Platform” (18 June 2025). https://blog.alignedlayer.com/introducing-aligned-raas/ ↩
55. Sovra, product pages: https://sovra.io/ ; https://sovra.io/sovraid ; https://sovra.io/sovrarag ↩
56. Fernández, D., Cepeda, J., Garza, A., Jolias, L., Carrone, F., Onorato, M., Catalán, R. and Kingston, D., “Propia: A Public Trust Layer for Verifiable Digital Credentials. Protocol Whitepaper” (13 August 2026), a collaboration between Sovra, Lambda and Aligned. https://propia.id/whitepaper/ ↩
57. Propia was published in the final week of this research period. At the close of research there was no registry in production, no adopting government and no governance body constituted. The specification is described here and not assessed. ↩
58. Sovra, SovraID API guide. https://github.com/sovrahq/id-docs/blob/main/docs/guides/api-sovra-id.md ↩
59. Sovra’s Knowledge section contains no reference to post-quantum cryptography, ISO 18013-5, mDL, AnonCreds, BBS+ or Trust over IP, and does not address the agentic layer announced on the company’s home page. https://sovra.io/knowledge ↩
60. Sovra’s Knowledge section contains no reference to post-quantum cryptography, ISO 18013-5, mDL, AnonCreds, BBS+ or Trust over IP, and does not address the agentic layer announced on the company’s home page. https://sovra.io/knowledge ↩
61. AGESIC (Presidencia de Uruguay) and Blerify, “Caso de estudio AGESIC Uruguay: ID Digital ISO/IEC 18013-5, ID Wallet y contratos inteligentes para lista de confianza regional descentralizada” (2026). https://blerify.com/research-agesic.html ↩
62. Blerify, “DID Method did:lac1,” developer documentation. https://docs.blerify.com/trust-registry/did-method/did-method-did-lac1 The method is presented as compliant with the W3C DID Core specification and as building on the ethr DID method and the LACChain DID method, adding backwards revocation time support, direct DID registry resolution and DID migration through the alsoKnownAs attribute. ↩
63. AGESIC (Presidencia de Uruguay) and Blerify, “Caso de estudio AGESIC Uruguay: ID Digital ISO/IEC 18013-5, ID Wallet y contratos inteligentes para lista de confianza regional descentralizada” (2026). https://blerify.com/research-agesic.html ↩
64. AGESIC (Presidencia de Uruguay) and Blerify, “Caso de estudio AGESIC Uruguay: ID Digital ISO/IEC 18013-5, ID Wallet y contratos inteligentes para lista de confianza regional descentralizada” (2026). https://blerify.com/research-agesic.html ↩
65. AGESIC (Presidencia de Uruguay) and Blerify, “Caso de estudio AGESIC Uruguay: ID Digital ISO/IEC 18013-5, ID Wallet y contratos inteligentes para lista de confianza regional descentralizada” (2026). https://blerify.com/research-agesic.html ↩
66. AGESIC (Presidencia de Uruguay) and Blerify, “Caso de estudio AGESIC Uruguay: ID Digital ISO/IEC 18013-5, ID Wallet y contratos inteligentes para lista de confianza regional descentralizada” (2026). https://blerify.com/research-agesic.html ↩
67. AGESIC (Presidencia de Uruguay) and Blerify, “Caso de estudio AGESIC Uruguay: ID Digital ISO/IEC 18013-5, ID Wallet y contratos inteligentes para lista de confianza regional descentralizada” (2026). https://blerify.com/research-agesic.html ↩
68. AGESIC (Presidencia de Uruguay) and Blerify, “Caso de estudio AGESIC Uruguay: ID Digital ISO/IEC 18013-5, ID Wallet y contratos inteligentes para lista de confianza regional descentralizada” (2026). https://blerify.com/research-agesic.html ↩
69. AGESIC (Presidencia de Uruguay) and Blerify, “Caso de estudio AGESIC Uruguay: ID Digital ISO/IEC 18013-5, ID Wallet y contratos inteligentes para lista de confianza regional descentralizada” (2026). https://blerify.com/research-agesic.html ↩
70. AGESIC (Presidencia de Uruguay) and Blerify, “Caso de estudio AGESIC Uruguay: ID Digital ISO/IEC 18013-5, ID Wallet y contratos inteligentes para lista de confianza regional descentralizada” (2026). https://blerify.com/research-agesic.html ↩
71. Blerify, “Panamá Conecta · Sertracen” case study. https://blerify.com/caso.html?c=sertracen ↩
72. Telemetro, “Licencia digital Panamá: Sertracen habilita opción para gestionar el trámite” (May 2026). https://www.telemetro.com/nacionales/licencia-digital-panama-sertracen-habilita-opcion-gestionar-el-documento-movil-n6079483 ↩
73. La Estrella de Panamá, “AIG: licencia digital en Panamá está lista, pero falta el freno legal” (2 August 2026). https://www.laestrella.com.pa/panama/nacional/aig-licencia-digital-en-panama-esta-lista-pero-falta-el-freno-legal-PF24566845; La Prensa, “Esto es lo que está retrasando la licencia de conducir digital en Panamá” (15 May 2026). https://www.prensa.com/sociedad/esto-es-lo-que-esta-retrasando-la-licencia-de-conducir-digital-en-panama/ ↩
74. La Estrella de Panamá, “AIG: licencia digital en Panamá está lista, pero falta el freno legal” (2 August 2026). https://www.laestrella.com.pa/panama/nacional/aig-licencia-digital-en-panama-esta-lista-pero-falta-el-freno-legal-PF24566845; La Prensa, “Esto es lo que está retrasando la licencia de conducir digital en Panamá” (15 May 2026). https://www.prensa.com/sociedad/esto-es-lo-que-esta-retrasando-la-licencia-de-conducir-digital-en-panama/ ↩
75. Diario El Mundo, “Uso del DUI digital requerirá reformas legislativas para su reconocimiento” (July 2026). https://diario.elmundo.sv/politica/uso-del-dui-digital-requerira-reformas-legislativas-para-su-reconocimiento-en-todos-los-tramites ↩
76. Blerify, “Casos de éxito.” https://blerify.com/historias.html ↩
77. Blerify, “Casos de éxito.” https://blerify.com/historias.html ↩
78. RENIEC, “Lineamientos para la emisión y uso del Documento Nacional de Identidad Digital (DNId),” Dirección de Certificación y Servicios Digitales, first version (published 31 July 2026). The guidelines specify the W3C Verifiable Credentials Data Model 1.1 and 2.0, W3C Decentralized Identifiers 1.0, and OpenID4VCI for issuance. https://www.gob.pe/institucion/reniec/informes-publicaciones/8432059-lineamientos-para-la-emision-y-uso-del-documento-nacional-de-identidad-digital-dnid ↩

79. Contreras García, V., "World impulsará la autenticación humana en startups de LATAM," DPL News (25 June 2025). https://dplnews.com/world-impulsara-autenticacion-humana-startups-latam/ The figure is attributed to Tools for Humanity as an institution at the Build ↩

80. Country figures as stated by Tools for Humanity representatives. Argentina: iProUP (July 2025). https://www.iproup.com/innovacion/58360-world-alcanza-los-14-millones-de-humanos-verificados-a-2-anos-de-su-lanzamiento Colombia: Área Cúcuta (July 2025). https://www.areacucuta.com/world-alcanza-14-millones-de-humanos-verificados-y-se-consolida-como-la-red-mas-grande-del-planeta/ Mexico, per Miguel Rocha, regional director for Mexico and Central America: BeInCrypto en Español (July 2025). https://es.beincrypto.com/world-cuenta-13-millones-usuarios-mexico-18-millones-colombia/ Peru, per Javier Turián, LATAM communications: Gestión (25 March 2025). https://gestion.pe/tu-dinero/identidad-digital-un-millon-de-peruanos-la-acreditan-cuanto-cobran-hay-riesgo-noticia/ Brazil: Mobile Time (5 December 2025). https://www.mobiletime.com.br/noticias/05/12/2025/tools-for-humanity-iris/ Each source distinguishes World App accounts from Orb-verified humans; the two differ by roughly a factor of two in every country reported. ↩

81. Least Authority, "Worldcoin Protocol Cryptography Final Audit Report" (July 2023). https://leastauthority.com/wp-content/uploads/2023/07/Worldcoin_Protocol_Cryptography_Final_Audit_Report.pdf ↩

82. World, "World ID Contracts Reference", https://docs.world.org/reference/contracts ↩

83. World, "The new World ID and the partners bringing proof of human to the internet" (17 April 2026). https://world.org/es-es/blog/announcements/the-new-world-id-and-the-partners-bringing-proof-of-human-to-the-internet ↩

84. Adler, S. et al., "Personhood credentials: Artificial intelligence and the value of privacy-preserving tools to distinguish who is real online" (2024). https://arxiv.org/abs/2408.07892 ↩

85. Ho, C., "AI And Blockchain Can Mitigate Fraud Risk Caused By Deepfakes," Forbes (6 July 2024). https://www.forbes.com/sites/digital-assets/2024/07/06/ai-and-blockchain-synergies-mitigate-risk-of-deepfakes-in-kyc/ ↩

86. World, "Introducing the World Partner Markets Program" (30 September 2025). https://world.org/blog/world/introducing-the-world-partner-markets-program ↩

87. McConvey, J. R., "World wants big retail partners to house more of its Orbs," Biometric Update (2 October 2025). https://www.biometricupdate.com/202510/world-wants-big-retail-partners-to-house-more-of-its-orbs ↩

88. World, Rappi partnership page. https://world.org/rappi ↩

89. World, "CONMEBOL Sudamericana" partnership page. https://world.org/es-la/conmebol-sudamericana ↩

90. Wiggers, K., "World partners with Tinder, Visa to bring its ID-verifying tech to more places," TechCrunch (30 April 2025). https://techcrunch.com/2025/04/30/world-partners-with-tinder-visa-to-bring-its-id-verifying-tech-to-more-places/ ↩

91. World, "How do I verify my age on Tinder using my World ID?," World support centre. https://support.world.org/hc/en-us/articles/46009614590227-How-do-I-verify-my-age-on-Tinder-using-my-World-ID ↩

92. World, "The new World ID and the partners bringing proof of human to the internet" (17 April 2026). https://world.org/es-es/blog/announcements/the-new-world-id-and-the-partners-bringing-proof-of-human-to-the-internet ↩

93. Tinder, "Tinder to Expand Facial Verification Feature Across the U.S., Setting a New Standard for Dating Safety" (22 October 2025). https://www.tinderpressroom.com/2025-10-22-Tinder-to-Expand-Facial-Verification-Feature-Across-the-U-S-,-Setting-a-New-Standard-for-Dating-Safety ↩

94. Coverage of the Lift Off announcement: Lucas Ropek, TechCrunch (17 April 2026), https://techcrunch.com/2026/04/17/sam-altmans-project-world-looks-to-scale-its-human-verification-empire-first-stop-tinder/; Kali Hays, "Tinder and Zoom offer 'proof of humanity' eye-scans to combat AI," BBC (17 April 2026), https://www.bbc.com/news/articles/cp9vppem4evo; Wired (17 April 2026), https://www.wired.com/story/gazing-into-sam-altmans-orb-now-proves-youre-human-on-tinder/; El Universal (17 April 2026), https://www.eluniversal.com.mx/techbit/world-id-lanza-app-de-identidad-humana-contra-deepfakes-y-reventa-de-boletos/. On the pause of sign-ups in Latin America, see Rest of World (27 April 2026), https://restofworld.org/2026/sam-altman-worldcoin-zoom-tinder-partnerships/ ↩

95. Salazar Herrada, E., "Blockchain y elecciones 2026: cómo se desarrollará en Perú el voto digital obligatorio para militares y ciudadanos en el extranjero," Infobae (2025). https://www.infobae.com/peru/2025/06/06/blockchain-y-elecciones-2026-como-se-desarrollara-en-peru-el-voto-digital-obligatorio-para-militares-y-ciudadanos-en-el-extranjero/ ↩

96. Cholakian Herrera, L., "Worldcoin is surging in Argentina thanks to 288% inflation," Rest of World (1 May 2024). https://restofworld.org/2024/worldcoin-argentina/ ↩

97. Guo, E. and Renaldi, A., "Deception, exploited workers, and cash handouts: How World recruited its first half a million test users," MIT Technology Review (6 April 2022). https://www.technologyreview.com/2022/04/06/1048981/worldcoin-cryptocurrency-biometrics-web3/ ↩

98. Bhattacharya, A., "U.S. companies back Sam Altman's World ID even as much of the world pushes back," Rest of World (27 April 2026). https://restofworld.org/2026/sam-altman-worldcoin-zoom-tinder-partnerships/ ↩

99. The four elements track the consent standards in force in the region. Brazil's LGPD (Lei 13.709/2018) classifies biometric data as sensitive personal data (art. 5, II), defines consent as a free, informed and unequivocal statement for a determined purpose (art. 5, XII), requires it to be specific and highlighted when the data is sensitive (art. 11, I), allows revocation at any time through a free and facilitated procedure (art. 8, §5), and grants rights of confirmation, access, correction, anonymisation, blocking and deletion (art. 18). Argentina's Ley 25.326 requires free, express and informed consent given in writing (art. 5), imposes a prior duty to inform on purpose, recipients and rights (art. 6), bars anyone from being compelled to supply sensitive data (art. 7), and grants access (art. 14) and rectification, up-

dating or deletion (art. 16). Brazil's ANPD applied this standard to World directly, ordering Tools for Humanity to stop offering cryptocurrency in exchange for iris scans on the ground that payment interferes with the free formation of consent, and rejecting the company's appeal on 11 February 2025. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm ; https://servicios.infoleg.gob.ar/infolegInternet/anexos/60000-64999/64790/texact.htm ; https://agenciabrasil.ebc.com.br/geral/noticia/2025-02/empresa-suspende-coleta-de-iris-de-brasileiros ↩

100. AIAAIC Repository, "Argentina opens investigation into Worldcoin," AI, Algorithmic and Automation Incidents and Controversies. https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/argentina-opens-investigation-into-worldcoin ↩
101. Animal Político, "México niega información sobre Worldcoin; Colombia, Filipinas y Brasil escalan medidas regulatorias" (4 December 2025). https://grupoanimal.mx/sociedad/mexico-reserva-informacion-worldcoin-medidas-regulatorias. On the dissolution of the INAI and the transfer of its data protection powers to the Secretaría Anticorrupción y Buen Gobierno, see https://www.infobae.com/mexico/2025/05/10/inai-desaparece-secretaria-anticorrupcion-y-buen-gobierno-asume-funciones/. Earlier regional coverage of the regulatory scrutiny: https://www.biometricupdate.com/202405/regulators-hound-worldcoin-digital-id-project-as-it-expands-in-latin-america ↩
102. DataGuidance, "Chile: Supreme Court rules Worldcoin violated the right to privacy." https://www.dataguidance.com/news/chile-supreme-court-rules-worldcoin-violated-right ↩
103. Electronic Privacy Information Center, "Paying for Iris Scans: AI-Fueled Surveillance Harms." https://epic.org/paying-for-iris-scans-ai-fueled-surveillance-harms/ ↩
104. World, "World in Brazil: Facts and Benefits of an Important New Technology" (27 January 2025, updated 6 August 2025). https://world.org/es-es/blog/policy/world-brazil-facts-benefits-important-new-technology ↩
105. Mobile Time, "Tools for Humanity vai expandir sua rede no Brasil em 2026" (5 December 2025), interview with Juliana Felippe, General Director of Tools for Humanity Brazil. https://www.mobiletime.com.br/noticias/05/12/2025/tools-for-humanity-iris/ ↩
106. Câmara Municipal de São Paulo, "CPI da Íris: relatório final é apresentado e aprovado pelo colegiado" (7 April 2026). https://www.saopaulo.sp.leg.br/blog/cpi-da-iris-relatorio-final-e-apresentado-e-aprovado-pelo-colegiado/ ↩
107. Hardware.com.br, "Ministério Público de SP processa Tools for Humanity e AWS em R$ 240 milhões por escaneamento de íris." https://www.hardware.com.br/noticias/ministerio-publico-sp-processa-tools-for-humanity-aws-240-milhoes-escaneamento-iris/ ↩
108. Coverage of the Lift Off announcement: Lucas Ropek, TechCrunch (17 April 2026), https://techcrunch.com/2026/04/17/sam-altmans-project-world-looks-to-scale-its-human-verification-empire-first-stop-tinder/; Kali Hays, "Tinder and Zoom offer 'proof of humanity' eye-scans to combat AI," BBC (17 April 2026), https://www.bbc.com/news/articles/cp9vppem4evo; Wired (17 April 2026), https://www.wired.com/story/gazing-into-sam-altmans-orb-now-proves-youre-human-on-tinder/; El Universal (17 April 2026), https://www.eluniversal.com.mx/techbit/world-id-lanza-app-de-identidad-humana-contra-deepfakes-y-reventa-de-boletos/. On the pause of sign-ups in Latin America, see Rest of World (27 April 2026), https://restofworld.org/2026/sam-altman-worldcoin-zoom-tinder-partnerships/ ↩
109. Xataka México (22 August 2025), reporting on the Animal Político investigation. https://www.xataka.com.mx/otros-1/700-mil-personas-mexico-dieron-su-iris-bitcoins-ahora-no-hay-quien-proteja-sus-datos-animal-politico ↩
110. INATBA, "AI & Blockchain Convergences" (July 2024). https://inatba.org/wp-content/uploads/2024/07/AI-BC-Report-2.pdf ↩
111. Lenin Fabricio Rodriguez Yanez, "From AI Assistants to AI Agents: A New Operational Layer for Digital Government," Inter-American Development Bank (22 June 2026). https://www.iadb.org/en/blog/modernization-state/public-administration/ai-assistants-ai-agents-new-operational-layer-digital-government ↩
112. South, T., Marro, S., Hardjono, T., Mahari, R., Deslandes Whitney, C., Greenwood, D., Chan, A. and Pentland, A., "Authenticated Delegation and Authorized AI Agents" (arXiv:2501.09674, January 2025); published as "Position: AI Agents Need Authenticated Delegation," ICML 2025. https://arxiv.org/abs/2501.09674. The quoted passages are in section 3.5. ↩↩
113. Senarath, T. S., "OAuth 2.0 Extension: On-Behalf-Of User Authorization for AI Agents," IETF Internet-Draft draft-oauth-ai-agents-on-behalf-of-user-00 (2 May 2025, expired). https://datatracker.ietf.org/doc/html/draft-oauth-ai-agents-on-behalf-of-user-00 ↩
114. Google Cloud, "Announcing Agent Payments Protocol (AP2)" (16 September 2025). https://cloud.google.com/blog/products/ai-machine-learning/announcing-agents-to-payments-ap2-protocol ↩
115. AP2 specification and repository. https://ap2-protocol.org/specification/; https://github.com/google-agentic-commerce/AP2. The current specification groups the Intent and Cart Mandates under a Checkout Mandate; the mechanism, a signed credential at each stage chained to the previous one, is unchanged. ↩
116. FIDO Alliance, "FIDO Alliance to Develop Standards for Trusted AI Agent Interactions" (28 April 2026). https://fidoalliance.org/fido-alliance-to-develop-standards-for-trusted-ai-agent-interactions/ ↩
117. OpenID Foundation, "New whitepaper tackles AI agent identity challenges" (7 October 2025). https://openid.net/new-whitepaper-tackles-ai-agent-identity-challenges/; Artificial Intelligence Identity Management Community Group, https://openid.net/cg/artificial-intelligence-identity-management-community-group/ ↩
118. Cloudflare, "Signed Agents: verifying AI agent identity on the open web." https://blog.cloudflare.com/signed-agents/ ; IETF Web Bot Auth draft, building on RFC 9421 HTTP Message Signatures. ↩
119. Cloudflare, "Signed Agents: verifying AI agent identity on the open web." https://blog.cloudflare.com/signed-agents/ ; IETF Web Bot Auth draft, building on RFC 9421 HTTP Message Signatures. ↩
120. W3C, "AI Agent Protocol Community Group." https://www.w3.org/groups/cg/agentprotocol/ ↩

121. W3C, “Agent Identity Registry Protocol Community Group” (proposed 22 April 2026). https://www.w3.org/community/agent-identity/ ↩

122. World, “Now available: AgentKit, proof of human for the agentic web” (17 March 2026). https://world.org/blog/announcements/now-available-agentkit-proof-of-human-for-the-agentic-web ↩

123. World, “Okta Verify: Proof of Human for the Agentic Web” (23 July 2026). https://world.org/blog/announcements/okta-verify-proof-of-human-agentic-web. Product site and beta waitlist at https://humanprincipal.ai (Auth0Lab, Okta Inc.). Both sources are the vendor and its partner describing their own product. ↩

124. Okta, “Okta announces new blueprint for the secure agentic enterprise” (16 March 2026). https://www.okta.com/newsroom/press-releases/showcase-2026/ ↩

125. Vercel, “Human-in-the-Loop with Chat SDK and Workflow SDK,” Vercel Knowledge Base (updated 17 June 2026). https://vercel.com/kb/guide/human-in-the-loop-with-chat-sdk-and-workflow-sdk ↩

126. World, “Vercel Workflows: Human in the Loop for Agents” (22 July 2026). https://world.org/blog/announcements/vercel-workflows-human-in-the-loop ↩

127. Marco De Rossi, Davide Crapis, Jordan Ellis and Erik Reppel, “ERC-8004: Trustless Agents,” Ethereum Improvement Proposals (created 13 August 2025; status Draft as of August 2026). https://eips.ethereum.org/EIPS/eip-8004. Registry deployments and author affiliations: https://github.com/erc-8004/erc-8004-contracts ↩

128. Registry deployments across networks: https://github.com/erc-8004/erc-8004-contracts. Mainnet launch coverage: CoinDesk, 28 January 2026, https://www.coindesk.com/markets/2026/01/28/ethereum-s-erc-8004-aims-to-put-identity-and-trust-behind-ai-agents. Registration counts by chain: The Defiant, https://thedefiant.io/news/defi/bnb-smart-chain-becomes-home-to-most-erc-8004-ai-agents ↩

129. Lorenz Lehmann, “EIP-8004,” growthepie (17 February 2026). https://www.growthepie.com/quick-bites/eip-8004 ↩

130. Xihan Xiong, Zelin Li, Wei Wei, Qin Wang, William Knottenbelt and Zhipeng Wang, “Can Trustless Agents Be Trusted? An Empirical Study of the ERC-8004 Decentralized AI Agent Ecosystem,” arXiv:2606.26028 (24 June 2026, revised 8 July 2026). https://arxiv.org/abs/2606.26028 ↩

131. Rischan Mafrur and Priagung Khusumanegara, “From Agent Identity to Agent Economy: Measuring the Operational Readiness of ERC-8004 AI Agents,” arXiv:2606.12128 (10 June 2026). https://arxiv.org/abs/2606.12128 ↩

132. IAPP, “Notas de la IAPP América Latina: avances en la implementación de normas sobre IA y protección de datos.” https://iapp.org/news/a/notas-de-la-iapp-america-latina-avances-en-la-implementaci-n-de-normas-sobre-ia-y-protecci-n-de-datos ↩

133. ISO, “ISO/IEC JTC 1/SC 42 Artificial intelligence: Participation.” https://www.iso.org/committee/6794475.html?view=participation (accessed September 2026). Participating members from the region: ABNT (Brazil), ICONTEC (Colombia), INDOCAL (Dominican Republic), INEN (Ecuador), INACAL (Peru), UNIT (Uruguay), BBSQ (Bahamas); observing: IRAM (Argentina), IBNORCA (Bolivia), INN (Chile), DGN (Mexico), SLBS (Saint Lucia). On the Dominican committee, which brings together INDOCAL, CODOCA, OGTIC, the ICT Chamber and ANEIH: INDOCAL, “Indocal presenta conformación de Comité Técnico de Normalización sobre Inteligencia Artificial” (30 March 2026). https://indocal.gob.do/indocal-presenta-conformacion-de-comite-tecnico-de-normalizacion-sobre-inteligencia-artificial/ ↩