SDS 2026 · 13th IEEE Swiss Conference on Data Science · 6 MAY 2026

ICEBERG
RESEARCH GROUP

The Architecture of Digital Trust

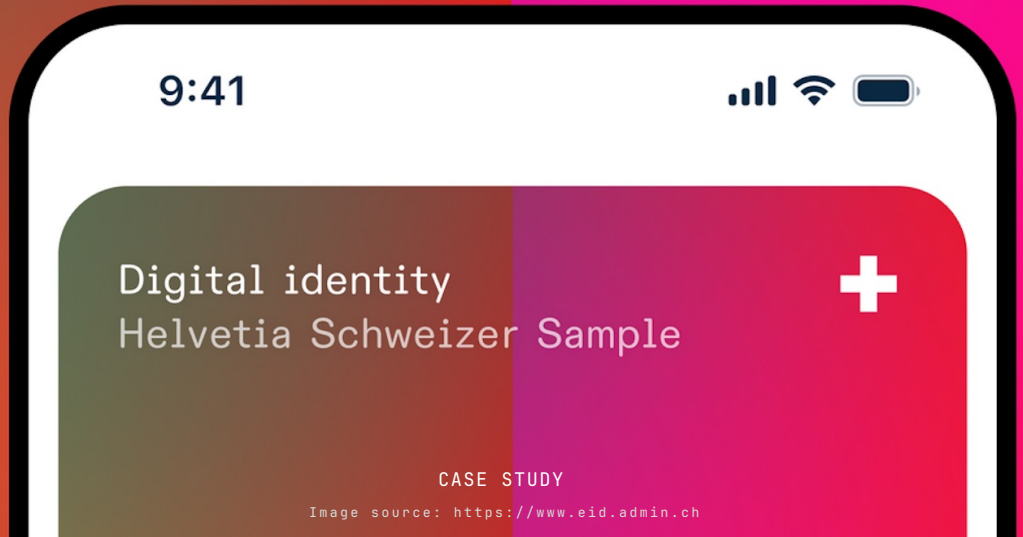
A Multi-Level Framework for Bridging the AI Value Gap.

Daniel Glinz · iceberg.digital · Zurich · ORCID 0009-0003-1706-6102

In 2021, a technically sound e-ID was rejected by 64.4%

2021:
64.4% No
35.6% Yes

2025:
49.6% No
50.4% Yes



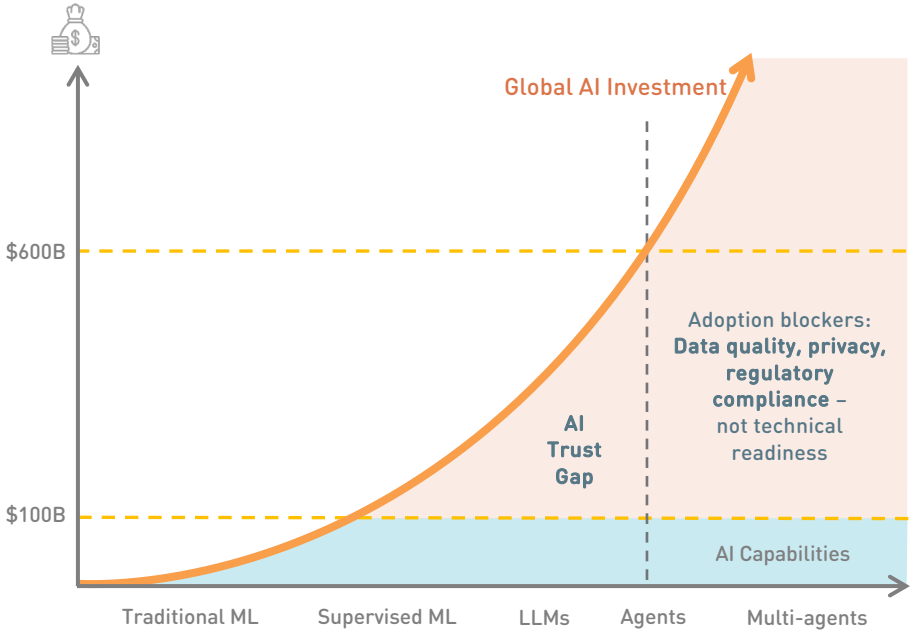
CASE STUDY

Image source: https://www.eid.admin.ch

AI investment is rising. Value realization is not

- Dominant barriers are not technical readiness or business case validation
- They are data quality, data privacy, and regulatory compliance (SFTI, 2025)

→ Whether the organization, its regulators, and its customers can justify trusting the system



Source: inspired by LatticeFlow, 2025 and Gartner, 2025

Three paradoxes characterize AI trust, and they cannot be solved technically

None of these paradoxes are bugs. They are structural features of human–AI interaction that persist regardless of model accuracy or interface polish.

→ which is why architectural response, not technical fix, is the right frame.

01

AI Authorship Effect

Identical content rated less trustworthy and less authentic the moment it's labeled AI-generated (Kirk & Givi, 2025).

02

AI Intimacy Paradox

Users disclose more because interactions feel safe, while distrusting the organizations that operate them (Turkle, 2015).

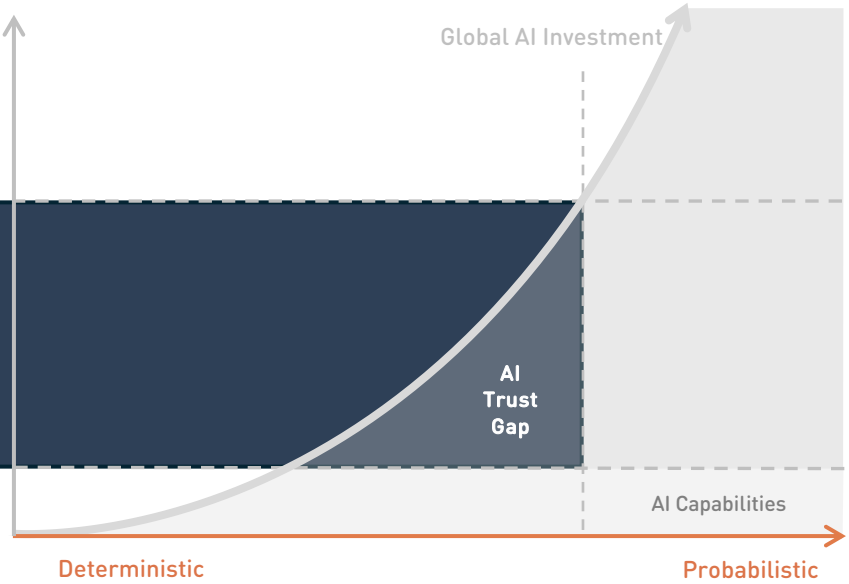
03

Agency Paradox

Users seek cognitive relief from AI, yet feel threatened when it overtakes identity-relevant tasks (Nowotny, 2021).

Architectures built for deterministic systems don't hold up against probabilistic, emergent AI

- The AI value gap is a trust deficit that current architectures are not equipped to close.
- Closing it requires an architectural approach that treats digital trust as a verifiable system property.



Grounded-Theory Design Science

A defined coding procedure sits within a defined design program

Standard Design Science Research (Hevner et al., 2004) informs artifact design through case studies, surveys, or expert interviews and inherits the perspectival limits of those data sources.

Grounded-theory literature review (Wolfswinkel, Furtmueller & Wilderom, 2013) produces categories and constructs but is not committed to a deliverable artifact.

DSR · 01

Problem

AI value gap framed as trust deficit

DSR · 02

Design

Four-layer iceberg, derived by necessity-and-sufficiency analysis against the corpus

DSR · 03

Operationalization

GT-LR nested here (Wolfswinkel 5 stages):

1. Define
2. Search
3. Select
4. Analyze
5. Present

DSR · 04

Evaluation

Four-criterion evaluation:

1. literature-based grounding
2. constant-comparison boundary checks
3. corpus-level plateau analysis
4. cases for explanatory adequacy

DSR · 05

Refinement

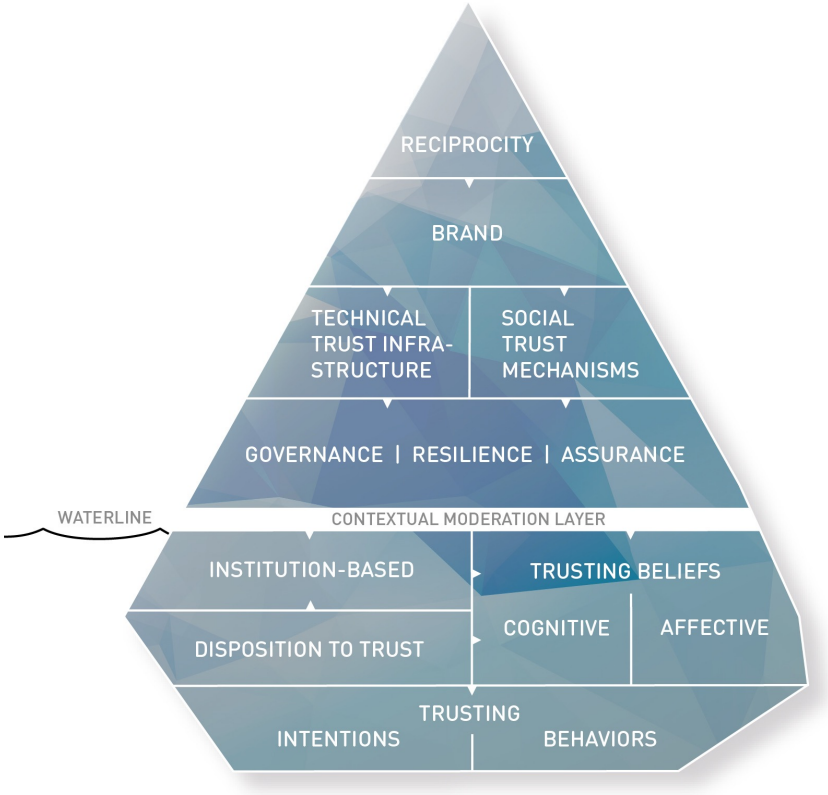
Iterative refinement of constructs & cues

The iceberg framework. Above and below the waterline

The iceberg is not decoration.

Cues an organization exposes (brand, disclosures, marketing) sit above the waterline. Trust formation is also shaped by cognitive structures below it (institution-based trust, dispositional trust, affect-based trust beliefs), which organizations cannot directly intervene on and must instead address indirectly through sustained behavioral consistency over time.

INTERDEPENDENT LAYERED EMERGENT



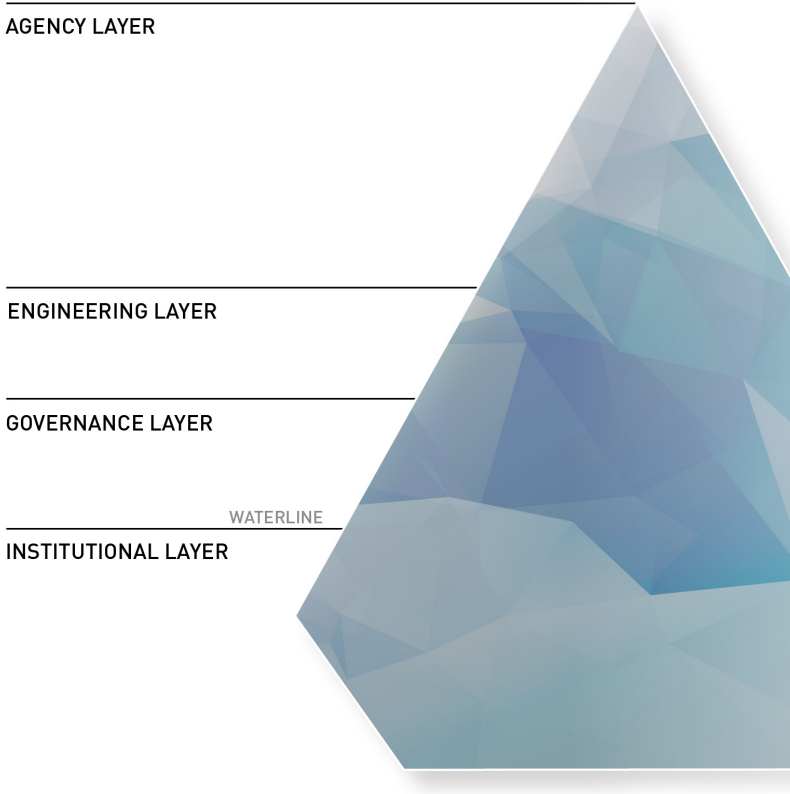
Four layers.
Each with its own work

A — Humans read digital systems for competence, sincerity, and intent.
The three paradoxes all operate here.
Trust erosion is relational, not technical.

E — Trust as a measurable, cryptographically verifiable system property.
DIDs, Verifiable Credentials, C2PA …
AI failure modes require continuous evaluation, not one-time certification.

G — Traditional GRC assumes deterministic systems. AI is probabilistic, emergent, and fast-moving. Replaced by adaptive governance, organisational resilience, and continuous digital assurance.

I — eIDAS 2.0, EU AI Act, FINMA 2024. Trust cannot rest on private good intentions alone. Public trust is anticipatory: people evaluate what the system could become, not only what it is.



One framework, five actionable principles

The framework's prescriptive payload

NO.	PRINCIPLE	KEY INSIGHT	LAYER FOCUS	PRIORITY
01	**Layered Architecture**	Trust emerges from alignment across all layers.	All four	Foundation
02	**Forward-Looking Trust**	Users evaluate what the system could become.	Institutional	Critical
03	**Productive Friction**	Intentional pauses build calibrated trust.	Agency	High

04 **Paradox Management** Trust paradoxes require navigation, not resolution. — Ongoing

05 **Ecosystem Integration** Trust infrastructure requires collective action. — Strategic

The rejection, layer by layer

AGENCY-LAYER

WEAK

Authorship cues fired ambiguously

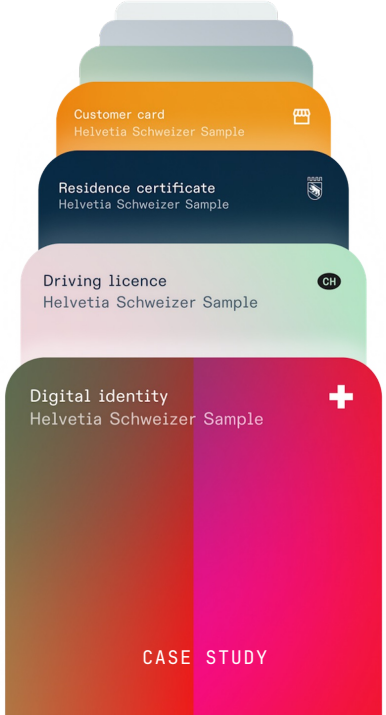
Unclear who was accountable

ENGINEERING -LAYER

SOLID

Technical design

Cryptographically sound



GOVERNANCE -LAYER

VAGUE

Accountability pathway

No binding oversight loop

INSTITUTIONAL -LAYER

FAIL

Future-safety guarantees

No constraint on future transformation

Image source: https://www.eid.admin.ch

Three honest limitations, each one opening the next phase

The framework is conceptual. These limitations are not concessions; they define the empirical research program that follows from the current work.

01

Empirical Validation

Conceptual framework. Validation across diverse deployment contexts still required.

02

Single-Coder Derivation

Construct derivation by a single coder. Inter-rater reliability not yet assessed.

03

Cultural Variation

Higher AI trust in China than Western samples. Context-specific adaptation likely required.

An empirical validation pipeline applies the cue catalog to documented incidents and best practices through a four-eyes classification protocol (independent classification by 2 LLM coders, adjudication by a 3rd on disagreement, inter-rater reliability target Gwet AC1 ≥

Scale AI responsibly. Architect digital trust

01 **FOUR-LAYER ARCHITECTURE** — Agency, Engineering, Governance, and Institutional layers integrated in one coherent model. Trust failures happen at the seams. Operationalization into 10 constructs and 127 cues.

02 **FIVE DESIGN PRINCIPLES** — Layered, Forward-Looking, Productive Friction, Paradox Management, Ecosystem Integration. Prioritized, actionable.

03 **TRUST FAILURES AT THE SEAMS** — Across Swiss e-ID, Coca-Cola, Apple, Deloitte …, the framework localizes each failure to specific layer-cue combinations and shows that strength in three layers does not compensate for weakness in the fourth.

04 **METHODOLOGICAL AUDIT TRAIL** — 56-source corpus, four-criterion evaluation, four-eyes validation pipeline, queryable bias ledger.

05 **THE BROADER ARGUMENT** — Closing the AI value gap is not about scaling AI faster. It is about scaling it responsibly, with digital trust treated as architected infrastructure.

06 **NEXT** — Validation pipeline (research), assessment products (IcebergScore: API Scan, Platform Audit, Vendor Assessment), and agency-side tools. The same 10 constructs and 127 cues power all three.

Thank you

“Thus, for the first time since the Enlightenment, the question arises again whether human beings are well advised to rely solely on their intellect and their capacity for rational thinking when making decisions.”

Source: Hüther, G. (2011)

Digital trust in AI systems is currently treated as a narrative or compliance claim rather than an architected, verifiable system property. This contributes to a persistent gap between AI potential and realized value.

What architecture engenders and sustains justifiable digital trust in AI systems?

A Multi-Level Digital Trust Framework, operationalized through 10 constructs and 127 trust cues, prescribing five design principles.

Why the AI value gap persists

01 · GOVERNANCE

Governance–AI Mismatch

GRC frameworks assume deterministic systems with traceable logic.

02 · OPERATIONS

Research–Practice Gap

Fairness metrics, causal inference, explainability hard to operationalise at scale.

03 · ETHICS

Principle–Action Gap

Ethical principles remain aspirational, not enforceable technical controls.

04 · ECONOMICS

Investment–ROI Gap

AI capabilities scale faster than organizational readiness to govern them.

05 · COMPLIANCE

Compliance–Trust Gap

Regulatory compliance does not automatically produce stakeholder trust.

06 · AUTHENTICITY

Quality Paradox

AI content may match human quality yet still be judged inauthentic.

07 · ECOSYSTEM

Flooding Problem

AI content overabundance erodes trust in all digital content.

COMMON ROOT

Architectural Fix

All seven close only through a layered architecture.

How the framework was built

01 INPUT

Corpus & sampling

56 primary sources (1964–2025)

8 regulatory frameworks (NIST AI RMF, EU AI Act, ALTAI, Three Lines, FINMA, eIDAS)

331 entries in wider bibliography

02 METHOD

GTDS hybrid

DSR Hevner et al. (2004) — 5 design-science cycles host build-and-evaluate

GT-LR Wolfswinkel et al. (2013) — 5-stage literature review hosts coding

Coding Strauss & Corbin (1998) — open · axial · selective; constant comparison per Glaser & Strauss (1967)

03 OUTPUT

Persistent artefacts

15 emergent categories (axial)

10 L1 constructs across 4 layers (selective)

127 L2 operationalising cues with literature anchors

04 VALIDATION

Evidence & limits

Plateau 14-of-15 at #14 (Lewicki et al., 1998); full 15-of-15 at #41 (Lankton et al., 2015)

Grounding Every construct anchored in primary literature; lineage documented in the concept matrix

Convergent Schlicker et al. (2025), Kim et al. (2008), Beldad et al. (2010), Kaplan et al. (2023)

AUDIT TRAIL · CONTAINED WITHIN A SINGLE COMPANION DOCUMENT

Resource table
every source with academic database, R1–R5 mapping, discipline, inclusion status

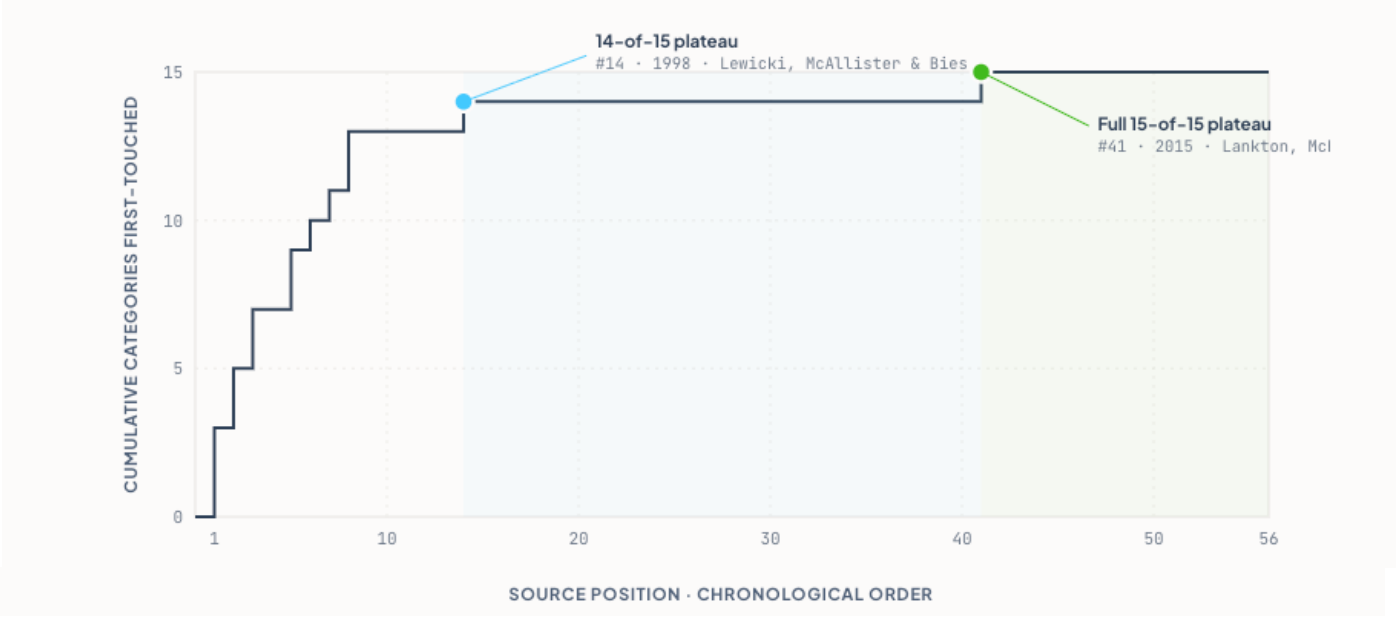
Concept matrix
65 rows × 18 columns mapping sources to constructs (Wolfswinkel advanced concept matrix)

Plateau analysis
first-touch category emergence at the 56-source corpus level — chronological year ordering

Limitations table
single-coder analysis, retrospective search-log reconstruction, deferred predictive validation

Conceptual coverage plateau

A cumulative category emergence across the 56–source primary corpus



From implicit trust to verifiable trust.

What breaks.

Implicit trust signals — brand reputation, platform dominance, perceived authority. In a world of AI-generated content, synthetic identities, and sophisticated deepfakes, these mechanisms prove fundamentally insufficient. The Entrust 2026 report logs deepfakes at ~20% of biometric fraud attempts, with national ID documents the most targeted type.

What replaces it.

Explicit, machine-verifiable guarantees. W3C Decentralized Identifiers and Verifiable Credentials for self-sovereign identity. C2PA cryptographic signing for content provenance. The Trust-over-IP stack with its Trust Spanning Protocol for interoperable messaging. Personhood credentials and formal delegation frameworks for AI agents. Trust becomes an architectural property, not a narrative claim.