Beyond the Black Box: Reclaiming Digital Trust

Stefano Falco 343739
Politecnico di Torino

ABSTRACT

Let's dwell on the idea of the Internet for a moment. It has become the main medium for essential services—from payments to healthcare—yet much of it still behaves like a **black box** from the user's point of view.
This report argues that opacity erodes trust and increases systemic vulnerability, and it proposes three levers to rebuild trust that can be operationalized and measured: **availability**, **controllability**, and **transparency**. Using selected case studies—presented in Section 3.1—we will analyze concrete mechanisms, limits, and trade-offs, and we'll outline enabling directions that can make trust verifiable rather than merely assumed. Throughout, we are going to demonstrate and emphasize that technical innovation must co-evolve with **user awareness** and rights that can actually be exercised.

1 INTRODUCTION

Inspired by the seminar held by *Prof. Paola Grosso* on 13th March 2025, this report will examine why mainstream internet services are perceived as opaque and how this perception undermines trust in increasingly pervasive digital infrastructures. As complexity grows, trust *"by delegation"* becomes fragile; instead, we argue for mechanisms that produce **inspectable evidence** of proper behavior.
We scope the discussion to developers, policymakers, educators, and end-users, and position the CIA features as design constraints rather than afterthoughts. As can be seen, **Figure 1** provides a conceptual map to place each mechanism along the stack (from routing to application logic to user interfaces).

1.1 Project's aim

We analyze how **accountability**, **availability**, **controllability**, and **transparency** are realized today, where they fail in practice, and which enabling technologies can close the gap. A central claim is that **user awareness** is not a soft add-on but a prerequisite: without it, even well-designed mechanisms fail to produce trustworthy outcomes in situ. The goal is to outline actionable guidance for systems where trust derives from properties that are **observable, testable, and enforceable**, aiming to a renewed digital trust.

2 THE "BLACK BOX" NATURE OF THE INTERNET

For most users, digital platforms operate on **trust without understanding**: algorithmic curation and automated decisions rarely expose inputs, assumptions, or error bounds. Even powerful trust-enhancing primitives (AI, cryptography, blockchain) remain conceptually distant, which results in this *"black box"* effect, characterized by:

I) **Loss of Agency**: Meaning users feel they have little control over the services they use.
II) **Dependence on Intermediaries**: Trust is outsourced to companies and platforms, often without the means to verify their claims.
III) **Greater Exposure to Manipulation**: Lack of transparency can allow exploitation, misinformation, and unfair practices to go unnoticed.

Figure 1: Overview of the internet architecture

2.1 Growing Dependency on Digital Services

Take a moment thinking about it: governments, businesses, and individuals alike **assume** constant online availability, secure transactions, and reliable information. But reliance on always-on connectivity creates **trust by necessity** more than **trust by evidence**. We highlight three stressors:

- **Potential security vulnerabilities**: Many systems are susceptible to cyberattacks, exposing personal data and critical infrastructures.
- **Platform concentration**: Small number of corporations introducing single points of control without sufficient oversight.
- **Data exploitation**: Users often unknowingly exchange personal information for access to services, with limited understanding of how their data is actually managed.

3 THREE MAIN CHALLENGES

As mentioned earlier in Section 1.1, to truly imagine and define a controllable and transparent services, we must place **the user at the center** of all digital system designs. Technology alone cannot guarantee trustworthy outcomes: users need decision points, intelligible controls, and rights that translate to concrete system behavior. We cannot forget that who uses these services is actually a human being, therefore empowering users means not only creating "better systems" but also promoting digital literacy, autonomy, and

critical engagement with technology. Therefore, to begin with, it is essential to to define the terms that we will use most often:

Availability. The assurance that systems, services, and data are accessible and usable when needed. Availability protects against downtime caused by attacks (*like DDoS*) or failures, ensuring continuous operation and service delivery.

Controllability. The ability for users or administrators to govern how their data, systems, and interactions are managed. Controllability means having the tools and permissions to configure, limit, or revoke access according to personal or organizational needs.

Transparency. The principle by which system operations, data usage, and decision-making processes should be visible and understandable. Transparency builds trust by allowing users to see how their information is handled and how system behaviors are determined.

3.1 Selection of study cases

For each of these properties, I have selected an example of **an existing platform or infrastructure** that **efficiently embodies them**. The aim is not only to showcase how digital services can be efficiently provided for users but also to highlight their limitations. This analysis will also help clarify the distinctions between these examples and the final proposals.

4 TRUSTWORTHINESS IN TODAY'S ECOSYSTEM

Trustworthiness emerges from the interaction of network engineering, security controls, governance of data, and usable interfaces. The three case studies embody distinct trust constructions: global resilience (*Cloudflare*), user-centric control and portability (*Mastodon*), and verifiable end-to-end assurances (*Signal*). We tie each to where it lives in **Figure 1** and to which user-visible guarantees it actually improves.

4.1 Cloudflare

Cloudflare operates a global Anycast network designed for **high availability and low latency**; each node can serve the full stack, eliminating intra-DC dependencies and reducing single points of failure.Redundancy, failover and load-balancing distribute traffic and absorb failures while **Cloudflare Tunnels** enable origin exposure minimization via outbound-only links.
Measurable effects include improved error budgets, shorter RTO/RPO in failure events, and better DDoS absorption headroom. **Figure 2** (*Distributed Cloud Bot Defense*) illustrates one control plane that—while security-oriented—also contributes to availability by preventing resource exhaustion at the edge.
There are some limitations as well, in particular residual centralization and vendor lock-in risk, which we note in **Section 6** when discussing decentralization. To summarize its *multi-layered strategy*:

- **Global Anycast Network**: Its extensive global presence utilizes an Anycast network, which means that traffic is routed to the nearest data center, minimizing latency and improving response times.
- **Redundancy and Failover**: Each server in Cloudflare's network runs the entire software stack, so requests are handled locally rather than being routed across different services in a data center.
- **Load Balancing**: Service used to distribute traffic evenly across groups of servers, preventing any single server from being overwhelmed.
- **Cloudflare Tunnel**: This feature provides a secure way to connect origin servers to Cloudflare without exposing them to the public internet.

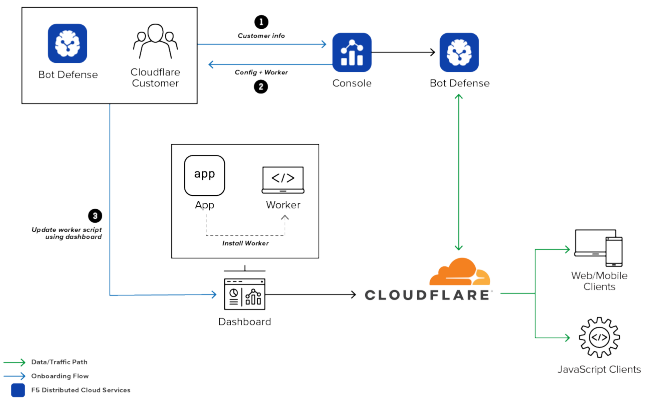

Figure 2: Cloudflare Distributed Cloud Bot Defense

4.2 Mastodon

Mastodon's federated architecture (*ActivityPub*) lets users select or self-host instances, aligning governance and moderation with community norms. **Practical controllability** appears as data export and import, account migration, block/mute lists, and instance-level policy diversity.
The trade-off is heterogeneity in QoS, moderation, and reliability across instances; user burden increases as control increases. We reference **Figure 1** to show how federation shifts power from a centralized application backend to multiple administrative domains.
Again, let's sum up the key ways Mastodon's architecture and design prioritize user control:

- **Decentralization and Server Choice**: Users choose a specific server to join based on their interests, values, or the server's moderation policies.
- **Data Ownership and Portability**: While the server administrator technically hosts the data, Mastodon emphasizes user autonomy and provides tools for users to export their data.
- **Moderation Flexibility and User Tools**: Each Mastodon server independently sets its own moderation policies and enforces them locally. This community-based moderation allows for more tailored and responsive handling of unwanted behavior compared to a single global policy.

4.3 Signal

Signal exposes client and server codebases as open source, enabling independent inspection and reproducibility of security claims. End-to-end encryption with Safety Numbers (plus change-key notifications) supports **out-of-band verification** and credible **MITM** detection.
The minimal-metadata posture reduces linkability risk at the service layer. **Figure 3** visualizes the metadata policy; we explicitly use it to argue how transparency of what is not retained is as important as transparency of code. Caveat: transparency presumes expertise—open code is necessary but not sufficient without community audit capacity.

The **Signal Protocol** is recognized by cybersecurity researchers as the gold standard for asynchronous messaging encryption (*OWASP's Secure Messaging Applications Guide 2020*).
This endorsement from the security community reinforces the transparency and robustness of their encryption.

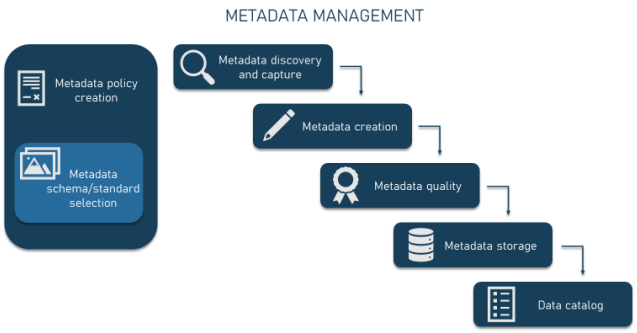

Figure 3: Signal Minimal Metadata Policy

5 USER AWARENESS

"In a world where we entrust machines with our most private thoughts, not understanding them is a form of surrender."
— **Shoshana Zuboff**, The Age of Surveillance Capitalism

The title of this article features a key term for our discussion, and we have already emphasized the importance of defining the terms we will use.
The concept of **Digital Trust** refers to the confidence users place in online systems to function securely, transparently, and fairly. Achieving strong digital trust requires more than advanced encryption or decentralized architectures; it demands an **active, aware, and empowered user base**. Without user awareness, even the most transparent technologies can become opaque and misunderstood, and controllability features may remain unused or misapplied.

In Section 2 we already mentioned these risks: the true antidote to the *"black box"* internet is a combination of transparent technology and educated users.
But how can users be trained in this sense?

Digital Literacy Programs. Public and private institutions should integrate cybersecurity education into general digital literacy initiatives, starting from basic school curriculums to professional training programs.

Gamification of Security Practices. Using gamified apps or simulations to teach cybersecurity basics (e.g., recognizing phishing, using two-factor authentication) can make learning more engaging for non-experts.

Microlearning and Continuous Updates. Short, regular security tips delivered through the platforms users already engage with can reinforce good practices over time, instead of relying on one-off, forgettable training sessions.

Peer-to-Peer Knowledge Sharing. Building community-based knowledge-sharing initiatives, like digital security workshops or *"cyber ambassadors"* within organizations, helps normalize cybersecurity conversations and spreads awareness more naturally.

Users need to understand how to manage their identities, data, and digital interactions in order to fully benefit from the tools and services available on the Internet. In doing this, **technical solutions alone are not sufficient**: education and digital literacy initiatives are equally important in empowering users to shift from passive consumption to competent verification.

6 REIMAGINING DIGITAL TRUST

Decentralized architectures can reduce opacity and single-control risks. Therefore, we will analyze **Web3** (operational transparency and distribution of control), **SSI** (user-centric, portable identity with selective disclosure), and **ZKPs** (provable statements without data exposure). Each is positioned as an enabler of availability/controllability/transparency, not a silver bullet; governance and usability of course remain critical constraints.

6.1 Web3 Initiatives

The core idea of *Web3* is a shift away from a central authority toward a network of informed participants. Aiming for a more **user-friendly and equitable digital landscape**, *Web3* leverages many decentralized technologies, primarily like blockchain.
In **Figure 4** summarizes the architecture; we use it to distinguish data-plane decentralization (e.g., IPFS) from control-plane decentralization (e.g., contract-governed services).
This transition to greater user control over their data, digital assets, and overall online experience is supported by smart contracts and technologies that facilitate peer-to-peer interaction, **relying much less on intermediaries**. That's why we can focus about:

- **Availability**: Data and services are hosted across decentralized networks (like IPFS or blockchain nodes), reducing single points of failure and censorship risks.
- **Accountability**: Blockchain's public ledgers make transactions and activities traceable and verifiable by anyone, preventing hidden manipulations.

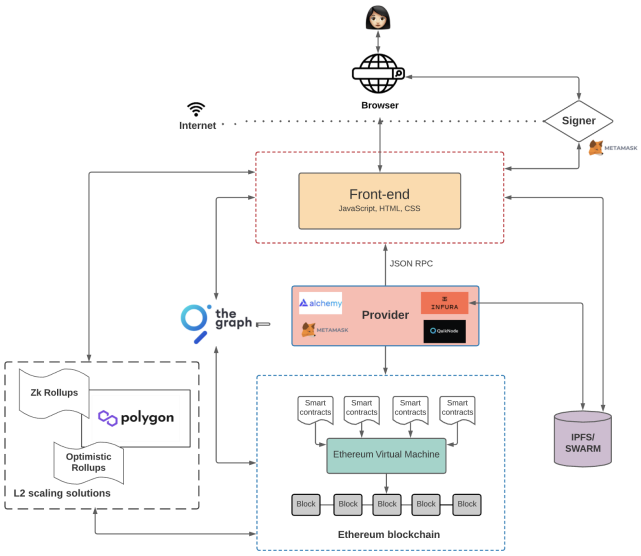

Figure 4: Overview of Web3 architecture

- **Controllability**: Users control their digital wallets, content, and interactions directly without relying on centralized companies (no third-party lock-in).
- **Transparency**: Smart contracts (self-executing code) are publicly auditable, ensuring that processes are predictable and visible.

6.2 Self-Sovereign Identity (SSI)

Self-Sovereign Identity represents a paradigm shift in how individuals manage their own digital identities and data. The core concept of *SSI* is that individuals should have ownership and control over their digital identities, **without relying on centralized authorities** such as governments or corporations.
This is typically achieved through the use of decentralized identifiers (*DIDs*), which are unique, globally resolvable identifiers that can't be controlled by any single organization.

Alongside DIDs, SSI relies on *verifiable credentials* (*VCs*), which are **digital representations of identity attributes or qualifications**, such as your ID card, your driver's license, university degree and so on all in one. These credentials are cryptographically signed by the issuer, making them tamper-proof and verifiable by relying parties. Let's have a look at the benefits in terms of:

- **Availability**: Identities are portable across platforms — even if a service shuts down, the user's identity persists independently.
- **Accountability**: SSI frameworks record when and how identity credentials are issued and verified, creating a clear, auditable trail.
- **Controllability**: Users decide what identity information to share, with whom, and for how long, often through selective disclosure techniques.
- **Transparency**: Credential issuers (like universities, governments) and their validations are visible and verifiable without compromising user privacy.

6.3 Zero-Knowledge Proofs (ZKPs)

Zero-Knowledge Proofs are cryptographic techniques that allow a prover to convince a verifier of a statement's truth while revealing nothing beyond validity. As a reference, **Figure 5** grounds the protocol phases we reference (setup, proving, verification) and where computational cost concentrates.
This is achieved through a series of interactions between the prover and the verifier, often involving **mathematical challenges and responses**, that provide statistical confidence in the truth of the statement.
Thanks to its key properties, first of all the **zero-knowledge** property (by which the verifier learns nothing beyond the validity of the statement), these techniques support the user in terms of:

- **Availability**: ZKPs enable trusted verification even across decentralized networks where trust between parties is minimal, boosting accessibility and system robustness.
- **Accountability**: Verifiers can be sure of a statement's truth without accessing sensitive data, reducing the need for risky data storage.
- **Controllability**: Users share only the outcome of the proof instead of surrendering full datasets like IDs or passwords.
- **Transparency**: The rules of verification are publicly known, and the mathematical proofs are checkable by any party, ensuring honesty without invasive surveillance.

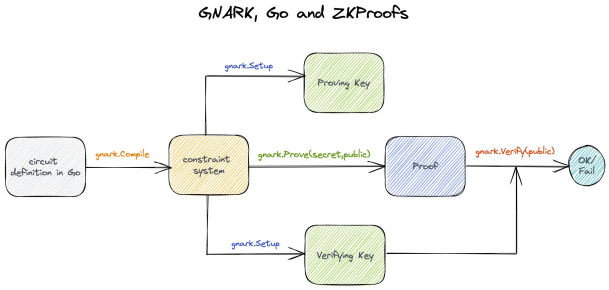

Figure 5: ZKProofs working schema with Gnark and Go

7 CONCLUSION

We unscrupulously entrust our security, our data, and our work resources to designated technologies, without really knowing if and how they will protect our sensitive information. This is where the concept of *transparency* comes into play.
But building a safer and more transparent internet is not solely the task of developers, regulators, or companies; **it is a shared responsibility between technology creators and users**. Functionality and trust must advance together.
Decentralization, SSI, and ZKPs are promising but only deliver if paired with enforceable **user rights** and **sustained awareness efforts**. The proposed path blends technical hardening with user empowerment, moving away from opaque delegation and toward an internet that is open, auditable, and human-centered.