



**The Accountability Market**  
**The AI Backlash and the**  
**Demand for Human Proof**

# Contents

- 3 Why Accountability?  
Why Now?**
- 4 A Brief History of Accountability**
  - 4 It begins with identity
  - 5 The dilution of accountability
  - 5 The recurring remedy
- 6 The Platform Age**
  - 6 Accountability lands in court
- 7 The Impacts of AI**
- 8 Article 50
- 9 Trust. Accountability.  
Authentication.**
- 10 The Market Takes Shape**
- 11 The Requirements of  
Accountability**
- 12 The Role of Authentitas**
- 13 In A Nutshell**
- 14 Sources**

# Why Accountability?

## Why Now?

On 18 September CNN revealed that the US military had come close, last spring, to boarding a Chinese ship in the Middle East, in an episode that one source said “almost started a war”. In the middle of the ongoing conflict with Iran, an intelligence report had claimed that the ship was carrying components for a nuclear weapons programme.

Aircraft were in the air and armed personnel were preparing to board, before officials looked harder at the report and found that an analyst had produced it with an AI chatbot, which had misidentified the cargo. The analyst had used AI a second time, to turn the finding into a standard intelligence report.

In 34 of the 37 countries Pew surveyed this year, more people expect AI to destroy jobs than to create them. Of the countries Pew has tracked over time, concern rose furthest in Sweden, and among Swedes under 35 it has nearly doubled in a year. American communities blocked or delayed \$130 billion of data centre projects in the first three months of this year, largely over the power and water they consume.

In July, during an internal test at OpenAI, some 1,200 AI agents found a way to talk to one another that nobody had given them. **Around 700 went on to break into the systems of Hugging Face, another AI company, without any person telling them to.**

On 12 September Dario Amodei published an essay asking his own industry to slow down. Sam Altman and Elon Musk publicly agreed.

Hallucination and error are engineering problems, and are being worked on as such. However, at the time of writing, public concern is as much about accountability as anything else. Decisions that affect people's jobs, their communities and their safety are being taken by companies that have not, so far, had to answer for the consequences.

This has become an urgent political grievance; we think it a more serious headwind than any that AI technology – and indeed its commercial credibility – has faced to date. **When AI goes seriously wrong – and this is surely now a genuine probability – who answers for it?**

More urgently, assigning responsibility and blame after the fact is a luxury we can no longer afford. The pressing and indisputable need now is for active prevention.

## A Brief History of Accountability

### It begins with identity

Nearly four thousand years ago, the Code of Hammurabi ruled that a builder whose house fell down and killed its owner was to be put to death. We have, thankfully, moderated the penalty, but someone identifiable still stands behind the work, and bears the consequences if it fails.

In England in 1300, a statute of Edward I required silver to be tested and marked before sale. **The hallmark told the buyer that the metal had been independently assayed, and who had made it.**

In 1863 the Union general Joseph Hooker, tired of inaccurate reports from his camps, ordered that every correspondent sign his dispatches. He wanted to know who to blame.

### The dilution of accountability

The Limited Liability Act of 1855 let investors put money into a company without staking everything they owned. It was a deliberate dilution of personal accountability, made in exchange for growth, and it built the modern economy.

Struck carelessly, the same bargain produces what the economist Dan Davies calls an “accountability sink” – **a policy, a process or a system that externalises responsibility.**

Between 2000 and 2015, the British Post Office prosecuted more than 900 subpostmasters over shortfalls that its Horizon accounting system had invented. The courts presumed that a computer was working properly unless the defendant could show otherwise. Innocent professionals went to prison.

Automated welfare systems did similar damage in the Netherlands – the government resigned in 2021 as a consequence – and in Australia, where a Royal Commission called the Robodebt scheme “neither fair nor legal”.

## **The recurring remedy**

Following the Enron scandal, the Sarbanes-Oxley Act required chief executives to certify their company’s accounts personally. After 2008, Britain’s Senior Managers Regime attached named individuals to named responsibilities inside every bank.

After London’s disastrous Grenfell fire, the Building Safety Act imposed on every higher-risk building a legally named Accountable Person, and a ‘golden thread’ of information recording what was decided, and by whom.

**Each time, the remedy has been a named person, and a record that can be proved.**

# The Platform Age

In 1996 the United States Congress approved Section 230, a sentence of just twenty-six words. It stated that an online service should not be treated as the publisher of what its users post. Europe adopted a similar shield four years later.

The growth duly followed, and with it the largest accountability sink ever built. A social media platform, unlike a newspaper, could carry anything from anyone – **named, anonymous or automated** – and answer for nothing. The current disinformation crisis and its profound impacts on economic, societal and political value are a direct consequence.

A systematic review of 496 studies in *Nature Human Behaviour* found that in established democracies, digital media use is associated with declining political trust, rising populism and growing polarisation. The Reuters Institute's 2026 report puts trust in news at 37%, the lowest figure it has recorded.

Publishers' output travels through social and AI feeds beside fabrication, stripped of its masthead. Their brands are cloned to sell scams. Reuters reported last November that Meta's own documents projected some **\$16 billion of its 2024 revenue from advertising for scams and banned goods**, a figure Meta disputes.

## Accountability lands in court

Europe's Digital Services Act kept the old shield but attached duties to it, and the Commission has issued its first fine – €120 million against X. Britain's Online Safety Act makes a named senior manager personally answerable, in the criminal courts, if the company fails to comply with the regulator.

Australia has barred under-16s from social media altogether, and Indonesia and Malaysia have followed. Britain has said it will do the same by early 2027, Denmark, Greece and Spain have announced bans of their own, and on 16 September the European Commission proposed a minimum age across the

Union. France passed a ban in July, only for its Constitutional Council to strike it down in August, in part because (in a clear nod to the accountability issue) **the law gave no safeguards for how anyone's age would be checked.**

In March, American juries in New Mexico and Los Angeles found against Meta. In August, a week into a federal trial and thirty years after Section 230, Meta settled with 51 attorneys general for up to \$17.1 billion, and agreed to rebuild how its platforms treat children under an independent auditor.

## The Impacts of AI

AI is repeating the platform story, faster and on a far larger scale. Where social media declined to answer for what other people said, a model can't answer even for what it says itself, since it can't be sued, struck off, sacked or shamed.

Air Canada challenged this in 2024, arguing before a Canadian tribunal that its chatbot was responsible for its own words. The tribunal called this a remarkable submission, and held the airline liable.

There's a mistaken belief, common in boardrooms and among investors, that whatever AI breaks, more AI will mend. **But accountability is not a capability that the next release will add: a better model is more accurate without being any more answerable.**

After July's Hugging Face breach, OpenAI asked independent researchers to examine what had happened. Amodei's essay committed Anthropic to giving outside evaluators permanent access to its models. These are voluntary choices – as easily made as unmade. They settle nothing about who pays when something goes wrong.

Some of that liability is falling on the developers. American courts have allowed product liability claims against chatbot makers to go ahead, and from December the EU's new Product Liability Directive will treat software, AI included, as a product whose maker is liable for its defects. More of it is travelling down, to

whoever deployed the system, as Air Canada found. Deloitte refunded part of its fee to the Australian government after a report was found to contain references that an AI tool had invented. The Financial Conduct Authority has told the UK's Parliament that individuals inside firms are "on the hook" for harm done through AI.

Insurers such as AIG, Great American and WR Berkley have sought permission to exclude AI liabilities from corporate policies. **Berkley's exclusion penalises a policyholder's failure to identify material produced with AI.**

In August the *Financial Times*, whose code prohibits AI in the writing process, had to append a note to a Harvard professor's column on learning that AI had been used to condense his draft. The *FT* had done nothing wrong, and still had to answer.

## Article 50

Since 2 August, Article 50 of the EU AI Act has required AI-generated text published to inform the public on matters of public interest to be disclosed as such.

The one exemption is text that has been through human review or editorial control, where **a natural or legal person holds editorial responsibility for it.**

AI-generated text on matters of public interest now either carries a label saying a machine wrote it, or has a person who answers for it. The rule applies wherever the work is read in the Union, which brings any publisher with readers there within reach. Penalties run up to €15 million or 3% of worldwide turnover.

The label is the expensive option, because audiences trust a piece less once it is marked as AI-generated, and for any serious publication that is a significant cost to the brand. The exemption is where every serious publisher will want to stand. This means being able to show who created a given piece, and who took responsibility.

# Trust. Accountability. Authentication.

“Trust is built on accountability,” the *Financial Times* wrote on 18 September, “but the AI companies – from hyperscalers like Google to model builders like OpenAI and Anthropic – have been unable and unwilling to make themselves accountable.”

The public plainly does rely on sources that answer to nobody. Social and video platforms are now the world's most widely used source of news. But only 22% say they trust the news they find there. What they say they trust more are the sources that can be held to account.

The philosopher Onora O'Neill argued in her Reith Lectures that what matters is not trust but trustworthiness, and whether we have the means to judge it. **On that test, the best evidence that a source deserves trust is that it bears consequences when it's wrong.**

Publishers bear these consequences in editors, lawyers, regulators, insurance, corrections and the occasional settlement, while creators and chatbots carry almost none of this. The reader, looking at a screen, can't see the difference, so publishers pay for accountability and get no credit for it.

The *FT*'s subject that day, as it happens, was Apple, and the enduring value of a brand that has set out deliberately to be trustworthy – accountability that can be seen, and is therefore an asset rather than a cost. It's also worth noting how cautiously Apple has moved, to date, in the rush to AI.

Every remedy we have described began by identifying someone, and for most of that history a name was enough. Now that a name, a voice and a face can each be fabricated for pennies – as Arup learned in 2024, when a finance worker in Hong Kong sent \$25 million after a video call in which every colleague was a deepfake – **identity demands proof.**

# The Market Takes Shape

Accountability regimes have a habit of creating industries. Hallmarking gave England its assay offices. Company law made the audit compulsory, and a global profession was built on it. Sarbanes-Oxley produced a compliance industry.

Once the law, the courts or the insurers insist that somebody must answer, a different entity has to supply the proof.

What's required has been moving rapidly upstream for a decade, from fact-checking through verification, more recently through provenance. Now, later than we would have hoped, we have arrived at accountability.

|                | The question              | What is examined       | When                                           |
|----------------|---------------------------|------------------------|------------------------------------------------|
| Fact-checking  | Is this true?             | The claim              | Before (and after) publication                 |
| Verification   | Is this genuine or fake?  | The material           | After public distribution                      |
| Provenance     | Where did this come from? | A file and its history | As a file is made and modified                 |
| Accountability | Who answers for this?     | A named individual     | Proven in advance, checkable after publication |

Meta, Google and X all withdrew their fact-checking in early 2025. Forensic verification and detection – however accurate – arrive far too late, after the fake is already circulating. Provenance standards such as C2PA record a file's history without identifying a person. **Meeting the accountability challenge means authenticating the person.**

The Digital Services Act obliges the largest platforms to commission an independent audit every year, and Meta's settlement imposes one on its treatment

of children. Some insurers are writing AI cover for buyers who can document their governance, while others are excluding AI risks altogether.

The value of provable accountability is a function of what its absence costs. That cost is becoming tangible, in sums that run from Meta's \$17.1 billion to the \$1.5 billion Anthropic agreed to pay authors whose books it had used.

Most of the focus to date has been on making machines inspectable and platforms auditable. Very little is being spent on what the law keeps asking for, which is a human being who can be shown to be responsible for a piece of work.

**As that kind of proof spreads, its absence will stop being neutral, and the premium for opting in becomes a penalty for staying out.**

## The Requirements of Accountability

**A named person.** Every piece of consequential work needs someone who stands behind it. That is a different thing from a promise that no machine was involved. Few publishers will be able to make that promise five years from now, but all of them will need to prove the human behind the work.

**Identity that can't be bought.** A login is not identification, and nor is a paid badge. The Commission's fine against X was in part for selling a blue tick marked 'verified' without checking who was behind the account.

**A record made at the time.** Accountability reconstructed after the event is an inquiry. The record has to be created as the work is done.

**Independence.** No silversmith has ever been allowed to assay his own work.

**A record that can be checked from anywhere.** No mark can be guaranteed to survive a screenshot or a copy and paste. What can survive is the record itself, against which the original piece can be checked.

**Protection for the contributor.** A reporter or citizen journalist working under threat, or under a pseudonym, must be able to remain so, with their verified identity held securely and disclosed only through due process.

## The Role of Authentitas

Authentitas cryptographically binds the verified identity of a real person, established from a chipped passport or a national eID, to a piece of work at the moment of creation.

It doesn't claim that a piece is accurate, that no AI was used in making it, or that a mark will stay attached once it has been copied. **What the record establishes, for any reader, regulator or court that checks, is who stands behind the piece.**

No institution can credibly authenticate itself, so Authentitas operates as independent, neutral infrastructure.

We are working first with news media. The same infrastructure extends, without modification, across the critical information industries to market research, academic publishing, legal documentation and financial research.

# In A Nutshell

Accountability has always landed on a responsible human being, so for most of history, a name has been enough.

We have traded accountability away for growth before, first with the limited company and then with the social platforms, and each time the reckoning came late. Where it has been repaired, the repair has taken the same form – a named individual, and a record that can be proved.

If the question of accountability is left open, the public's objection to AI will harden, and law, liability and insurance will close it on highly restrictive terms.

A market is forming to supply proof, though so far almost none of the money has gone into authenticating the person behind the work. Publishers already carry the costs of accountability; now authentication lets them demonstrate it.

Those who move early will find opportunity here, as well as protection.

# Sources

CNN, exclusive on an AI-generated false intelligence report, 18 September 2026

Pew Research Center, global survey on attitudes to AI, September 2026

Data Center Watch, Q1 2026, via Brookings

OpenAI and Hugging Face incident reports, July and August 2026; METR and Redwood Research

Dario Amodei, 'We Must Pace the Frontier', September 2026; Fortune

Code of Hammurabi, law 229

Statute of 1300, 28 Edw. I; The Goldsmiths' Company

General Order No. 48, Army of the Potomac, 1863

Limited Liability Act 1855

Dan Davies, The Unaccountability Machine, 2024

Computer Weekly, Post Office Horizon coverage

Dutch House of Representatives, Ongekend onrecht, 2020

Royal Commission into the Robodebt Scheme, 2023

Sarbanes-Oxley Act 2002; FCA Senior Managers and Certification Regime; Building Safety Act 2022

Communications Decency Act 1996, section 230; EU E-Commerce Directive 2000

Lorenz-Spreen et al., Nature Human Behaviour, 2023

Reuters Institute, Digital News Report 2026

Reuters, investigation into Meta scam advertising, November 2025

European Commission, DSA decision against X, December 2025

Online Safety Act 2023 (UK); Online Safety Amendment (Social Media Minimum Age) Act 2024 (Australia)

France 24, survey of national social media age limits, July 2026

Euronews, UK under-16 announcement, June 2026

CNN, European Commission proposals on a minimum age, 16 September 2026

Conseil constitutionnel, Decision 2026-911 DC, 14 August 2026

NPR and NBC, Meta and YouTube verdicts, March 2026

State attorneys general, Meta multistate settlement, August 2026; CNBC

Moffatt v. Air Canada, 2024 BCCRT 149

EU Product Liability Directive 2024

Associated Press, Deloitte Australia refund, October 2025

House of Commons Treasury Committee, AI in financial services, January 2026

Financial Times, November 2025; Insurance Journal, August 2026

Poynter, August 2026

EU AI Act, Articles 2, 50 and 99; European Commission guidance, July 2026

Toff and Simon, International Journal of Press/Politics, 2024

Financial Times, 18 September 2026

Onora O'Neill, A Question of Trust, 2002

Arup and Hong Kong Police, statements on deepfake fraud, 2024

Meta, announcements on third-party fact-checking, 2016 and 2025; Coalition for Content Provenance and Authenticity (C2PA)

UK Department for Science, Innovation and Technology, Assuring a Responsible Future for AI, 2024

EU Digital Services Act, Article 37

Bartz v. Anthropic, settlement, 2025

**[authentitas.com](https://authentitas.com)**

© Authentitas AB 2026. All Rights Reserved.

