Who Pays to Check the Machine

Verification Cost as the Binding Constraint on Generative AI in Professional Work, and Why the Firm Boundary Moves

Christopher I. V. Farmer

Independent researcher, Geneva, Switzerland

6 August 2026

Suggested citation. Christopher I. V. Farmer, 'Who Pays to Check the Machine' (2026) preprint.

Status. Preprint. Submitted for publication; not yet peer reviewed.

Keywords. verification cost; credence goods; costly state verification; artificial intelligence; productivity; retraction; research integrity; professional liability

Contents

Who Pays to Check the Machine

Verification Cost as the Binding Constraint on Generative AI in Professional Work, and Why the Firm Boundary Moves

Christopher I. V. Farmer

Abstract

Generative artificial intelligence has collapsed the cost of producing plausible assertion while leaving the cost of establishing whether an assertion is true substantially unchanged. This paper argues that the resulting asymmetry, and not the automatable share of tasks, governs the economic effect of the technology in the sectors where it is most heavily adopted; and it argues that the asymmetry does not resolve itself, because verification cost is not the kind of cost that amortises.

The argument has three parts. The first is a re-specification. In markets for goods whose quality the buyer cannot assess even after consumption—law, audit, medicine, compliance, scholarship—the binding constraint on what a production technology delivers is the cost of verification and not the cost of production, so a supply shock to the second does not become productivity growth. It becomes *verification rent*, a redistribution of surplus towards whoever can certify, and a re-allocation of liability towards whoever cannot. The growth literature has argued at length about how much labour the technology displaces. It has not asked what happens when the marginal cost of a claim falls to zero and the marginal cost of checking it does not.

The second part is the paper's central theoretical claim and it is a claim about cost structure. The dominant explanation of the present productivity shortfall treats the verification and governance practices that must be built around a new technology as *intangible capital*—a stock, accumulated once, thereafter yielding returns, and therefore transitional. This paper argues that verification cost divides, and that the division matters more than the total. Establishing that a cited authority *exists* is mechanisable and does amortise: it is a stock, and building the register is the investment. Establishing that a real authority *supports the proposition it is advanced for* is a fresh judgement about a fresh pairing every time the pairing is made. That is a flow, it is the larger part, and it does not amortise. A discriminating test is stated: does the measured verification burden in a sector fall with cumulative experience of the technology, or scale with the volume of output produced?

The third part is an original measurement, and its result is a negative one about measurement itself. Using the Retraction Watch corpus against Crossref annual publication counts, the paper computes, for every publication cohort from 2005 to 2023, the proportion of journal articles retracted within a fixed window after publication. Over two decades in which similarity detection, image forensics, post-publication review and paper-mill screening were all built out, the three-year detected retraction rate rose **6.3-fold**, and the rate for integrity-type retractions specifically rose **11.4-fold**, from 0.55 to 6.25 per ten thousand articles published. The integrity share of retractions rose from 38 per cent across the first five cohorts to 70 per cent across the last five.

That series admits of two readings. Either the production of defective work rose roughly elevenfold, in which case the record is deteriorating quickly; or production was roughly stable and detection improved elevenfold, in which case the 2005–2015 record contains an undetected residue of about ten times what was found. Both are serious and they are not the same. **Nothing observable from inside the system distinguishes them**, because separating them requires knowing the rate at which defective work is produced, which is

precisely the quantity the verification constraint makes unobservable. A verification apparatus operated for twenty years over a corpus it substantially controls cannot determine, from its own output, whether the underlying rate of defective production has risen or fallen. That is the paper's strongest single result, and it is a result about what the measurement cannot reach.

The paper closes with a proposal that follows from the analysis rather than from preference: a **verification duty** which *divides* the cost of checking rather than moving it. The component a register can discharge is allocated to the party that captures the saving from cheap production; the component no register can discharge is left, expressly and cumulatively, with the professional. It is framed as a statutory duty rather than an incremental extension of the common law, for reasons given at §20.1A. The proposal is positioned against the transparency obligations of the Artificial Intelligence Act, which came into application on 2 August 2026, and against the new Product Liability Directive, and the residual claim is stated at the width those instruments leave.

A companion paper applies the same cost structure to a different institution. Where this paper asks what happens to a market when producing a claim becomes cheaper than checking it, the companion asks what happens to a body of law when taking an action becomes cheaper than proving who took it. The two share a structure and not a cause, and neither depends on the other.

Keywords: verification cost; credence goods; costly state verification; artificial intelligence; productivity; retraction; research integrity; professional liability

---

Note on sources and method

Every authority in this paper has been checked against a primary or authoritative record before citation. Journal literature was resolved against the Crossref registry and accepted only where title, first author and year agreed; English authority was checked against a citation index of 75,837 judgments harvested from Find Case Law under a computational-analysis permission granted by The National Archives; law reviews, monographs and official instruments were verified individually against the publisher or the issuing body. Anything that could not be verified was removed rather than approximated.

This is stated not as a courtesy but because the paper's subject makes it material. A paper arguing that verification cost is the binding constraint of the period would be self-refuting if it declined to pay that cost. The apparatus is set out in the Appendix, together with the classes of failure that survived it—including failures in the checking tools themselves. They are reported because a paper making this argument should be judged partly on what its own verification missed.

The derivation code for the empirical work in Part IV, the registers, and the verification tooling accompany this paper as supplementary material and are itemised at Appendix §A.6, which names each file and states what it contains. Every figure reported is reproducible from the stated sources by the stated method.

Table of Cases

United Kingdom

Banca Nazionale del Lavoro SpA v Playboy Club London Ltd [2018] UKSC 43, [2018] 1 WLR 4041
Barton v Morris [2023] UKSC 3, [2023] AC 684
Caparo Industries plc v Dickman [1990] 2 AC 605 (HL)
Harber v Commissioners for His Majesty's Revenue and Customs [2023] UKFTT 1007 (TC)
Hedley Byrne & Co Ltd v Heller & Partners Ltd [1964] AC 465 (HL)
Khan v Meadows [2021] UKSC 21, [2022] AC 852
Manchester Building Society v Grant Thornton UK LLP [2021] UKSC 20, [2022] AC 783
R (Ayinde) v London Borough of Haringey [2025] EWHC 1040 (Admin)
R (Ayinde) v London Borough of Haringey; Al-Haroun v Qatar National Bank QPSC [2025] EWHC 1383 (Admin)

R (Bancoult) v Secretary of State for Foreign and Commonwealth Affairs (No 2) [2008] UKHL 61, [2009] 1 AC 453
R (Hamid) v Secretary of State for the Home Department [2012] EWHC 3070 (Admin)
Robinson v Chief Constable of West Yorkshire Police [2018] UKSC 4, [2018] AC 736
Steel v NRAM Ltd [2018] UKSC 13, [2018] 1 WLR 1190

United States

Mata v Avianca Inc 678 F Supp 3d 443 (SDNY 2023)
Park v Kim 91 F 4th 610 (2d Cir 2024)

Table of Legislation and Instruments

European Union

Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products [2024] OJ L2024/2853
Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L2024/1689

United States

Federal Rules of Civil Procedure, r 11
Rules of the United States Court of Appeals for the Second Circuit, r 46.2

Each instrument listed is cited in a footnote for the provision identified there, and no provision is relied on for a proposition it does not state.

Part I—Introduction

1. The problem stated

The cost of producing a plausible assertion and the cost of establishing whether it is true have moved apart. The institutions that price professional work, and that allocate liability when it goes wrong, were built when the two costs were close together, and they have not been re-specified.

The cost of producing a plausible assertion has collapsed. A competent-looking legal argument, a literature review, a compliance memorandum, a research paper with methods and results, can now be generated in seconds at negligible marginal cost. The cost of establishing whether such an assertion is *true* has not moved at all. Checking whether a cited case exists still requires consulting the report, and checking whether an experiment was run still requires the underlying data. The third case is different in kind: checking whether an argument holds requires a reader capable of having made it, and there is no register against which that can be resolved.

This paper is about what happens to a market when that gap opens, and its argument is that the effect is not primarily on employment, which is where the discussion has concentrated.[1] It is on price, on the allocation of liability, and on measured productivity, and the sign of the prediction is the opposite of the one the automation literature gives.

[1] Compare Daron Acemoglu, 'The Simple Macroeconomics of AI' (2025) 40 *Economic Policy* 13 with the transformative accounts discussed at Part II §7 below. The article was published on advance access in August 2024 and appears in the issue of January 2025; the issue year is given, per OSCOLA.

2. Why this is not simply the automation debate

The literature on the economic effects of artificial intelligence has organised itself around a single question: what share of tasks can the technology perform?[2] Answers differ by an order of magnitude, and the disagreement is substantive.[3] Experimental work on writing and support tasks finds real and rapid gains.[4]

That question is the right one for goods whose quality the buyer can observe. For goods whose quality the buyer cannot observe it is the wrong question, and it is wrong in a way that reverses the sign of the prediction. Where quality is unobservable, what limits useful output is not how much of the work the technology can perform but how much of it the technology can be trusted to have performed. Trust in this setting is a cost borne by an identified party and measured in hours, not a disposition.

The economics of that situation is old and well understood. Akerlof established that unobservable quality can collapse a market; Darby and Karni added the credence category, in which quality cannot be assessed even after consumption, and showed that the optimal amount of fraud in such a market is not zero.[5] Townsend formalised the contracting problem faced by a principal who can learn the truth of a report only by paying a fixed cost, and showed that the optimal contract verifies *stochastically* rather than always.[6] None of that is new. What is new is a technology that moves one of the two costs by an order of magnitude and leaves the other where it was.

3. What this paper does

Part II re-specifies the problem: it defines the two costs, states three propositions about what happens when their ratio moves, and derives the prediction that measured productivity can fall in a sector whose output volume is rising, as arithmetic rather than as mismeasurement.

Part III shows the same failure in two further markets—scholarly publishing and audit assurance—chosen because they differ in who pays the verifier, which turns out to be the variable that determines how badly the failure bites.

Part IV is the paper's empirical contribution and the test of its central claim. It measures, across every publication cohort from 2005 to 2023, the rate at which journal articles were retracted within a fixed window after publication, against the number published. The result is reported in full, including the two readings it admits of and the fact that the data cannot choose between them.

The counterarguments are met where they arise, at §8.1 to §8.4: the productivity J-curve, mismeasurement, the automation of verification, reputational and liability sorting, and Baumol's cost disease. The last of these is the strongest and is taken last.

Part V proposes a verification duty, positions it against the instruments that came into force while this was being written, and states what it does not achieve.

Part VI concludes, and sets out the three developments that would require the analysis to be revised.

4. What is new here, and what is not

A paper of this kind should say plainly which of its claims are original and which are restatements, because the distinction is easy to blur and expensive to get wrong.

[2] Acemoglu (n 1).

[3] See Part II §7.

[4] Shakked Noy and Whitney Zhang, 'Experimental evidence on the productivity effects of generative artificial intelligence' (2023) 381 *Science* 187; Erik Brynjolfsson, Danielle Li and Lindsey R Raymond, 'Generative AI at Work' (2025) 140 *Quarterly Journal of Economics* 889.

[5] George A Akerlof, 'The Market for "Lemons": Quality Uncertainty and the Market Mechanism' (1970) 84 *Quarterly Journal of Economics* 488; Michael R Darby and Edi Karni, 'Free Competition and the Optimal Amount of Fraud' (1973) 16 *Journal of Law and Economics* 67.

[6] Robert M Townsend, 'Optimal contracts and competitive markets with costly state verification' (1979) 21 *Journal of Economic Theory* 265.

4.1 What is not new

That verification cost matters in markets for goods whose quality cannot be observed is the founding result of the credence-goods literature and dates to Darby and Karni. **That a principal who must pay a fixed cost to learn the truth of a report faces a distinct contracting problem** is Townsend's costly state verification, formalised in 1979.[7] **That rents migrate to complementary assets** is Teece.[8] **That organisational boundaries sit where measurement cost changes** is Barzel, Alchian and Demsetz, and Holmström and Milgrom; that a technology altering monitoring cost moves the boundary is demonstrated empirically by Baker and Hubbard.[9] **That verification rather than capability is the binding constraint on what this technology delivers** has been reached independently and in parallel by Catalini, Hui and Wu, who model the collision between a falling cost to automate and a cost to verify bounded by human time.[10] Priority is not claimed for any of these and the convergence with the last is of more use to the argument than priority would have been.

The "verification ratio" introduced at §5 is a piece of exposition, not a discovery. It names the ratio of two costs the literature has long treated separately, and the claim should be put no higher: holding verification cost fixed, the ratio is a monotone transform of the production cost and carries no information the automatable-share view does not. What does the work throughout is the *level* of verification cost, and the ratio is a way of keeping it in view.

4.2 What is offered as new

First, the stock-versus-flow partition of verification cost, and the objection to the productivity J-curve that follows from it (Part II §8.1). The received explanation of the present productivity shortfall treats verification and governance practice as complementary intangible capital whose accumulation is transitional. This paper argues that verification cost divides into a component that behaves like a stock and a component that behaves like a flow, that the J-curve describes only the first, and that the second is the larger. Electrification required a factory to be rebuilt once. Establishing that a real authority supports the proposition it is advanced for is a fresh judgement every time the pairing is made, and no accumulated infrastructure retires it.

The cost taxonomy itself is not new—quality economics has separated prevention costs, which are investments, from appraisal costs, which recur per unit inspected, since the 1950s.[11] What is offered is the application to this technology, the identification of the boundary between the two components, a discriminating test, and at §16.3 the first attempt to measure the two components separately on a real corpus. That measurement is a proxy for the partition rather than the partition itself and its margin is sensitive to the window chosen; what it shows is that the rise in caught defects is concentrated almost entirely in the failure types the long-amortised tool does not reach, which is the direction the partition predicts and the direction the intangible-capital account does not.

Second, the identification result in the retraction cohort study (Part IV). The rate at which journal articles are caught and retracted within a fixed window after publication has risen sharply across two decades—6.3-fold in three years for retractions generally and 11.4-fold for integrity-type retractions specifically.

---

[7] Townsend (n 6).

[8] David J Teece, 'Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy' (1986) 15 *Research Policy* 285.

[9] Yoram Barzel, 'Measurement Cost and the Organization of Markets' (1982) 25 *The Journal of Law and Economics* 27; Armen A Alchian and Harold Demsetz, 'Production, Information Costs, and Economic Organization' (1972) 62 *American Economic Review* 777; Bengt Holmström and Paul Milgrom, 'Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design' (1991) 7 *Journal of Law, Economics, and Organization* 24; George P Baker and Thomas N Hubbard, 'Contractibility and Asset Ownership: On-Board Computers and Governance in US Trucking' (2004) 119 *Quarterly Journal of Economics* 1443.

[10] Christian Catalini, Xiang Hui and Jane Wu, 'Some Simple Economics of AGI' (2026) SSRN 6298838.

[11] The prevention–appraisal–failure taxonomy originates with Joseph M Juran and Armand V Feigenbaum in the 1950s; appraisal cost is by construction incurred per unit inspected, and prevention cost is an investment that amortises. The distinction drawn at Part II §8.1 is that one, applied to a general-purpose technology.

The rise is **not** offered as new. It is established: Steen, Casadevall and Fang asked why the number of retractions had risen and answered in favour of detection; Van Noorden reports the proportion of a year's papers subsequently retracted as having trebled in a decade; and Cokol and colleagues modelled the undetected residue two decades ago.[12] §15 positions the measurement against that work. What the construction here adds is a strict fixed-window comparison, which controls the right-censoring the cumulative measures do not, and a cohort-level split between total and integrity-type retraction.

What is offered as new is the **use to which the series is put**. It is that the rise is as consistent with a deterioration in the record as with an improvement in detection; that the two readings imply very different worlds; and that **no observation available from within the system separates them**, because separating them requires knowing what the apparatus missed. Steen and colleagues answered the question in one direction; this paper argues that the data does not permit the question to be answered, and sets out the partial discriminators—including one that leans against the detection reading and is not, so far as the author can establish, reported elsewhere—that narrow the range without closing it. A verification apparatus cannot measure its own coverage. That is the paper's thesis, established in the one credence market whose numerator is public.

Third, the firm-boundary prediction confined to its observable signature (§6.2). The mechanism is textbook and its application to this technology has been made independently and earlier by Hydari and Muzaffar.[13] What survives as new is one falsifiable proposition: that the signature of the shift is a movement in advisory work *away* from output-priced engagements and back towards time- and retainer-based pricing, because a fixed fee prices a deliverable and the deliverable is no longer the expensive part. That the underlying contract-form result is Bajari and Tadelis's is conceded at §6.2.[14] It should also be said plainly that the reported direction of travel is the opposite. Trade surveys record 72 per cent of United States firms offering alternative fee arrangements, rising to 91 per cent among firms of more than fifty lawyers and 96 per cent among firms of more than a hundred and fifty, with flat fees the most common form; and the trade press expects the technology's efficiencies to accelerate the shift through 2026.[15] The prediction here is therefore contrarian and may be wrong. Three things qualify the contrary evidence and none disposes of it. What is surveyed is what firms *offer* rather than what they bill. The same survey reports that 58 per cent of firms had seen no billing impact from the technology at the date of the survey, so the attribution of the shift to it is a forecast rather than a finding; and among the large firms furthest along in adoption, more said the effect was greater efficiency with no change to billable hours than said hours had fallen, which is the pattern the account here predicts and not the one a straightforward efficiency gain would produce. And an offer made before the verification cost has been felt is not evidence about where the cost settles. The prediction is about billed engagements in work whose output cannot be checked at arm's length, and it is offered because it is checkable on that measure.

Fourth, the division rather than reallocation of the verification duty (Part V). The proposals literature treats the question as who should bear the cost. The analysis here yields a different answer: the cost is not one thing, and the two components should sit with different parties. Limb three of the proposal is not a saving provision but the operative one.

4.3 What would falsify each claim

The **stock/flow partition** fails if the measured verification burden in a sector declines with cumulative experience of the technology rather than scaling with the volume of output produced. The test must be

[12] See Part IV §15 and the sources there.

[13] Muhammad Zia Hydari and Farooq Muzaffar, 'Redrawing the AI Map: A Theory of Accountability Boundaries in Agentic Ecosystems' (arXiv:2605.23179, 22 May 2026) https://arxiv.org/abs/2605.23179 accessed 4 August 2026.

[14] Patrick Bajari and Steven Tadelis, 'Incentives versus Transaction Costs: A Theory of Procurement Contracts' (2001) 32 *The RAND Journal of Economics* 387.

[15] Best Law Firms, 'Law Firms Embrace AFAs, But Clients Want More Flexibility' (11 November 2025) https://www.bestlawfirms.com/articles/law-firms-embrace-afas-but-clients-want-more-flexibility/7098 accessed 4 August 2026; 'With Law Firm AI Use on the Rise, Expect Alternative Fee Arrangements to Pick Up Steam in 2026' (*The American Lawyer*, 18 December 2025) https://www.law.com/americanlawyer/2025/12/18/its-real-now-with-law-firm-ai-use-on-the-rise-expect-alternative-fee-arrangements-to-pick-up-steam-in-2026/ accessed 4 August 2026.

applied to the irreducible component and not to the mechanisable one, which a register does amortise and which the proposal in Part V is designed to amortise.

The **cohort result** fails if an independent measure of the production rate of defective work becomes available and shows it flat, which would resolve the identification problem in favour of the detection reading. Nothing presently supplies one, which is the point.

The **firm-boundary prediction** fails on a continued shift towards fixed-fee and output-priced engagements in advisory work whose output cannot be checked at arm's length.

The **rent claim at §12** fails on the ratio of enterprise value in reference-data businesses to enterprise value in the professional-services firms that consume their data, which is observable and which the claim commits to.

Part II—The Verification Constraint

5. The asymmetry stated formally

Let a claim be any assertion whose economic value depends on its being true: a legal opinion, an audit sign-off, a diagnosis, a citation, a due diligence conclusion. Let g denote the marginal cost of generating such a claim to the standard at which a competent buyer cannot distinguish it from a true one, and let v denote the marginal cost of establishing whether it is in fact true.

Three propositions follow immediately, and none of them is controversial in isolation.

Proposition 1. For any claim, the socially efficient quantity is determined by the total cost $g + v$, not by g alone. A claim which is cheap to make and dear to check is not a cheap claim. Its cost has been relocated to the checker rather than reduced.

Proposition 2. The party bearing v is rarely the party bearing g. The producer of an assertion captures the saving on g; the recipient, the tribunal, the regulator or the counterparty bears v. This is a straightforward externality, and like most externalities it is invisible in the accounts of the party generating it.

Proposition 3. Where v cannot be borne—because the checker lacks the expertise, the access, or the time—the claim is not verified. It is accepted or rejected on some other basis: the reputation of the maker, the formatting of the document, the confidence of the delivery. None of these correlates with truth, and generative artificial intelligence has made all three cheap to imitate.

Define the **verification ratio** $= v / g$. The claim of this Part is that , and not the automatable share of tasks, is the parameter to watch in a technology that acts on g. The claim should be put no higher than that. Holding v fixed, is a monotone transform of $1/g$ and carries no information the automatable share does not; if v is allowed to move, does not determine the effect either, since a proportional fall in both leaves the ratio unchanged while transforming the economics. What does the work throughout this Part is the *level* of v, and is a way of keeping it in view rather than a new object.

Consider what happens as g falls towards zero with v held constant. rises without bound. Output volume rises, because production is now nearly free. Total verification cost rises in proportion to volume, because each new claim must be checked. And measured productivity—output of *verified* claims per unit of input—can fall, because the input required per verified claim now includes checking a growing volume of unverified ones. This is arithmetic rather than mismeasurement: the cost has moved from a stage the national accounts capture to a stage they do not. The point is a familiar one in the measurement literature, which has long recognised that output measures in the non-market and professional services are constructed from inputs and therefore record a quality change as no change at all.[16]

Call this **verification drag**. It is the mechanism by which a genuine and large improvement in production technology can register as a productivity slowdown, and it predicts something the current literature does not:

[16] Zvi Griliches, 'Productivity, R&D, and the Data Constraint' (1994) 84 *American Economic Review* 1.

that the sectors adopting the technology fastest will show the *worst* measured productivity performance, not the best, until verification is either automated or abandoned.

The alternative to bearing v deserves emphasis, because it is what actually happens. When verification cost exceeds what the checker can bear, the checker does not verify more slowly. The checker stops verifying and adopts a proxy. The proxy is cheap by construction and uncorrelated with truth by construction, and the equilibrium that results is one in which a great deal of confident, well-formatted, entirely unreliable work circulates at low cost and high volume. Part II §9 shows this happening in a setting where the outcome is a matter of public record.

6. Credence goods, and where the price is actually set

The economics of goods whose quality the buyer cannot assess is well developed and unusually well suited to this problem.

Nelson's distinction between search and experience goods, and Darby and Karni's addition of the credence category, established the taxonomy.[17] A search good can be inspected before purchase. An experience good reveals its quality on consumption. A credence good does neither: the buyer cannot determine quality even after consuming it, because doing so would require the same expertise that led them to buy in the first place. Darby and Karni's central result is that the optimal amount of fraud in such a market is not zero, because the cost of eliminating the last increment exceeds the loss it prevents.

Dulleck and Kerschbamer's survey provides the structure that matters here.[18] They show that efficient provision of credence goods requires some combination of three institutional conditions: **liability**, so that a provider who supplies the wrong service bears the loss; **verifiability**, so that what was supplied can be established after the fact; and **reputation with competition**, so that repeated dealing disciplines behaviour.

Their experimental work is where this paper must be careful, because it does not say what one might expect. Testing the institutions with 936 subjects, Dulleck, Kerschbamer and Sutter find that **liability has a crucial effect and verifiability at best a minor one**; reputation has little influence; and competition drives prices down without improving efficiency where liability is violated.[19] The efficient outcome is obtained by liability operating *without* the consumer verifying anything.

That result cuts against the loose claim that liability is simply a function of verifiability, and no such claim is made here. What survives, and what the argument needs, is narrower and follows from the mechanism rather than from the experiment. The liability that works in their design is liability for supplying the wrong service, which the seller knows and which the design makes enforceable by construction. The liability at issue in the markets this paper concerns is different in kind: it is liability established through litigation, on proof of breach, brought by a claimant who must first know that something was wrong. **That** form of liability is verifiability-dependent—not because liability is a weak institution, but because a claim that cannot be proved is not brought.

The distinction matters for Part V, and it helps rather than hinders. If liability works best when it attaches without requiring the buyer to verify, the correct response to expensive verification is not to demand more checking. It is to attach the duty where breach is observable without litigation—which is what the duty proposed at Part V does by requiring resolution at the moment of generation rather than proof of harm afterwards.

Emons reaches a compatible conclusion by a different route: the expert's incentive to defraud is governed by whether the client can subsequently establish what was needed, and where they cannot, competition alone does not discipline.[20]

[17] Phillip Nelson, 'Information and Consumer Behavior' (1970) 78 *Journal of Political Economy* 311; Darby and Karni (n 5).

[18] Uwe Dulleck and Rudolf Kerschbamer, 'On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods' (2006) 44 *Journal of Economic Literature* 5.

[19] Uwe Dulleck, Rudolf Kerschbamer and Matthias Sutter, 'The Economics of Credence Goods: An Experiment on the Role of Liability, Verifiability, Reputation, and Competition' (2011) 101 *American Economic Review* 526.

[20] Winand Emons, 'Credence Goods and Fraudulent Experts' (1997) 28 *The RAND Journal of Economics* 107.

The other two institutional conditions are also weaker than they appear. Reputation works, as Klein and Leffler and Shapiro showed, by making the present value of future custom exceed the one-period gain from cheating.[21] It therefore requires that cheating be detected with some probability, that detection be attributed to the cheat, and that the market observe the attribution. All three are verification problems. Tirole's analysis of collective reputation adds a further difficulty: where individual conduct cannot be separated from group reputation, the incentive to maintain quality is diluted across the group and free-riding is the equilibrium.[22] The credit rating agencies are the standing demonstration, and the literature on them states the position plainly about the mechanism: reputational discipline failed not because the agencies were unaware of their reputations but because the relevant conduct was not verifiable at the time it mattered.[23]

Licensing, the third response, is a verification technology and not an alternative to one. Leland's model of minimum quality standards is explicit that licensing raises average quality by excluding the low tail, at the cost of restricting supply and raising price.[24] What licensing certifies is the *provider*, on the theory that certifying each act of provision is too expensive. That theory holds only while the certified provider's output is a reliable signal of the certified provider's judgement. Where a licensed professional can generate output indistinguishable from their own considered work without applying any judgement to it, provider-level certification stops carrying information about act-level quality—and the entire architecture of professional regulation is provider-level.

This is the first substantive conclusion of the paper. **The three institutional responses to credence-goods failure—liability, reputation and licensing—all rest on verifiability, in the specific sense that each requires an outcome to be attributable to a supplier after the fact, and a technology that raises verification cost weakens all three at once.** Liability, reputation and licensing are not three independent safeguards. They are three applications of one, and they fail together.

6.1 The wider information-economics foundation

The credence-goods result is one branch of a larger body of work, and the branches converge on the same point.

Spence's signalling model established that where quality is unobservable, parties will expend resources on signals whose only function is to be costly, and that the resulting equilibrium is wasteful but stable.[25] Rothschild and Stiglitz showed that competitive markets under asymmetric information may have no equilibrium at all, and that screening devices distort the contracts actually offered.[26] Stiglitz's retrospective sets out how comprehensively these results displaced the assumption of costless information that classical price theory had rested on.[27]

The contracting literature reaches the same place from the incentive side. Holmström's analysis of moral hazard under imperfect observability establishes that the optimal contract depends on what is *observable*, not on what is done.[28] Grossman and Hart's treatment of the principal-agent problem formalises the resulting inefficiency.[29] Jensen and Meckling's account of agency costs, and Williamson's transaction-cost analysis of governance, both explain organisational form as a response to the cost of verifying performance.[30] The firm

[21] Benjamin Klein and Keith B Leffler, 'The Role of Market Forces in Assuring Contractual Performance' (1981) 89 *Journal of Political Economy* 615; Carl Shapiro, 'Premiums for High Quality Products as Returns to Reputations' (1983) 98 *Quarterly Journal of Economics* 659.

[22] Jean Tirole, 'A Theory of Collective Reputations (with Applications to the Persistence of Corruption and to Firm Quality)' (1996) 63 *The Review of Economic Studies* 1.

[23] See §8.3 below.

[24] Hayne E Leland, 'Quacks, Lemons, and Licensing: A Theory of Minimum Quality Standards' (1979) 87 *Journal of Political Economy* 1328.

[25] Michael Spence, 'Job Market Signaling' (1973) 87 *Quarterly Journal of Economics* 355.

[26] Michael Rothschild and Joseph Stiglitz, 'Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information' (1976) 90 *Quarterly Journal of Economics* 629.

[27] Joseph E Stiglitz, 'The Contributions of the Economics of Information to Twentieth Century Economics' (2000) 115 *Quarterly Journal of Economics* 1441.

[28] Bengt Holmström, 'Moral Hazard and Observability' (1979) 10 *The Bell Journal of Economics* 74.

[29] Sanford J Grossman and Oliver D Hart, 'An Analysis of the Principal-Agent Problem' (1983) 51 *Econometrica* 7.

[30] Michael C Jensen and William H Meckling, 'Theory of the firm: Managerial behavior, agency costs and ownership structure' (1976) 3 *Journal of Financial Economics* 305; Oliver E Williamson, 'Transaction-Cost Economics: The Governance of

exists, on this account, partly because verifying an employee is cheaper than verifying a contractor.

6.1A Costly state verification, and why voluntary disclosure does not solve this

Two results sit underneath everything in this Part and must be stated, because the argument is unintelligible without the first and indefensible without an answer to the second.

Townsend's costly state verification. A principal receives a report from an agent and can learn its truth only by paying a fixed cost. The optimal contract is not to verify always or never but to verify *stochastically*, or on a region of reported outcomes, and the resulting arrangement is second-best precisely because verification is expensive.[31] This is the formal statement of the problem this paper is about, forty-five years old, and the "verification ratio" of §5 is an expositional device for tracking one of its parameters rather than a new object. Two consequences follow that the paper relies on. Verification is optimally *incomplete*—so unchecked assertion is an equilibrium feature and not a failure of diligence. And the arrangement depends on the *cost*, so a change in that cost changes the contract, which is the whole mechanism claimed at §5 and §6.2.

Grossman and Milgrom's unravelling, and why it fails here. The natural objection is that no duty is needed because disclosure is self-enforcing. Where quality is verifiable at will and disclosure is costless, the highest type discloses to separate itself, the next-highest must then disclose to avoid being pooled with the residue, and the market unravels to full disclosure without any obligation at all.[32] If that held, Limb two of the duty would be otiose.

It does not hold, and the reasons identify precisely what a duty must supply. Unravelling requires that the receiver *knows the sender could have disclosed*; where the buyer does not know that verification state is a disclosable attribute, silence carries no adverse inference. It requires disclosure to be **verifiable**—a claim to have checked is worthless unless checkable, which is the problem restated one level up. It requires **common knowledge of the quality dimension**, and buyers of professional work do not observe "proportion of assertions resolved against a register" as a dimension at all. And it fails where the sender is uncertain what it holds, which is the position of a practitioner who does not know which of the assertions in a document were checked.

That diagnosis is constructive rather than defensive. A mandatory disclosure duty is warranted exactly where unravelling fails, and it should be designed to repair the specific failure: it must make the attribute **salient**, so silence carries inference; **standardised**, so the dimension is common knowledge; and, so far as the state of the art allows, **machine-checkable**. Limb two of the duty proposed at Part V is drafted to the first two expressly; the third it leaves to the state of the art, which is the point on which the Artificial Intelligence Act goes further, since art 50(2) requires a machine-readable marking.

A third result completes the picture. Lizzeri's analysis of certification intermediaries shows that an intermediary with a monopoly on verification will disclose the *minimum* consistent with capturing rent—typically that a threshold was met, not where in the distribution the subject sits.[33] That is a caution against the assumption at Part III §12.3 that verification rent accruing to register-holders is benign: the holder of the register has an incentive to reveal less than it knows.

6.2 A prediction the transaction-cost literature yields, applied to this shock

That body of work has an implication for the present problem which points in the opposite direction to the prevailing forecast. The mechanism is textbook and is not claimed as new: Barzel derives organisational form from measurement cost, Alchian and Demsetz from metering, and Holmström and Milgrom show that employment dominates contracting precisely where important dimensions of performance cannot be

Contractual Relations' (1979) 22 *The Journal of Law and Economics* 233.

[31] Townsend (n 6).

[32] Sanford J Grossman, 'The Informational Role of Warranties and Private Disclosure about Product Quality' (1981) 24 *The Journal of Law and Economics* 461; Paul R Milgrom, 'Good News and Bad News: Representation Theorems and Applications' (1981) 12 *The Bell Journal of Economics* 380.

[33] Alessandro Lizzeri, 'Information Revelation and Certification Intermediaries' (1999) 30 *The RAND Journal of Economics* 214.

measured; Baker and Hubbard demonstrate empirically that a technology which changed monitoring cost moved the boundary.[34] What is offered here is not the mechanism, nor the direction of the prediction, both of which are in the literature and one of which has been applied to this technology independently and earlier.[35] It is Proposition B below, which is the observable signature the other accounts do not name—though the *contract-form* result it rests on is also not new. Bajari and Tadelis show that as ex ante specification becomes harder, procurement shifts from fixed-price to cost-plus, which is the movement from output pricing to input pricing under exactly this pressure.[36] Proposition B is that result applied to advisory work under a shock that raises verification cost relative to production cost. It should also be said that the reported direction of travel is the opposite, and §4.2 sets out the survey evidence together with the three qualifications which stop it being decisive. The prediction is contrarian and may simply be wrong. It is offered because it is checkable, not because it is likely.

The Coasean question is why any activity is performed inside a firm rather than bought. Williamson's answer is that the boundary sits where the cost of governing a transaction internally falls below the cost of contracting for it, and a large component of that cost is the cost of establishing whether what was contracted for was in fact delivered.[37] Holmström's result is the same point in incentive terms: the optimal contract conditions on what is observable, so where performance is unobservable, contracting is inefficient and the activity is better placed under direct authority.[38] Employment is, on this account, partly a verification technology—a mechanism for obtaining output whose quality cannot be checked transaction by transaction.

If that is right, the boundary is a function of verification cost, and a technology that changes verification cost must move it. The prevailing expectation is that generative artificial intelligence causes **disaggregation**: tasks become automatable, automatable tasks become tradeable, and firms shrink towards a core of judgement. That expectation follows from the task framework and it is coherent on its own terms.

The verification analysis predicts the opposite for a specific and identifiable class of function. Where output can now be produced cheaply by a supplier and the buyer cannot verify it at arm's length, the transaction-cost logic says **internalise**. The firm should expand into precisely those functions whose output has become cheap to produce and remained expensive to check—not because it can produce them better, but because it can supervise them and cannot inspect them.

The prediction is sharper than a direction. Three testable propositions follow.

Proposition A. Outsourcing should retreat fastest in functions where the deliverable is an assertion whose truth is not observable on delivery—legal research, due diligence, market analysis, regulatory interpretation—and not in functions where the deliverable is verifiable on inspection, such as document processing or transcription, which should continue to externalise.

Proposition B. Where a function is retained externally despite being unverifiable, the contractual form should shift from output-based pricing towards input-based pricing or retainer, because paying for output requires measuring it. A move from fixed fees back to time-based billing in advisory work would be evidence for the mechanism; the reverse would be evidence against.

And the reverse is what is currently reported, which has to be said here rather than only at §4.2 where the claim's novelty is assessed. Trade surveys record a large and rising majority of United States firms offering alternative fee arrangements, with flat fees the commonest form, and the trade press expects the technology to accelerate that shift rather than reverse it.[39] Proposition B is therefore currently running against the observed direction of travel. Two things may be true at once—the aggregate mix may move towards fixed fees while the specific class of work this section identifies moves the other way—but that is a defence and not evidence, and the honest statement is that the proposition is presently unsupported and stated so that it can be tested rather than because it has been.

[34] Barzel, Alchian and Demsetz, Holmström and Milgrom, and Baker and Hubbard (n 9).
[35] Hydari and Muzaffar (n 13).
[36] Bajari and Tadelis (n 14).
[37] Williamson (n 30).
[38] Holmström (n 28).
[39] See §4.2 and the sources there (n 15).

Proposition C. The effect should be strongest where liability for the assertion sits with the buyer rather than the supplier, since a buyer who bears the consequence of an unverifiable claim has the strongest reason to bring its production under supervision. This is the point at which the analysis rejoins Part V: the allocation of liability determines the organisational response, so a verification duty placed on the supplier would reverse the prediction.

The proposition is not that firms will grow overall. Functions with verifiable output should continue to leave. The claim is about a **recomposition**—externalising what can be inspected, internalising what can only be supervised—and the observable signature is a change in the *mix* of what is bought in, not in its volume.

One caution. Teece's complementary-assets analysis predicts that returns accrue to whoever holds the assets an innovation needs, and would also predict integration in some of these cases.[40] The predictions overlap and would need to be distinguished empirically: the complementary-assets account predicts integration towards the scarce asset, the verification account predicts integration towards the unverifiable function, and these are different functions in most firms.

Sanchirico's analysis of evidence law supplies the final piece and is one the argument in Part V relies on.[41] His argument is that evidentiary rules are best understood not as devices for finding truth in the instant case but as devices for structuring incentives before it—the rules determine what parties do in anticipation of proof. That reframing is what makes an evidential presumption a policy instrument rather than a procedural detail, and it is why the duty proposed at Part V is framed as an obligation to resolve and disclose rather than as a prohibition on emission.

7. What the growth literature measures, and what it does not

The task-based framework that dominates the analysis of automation is a genuine intellectual achievement and is not in dispute here. Autor, Levy and Murnane established that technology substitutes for routine tasks and complements non-routine ones, which reoriented a literature that had been arguing about skill in the abstract.[42] Acemoglu and Restrepo formalised the displacement and reinstatement effects and gave the framework its modern shape.[43] Applied to generative artificial intelligence, the same apparatus produces Acemoglu's modest aggregate estimate: the exposed task share, multiplied by the achievable saving within those tasks, multiplied by the pass-through, yields a total factor productivity effect of well under one per cent over a decade.[44]

Against this, the transformative accounts model the technology as an input to idea production rather than to goods production, which changes the growth rate rather than the level.[45] Nordhaus tests the singularity hypothesis against the data and rejects it on current evidence.[46] Bloom, Jones, Van Reenen and Webb document the countervailing force: research productivity is falling sharply across fields, so that constant progress requires ever-larger research inputs.[47]

The empirical micro-evidence is more favourable than either macro position. Noy and Zhang find substantial time savings and quality improvement on mid-level writing tasks.[48] Brynjolfsson, Li and Raymond find large productivity gains among customer-support agents, concentrated among the least experienced.[49]

[40] Teece (n 8).

[41] Chris William Sanchirico, 'Character Evidence and the Object of Trial' (2001) 101 *Columbia Law Review* 1227.

[42] David H Autor, Frank Levy and Richard J Murnane, 'The Skill Content of Recent Technological Change: An Empirical Exploration' (2003) 118 *Quarterly Journal of Economics* 1279.

[43] Daron Acemoglu and Pascual Restrepo, 'The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment' (2018) 108 *American Economic Review* 1488; Daron Acemoglu and Pascual Restrepo, 'Automation and New Tasks: How Technology Displaces and Reinstates Labor' (2019) 33 *Journal of Economic Perspectives* 3.

[44] Acemoglu (n 1).

[45] See Nicholas Crafts, 'Artificial intelligence as a general-purpose technology: an historical perspective' (2021) 37 *Oxford Review of Economic Policy* 521, reviewing the transformative accounts.

[46] William D Nordhaus, 'Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth' (2021) 13 *American Economic Journal: Macroeconomics* 299.

[47] Nicholas Bloom and others, 'Are Ideas Getting Harder to Find?' (2020) 110 *American Economic Review* 1104.

[48] Noy and Zhang (n 4).

[49] Brynjolfsson, Li and Raymond (n 4).

Notice what every one of these studies has in common. Each measures the production of output. None measures the cost of establishing that the output is correct—and in the experimental settings, correctness is either observable at low cost (a resolved support ticket, a graded writing task) or is assumed. That is a reasonable design choice for the questions those papers ask. It is fatal when the results are extrapolated to sectors where correctness is precisely what cannot be cheaply observed.

Kabir and colleagues provide the exception that proves the point.[50] Examining answers generated for programming questions, they find that a little over half contained incorrect information, that the incorrect answers were nonetheless preferred by users in a substantial share of cases, and that users failed to identify the errors when the answer was articulate and well-structured. Its interest here is not the programming context but the measurement: it records v under conditions in which g has collapsed, and the finding is that the checker's error rate rises with the fluency of the claim.

The technical literature explains part of why this is structural rather than transitional, and why one part of it is not. Kalai and Vempala prove that a calibrated language model must hallucinate on *arbitrary* facts: where a fact appears once in the training data and its veracity cannot be determined from that data, the rate of fabrication is bounded below by a Good-Turing quantity, so suppressing it entirely requires giving up calibration.[51] That is a result about a class of system rather than a particular one.

Their result is expressly limited, and the limit is the more useful half here. They observe that there is no statistical reason for pretraining to produce hallucination on facts appearing more than once in the training data—giving *references to publications such as articles and books* as their example—nor on systematic facts, and that architecture and post-training may mitigate both. Fabricated citation is therefore **not** shown to be irreducible. It is the class of error most open to mechanisation, which is why an instrument that resolves citations against a register can be built at all, and is why Part V proposes that one be required. The irreducible component of verification cost lies elsewhere: not in whether an authority exists, but in whether it supports the proposition it is cited for, which no register can answer. Ji and colleagues' survey documents the same phenomenon across generation tasks.[52] Bender and colleagues anticipated the mechanism: a system optimised to produce text that patterns like true text will produce text that patterns like true text, whether or not it is.[53] Dahl and colleagues measure the consequence in the domain where it is most consequential, finding hallucination rates on legal queries that range from substantial to alarming depending on the specificity of the question, and—critically—highest on questions about lower courts and less prominent jurisdictions, which is exactly where independent verification is hardest.[54]

The last point inverts the intuition. Fabrication is not distributed randomly across queries; it concentrates where source material is sparse, which is where checking is most expensive. That is not a coincidence: the same sparsity of source material that makes a question hard to answer reliably makes the answer hard to check. g and v are correlated in the wrong direction.

7.1 The labour-market literature has the same blind spot

The applied work on exposure is now substantial and it inherits the specification wholesale. Felten, Raj and Seamans construct occupational exposure measures from the alignment of model capabilities to task descriptions.[55] Webb infers exposure from the overlap between patent text and job descriptions.[56] Eloundou and colleagues estimate that around 80 per cent of the United States workforce has at least ten per cent of

[50] Samia Kabir and others, 'Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions' (2024) *Proceedings of the CHI Conference on Human Factors in Computing Systems* 1.

[51] Adam Tauman Kalai and Santosh S Vempala, 'Calibrated Language Models Must Hallucinate' (2024) *Proceedings of the 56th Annual ACM Symposium on Theory of Computing* 160.

[52] Ziwei Ji and others, 'Survey of Hallucination in Natural Language Generation' (2023) 55 *ACM Computing Surveys* 1.

[53] Emily M Bender and others, 'On the Dangers of Stochastic Parrots' (2021) *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 610.

[54] Matthew Dahl and others, 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models' (2024) 16 *Journal of Legal Analysis* 64.

[55] Edward Felten, Manav Raj and Robert Seamans, 'Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses' (2021) 42 *Strategic Management Journal* 2195.

[56] Michael Webb, 'The Impact of Artificial Intelligence on the Labor Market' (2019) SSRN 3482150.

its tasks affected.[57] These follow the tradition running from Frey and Osborne's computerisation estimates through Arntz, Gregory and Zierahn's task-based correction of them.[58]

Every one of these measures *exposure of tasks to substitution*. None measures whether the substituted output is checkable, and the distinction is not pedantic: two occupations with identical exposure scores will have entirely different economics if one produces claims whose truth is observable on delivery and the other does not. A radiographer's read is checked against the eventual pathology; a compliance opinion is checked against nothing until something goes wrong, if then.

The firm-level evidence is more interesting, because on its face it points the other way. Babina and colleagues find AI-investing firms growing faster and innovating more.[59] That is prima facie contrary evidence for the prediction at §5 that the sectors adopting fastest will show the *worst* measured productivity performance, and it should be met rather than absorbed. The answer is that firm-level growth and sector-level measured productivity are different objects and can move in opposite directions: a firm that captures share by producing more output per unit of input grows, while the sector's measured productivity falls if the additional output requires verification labour that is counted as input and produces no additional measured value. The two are reconciled if the gains are *distributional within the sector* rather than aggregate—which is the same claim §12 makes about rent—and that is the discriminating test. If AI-investing firms grow while sector output per hour stagnates, the account here is supported. If sector productivity rises with adoption, the account is not falsified outright but is pushed onto its second branch: measured productivity also rises where verification is *abandoned* rather than performed, because the checking labour is never counted as input while the additional output is counted. The two branches are distinguished by whether the error rate in the sector's output rises at the same time, which is observable in litigation, restatement and retraction data and is where §9 to §12 look. Bessen's argument that demand elasticity determines whether automation destroys or creates employment identifies a genuine mechanism, and Autor's proposal that the technology could rebuild middle-skill work rests on the claim that it lets non-experts exercise expert judgement.[60] That last claim is precisely where verification cost bites: allowing a non-expert to *produce* expert output is not the same as allowing them to *check* it, and it is the checking that the expert was being paid for. Otis and colleagues find the gains from generative assistance distributed unevenly across entrepreneurs in a way consistent with this—the benefit accrues to those already able to evaluate the output.[61]

None of this is a criticism of the studies, which answer the questions they set. It is an observation that the aggregate conclusion drawn from them—that exposure predicts economic effect—omits the variable that determines whether exposed output has value.

8. The counterarguments

Five objections are addressed in this section: the productivity J-curve, mismeasurement, the automation of verification, reputational and liability sorting, and Baumol's cost disease.

8.1 The J-curve: this is an adjustment lag, not a constraint

Brynjolfsson, Rock and Syverson's productivity J-curve is the most serious version of the "wait" argument.[62] General-purpose technologies require complementary intangible investment—reorganisation, retraining, process redesign—that is expensed rather than capitalised, so measured productivity is understated during the

[57] Tyna Eloundou and others, 'GPTs are GPTs: Labor market impact potential of LLMs' (2024) 384 *Science* 1306.

[58] Carl Benedikt Frey and Michael A Osborne, 'The future of employment: How susceptible are jobs to computerisation?' (2017) 114 *Technological Forecasting and Social Change* 254; Melanie Arntz, Terry Gregory and Ulrich Zierahn, 'Revisiting the risk of automation' (2017) 159 *Economics Letters* 157.

[59] Tania Babina and others, 'Artificial intelligence, firm growth, and product innovation' (2024) 151 *Journal of Financial Economics* 103745.

[60] James Bessen, 'AI and Jobs: The Role of Demand' (2018) NBER Working Paper 24235; David Autor, 'Applying AI to Rebuild Middle Class Jobs' (2024) NBER Working Paper 32140.

[61] Nicholas Otis and others, 'The Uneven Impact of Generative AI on Entrepreneurial Performance' (2024) 2024 *Academy of Management Proceedings* 21128.

[62] Erik Brynjolfsson, Daniel Rock and Chad Syverson, 'The Productivity J-Curve: How Intangibles Complement General Purpose Technologies' (2021) 13 *American Economic Journal: Macroeconomics* 333.

build-out and overstated later. On this account the current picture is simply the trough, and David's account of the dynamo supplies the historical template.[63] Crafts places artificial intelligence in that lineage explicitly.[64]

This objection is correct about the general mechanism and wrong about the present case, for a reason internal to its own logic. The J-curve describes a *transitional* accumulation of complementary capital which subsequently yields returns. Verification cost is not complementary capital. It is a recurring per-unit cost of the output, and it recurs at the same rate on the millionth claim as on the first. Electrification required factories to be rebuilt once. Verification requires each assertion to be checked every time it is made.

One qualification is required and it is the objection this argument must answer, because the present paper supplies the counter-example. A citation *register* is a stock: it is built once, at an up-front cost, and thereafter lowers the per-claim cost of establishing that an authority exists. To that extent verification does amortise, and §20.4 records that the author built such a register. The claim here is therefore not that verification cost as a whole is a flow. It is that verification cost has two components which behave differently. Establishing that a cited authority *exists* is mechanisable and does amortise; establishing that it *supports the proposition it is cited for* is not, and does not, because it is a fresh judgement about a fresh pairing of authority and claim every time the pairing is made. §7 gives the reason the second component is the larger one: the errors that survive mechanisation are concentrated where sources are sparse, which is where checking was already dearest. Part IV §16.3 tests the division on a real corpus by splitting the retraction series on what similarity and image screening reaches, and finds the rise concentrated almost entirely outside it. The stock/flow argument is made about the second component, and the falsification test at §4.3 should be read as applying to it.

The distinction is testable. Complementary-investment stories predict that the productivity cost falls as the stock of intangible capital accumulates. The verification account predicts that it scales with volume and does not fall with experience—and that it worsens, because as fabrication becomes rarer per claim while volume rises, the expected value of checking any particular claim falls while the cost of missing the rare bad one stays constant. That is the classic structure of a monitoring problem with a low base rate, and low base rates make monitoring harder, not easier.

8.1A The measurement problem is real but does not save the objection

A related defence holds that the difficulty is one of measurement: national accounts capture poorly what is being produced, so the effect exists and is not being seen.

The case is respectable. Solow's residual has always absorbed everything not otherwise explained, and its interpretation has been contested since it was introduced.[65] Syverson's assessment of mismeasurement explanations for the productivity slowdown examines the candidates carefully and finds them insufficient to account for the observed gap—the required mismeasurement is larger than the plausible bounds.[66] Gordon's account attributes the slowdown to the exhaustion of a particular class of innovation rather than to measurement.[67] Brynjolfsson, Rock and Syverson's treatment of the artificial-intelligence paradox canvasses the four candidate explanations—false hopes, mismeasurement, redistribution, and implementation lags—and favours the last.[68]

The verification account is compatible with all of these and reducible to none. Where it differs is in identifying a cost that is genuinely *incurred* rather than merely unmeasured. Time spent checking a fabricated citation is real labour producing no output; it appears in the input side of the accounts and nowhere on the output side, which is not mismeasurement but correct measurement of a genuine loss. Agrawal, Gans and Goldfarb's

[63] Discussed in Crafts (n 45).

[64] Crafts (n 45).

[65] Robert M Solow, 'Technical Change and the Aggregate Production Function' (1957) 39 *Review of Economics and Statistics* 312.

[66] Chad Syverson, 'Challenges to Mismeasurement Explanations for the US Productivity Slowdown' (2017) 31 *Journal of Economic Perspectives* 165.

[67] Robert Gordon, *The Rise and Fall of American Growth* (Princeton University Press 2016).

[68] Erik Brynjolfsson, Daniel Rock and Chad Syverson, 'Artificial Intelligence and the Modern Productivity Paradox' in *The Economics of Artificial Intelligence* (University of Chicago Press 2019) 23.

distinction between prediction and judgement is the closest existing framing: they argue that machine prediction raises the value of complementary human judgement.[69] The argument here is narrower and less optimistic—the complementary activity is not judgement about what to do, but verification of what has been asserted, and unlike judgement it produces nothing new.

Nor is misallocation the explanation. Hsieh and Klenow's work established how large the aggregate consequences of resource misallocation can be, and it is a candidate for any productivity puzzle.[70] But misallocation describes resources deployed in the wrong place; verification drag describes resources deployed in the right place on a task that has been created by the technology itself. The two are separable, and Aghion and Howitt's creative-destruction framework is the reason the distinction matters: their mechanism is that new activity displaces old, whereas verification is new activity that displaces nothing.[71]

8.2 Verification will be automated too

The natural rejoinder is that the same technology that generates claims can check them, so v will fall to meet g. This is the objection that most people reach for and it is weaker than it appears, for three reasons.

First, the empirical record is poor. Detection tools for machine-generated text perform badly and inconsistently, and the systematic evaluations find them unreliable enough that institutional reliance on them is unwarranted.[72] Worse, they are biased: Liang and colleagues show that detectors systematically misclassify writing by non-native English speakers as machine-generated, which converts a verification tool into a discrimination mechanism.[73]

Second, checking a claim against the world is not the same operation as generating a claim that resembles true claims. A system that checks a citation must possess an index of what exists. That index is not a model capability; it is a corpus, and corpora are expensive, jurisdictionally fragmented, and in the case of law frequently unavailable for computational use without permission. This paper's own apparatus makes the point concretely: verifying English authority required assembling an index of 75,837 judgments under a specific licence, and that index still cannot confirm a House of Lords citation, because the corpus does not reach the House of Lords. The limitation is an absence of data rather than a modelling failure, and capability does not substitute for a corpus.

Third, where automated verification does work it works by *reference to an authoritative register*, which relocates rather than removes the problem. The register must be maintained, funded, and trusted. That is a governance question, and it is the question Part V addresses.

8.3 Reputation and liability will re-price the risk

If verification is expensive, the argument runs, buyers will pay a premium for providers who are known to verify, and the market will resolve itself. This is Shapiro's mechanism and Klein and Leffler's, and it is sound where its conditions hold.[74]

Its conditions do not hold here. Reputation disciplines conduct that is observed with some probability. The whole difficulty is that unverified work is *indistinguishable from verified work at the point of purchase and frequently afterwards*. A provider who verifies and a provider who does not produce identical-looking output; the difference emerges only when a specific claim is challenged, which is rare, late, and idiosyncratic. That is precisely the condition under which Tirole's collective reputation result applies: individual conduct is not separable, the reputational externality is shared, and free-riding dominates.[75]

[69] Ajay Agrawal, Joshua S Gans and Avi Goldfarb, 'Prediction, Judgment, and Complexity' in *The Economics of Artificial Intelligence* (University of Chicago Press 2019) 89.

[70] Chang-Tai Hsieh and Peter J Klenow, 'Misallocation and Manufacturing TFP in China and India' (2009) 124 *Quarterly Journal of Economics* 1403.

[71] Philippe Aghion and Peter Howitt, 'A Model of Growth Through Creative Destruction' (1992) 60 *Econometrica* 323.

[72] Debora Weber-Wulff and others, 'Testing of detection tools for AI-generated text' (2023) 19 *International Journal for Educational Integrity* 26.

[73] Weixin Liang and others, 'GPT detectors are biased against non-native English writers' (2023) 4 *Patterns* 100779.

[74] Klein and Leffler (n 21); Shapiro (n 21).

[75] Tirole (n 22).

The credit ratings literature is the natural test case and it does not support the optimistic view. Bolton, Freixas and Shapiro model the ratings market and find that reputational concerns are insufficient to prevent inflation where the issuer selects the rater and the error is revealed only in a downturn.[76] Mathis, McAndrews and Rochet reach the same conclusion.[77] The auditing evidence is stronger still: DeFond and Zhang's review documents the persistent difficulty of demonstrating that audit quality is observable to the market at all.[78]

8.4 This is simply Baumol

The objection considered in this subsection is the strongest of the five, and it is that none of this is new. Baumol's cost disease already describes sectors in which productivity growth is structurally slow because the output is labour-intensive and resists mechanisation, and law, medicine, education and audit are its canonical examples.[79] On this view the verification constraint is a re-description of a known phenomenon in fashionable terms.

The objection is well aimed and requires a precise answer. Baumol's mechanism is that the *productive* stage of certain services resists mechanisation, so unit costs rise relative to sectors where it does not. The mechanism here is the opposite: the productive stage has been mechanised, thoroughly and suddenly, while a different stage has not. These produce different predictions and can be distinguished empirically.

Baumol predicts that output per worker in the affected sector stays flat while wages rise with the general level. The verification account predicts that *output volume per worker rises sharply* while *verified* output per worker stays flat or falls, and that the wedge between them shows up as growth in assurance, review and quality-control functions—headcount that produces no new claims and exists only to check existing ones. Under Baumol, the sector's cost structure shifts towards its existing labour. Under verification drag, it shifts towards a *new category* of labour that did not previously exist at scale.

There is a second distinguishing prediction. Baumol's disease is symmetric across firms in the affected sector. Verification drag is not: it falls hardest on the party that cannot pass the cost on. Where a verification duty is legally allocated—as it is to the practitioner before a court, and as it is not to the vendor of the tool—the drag concentrates on the allocated party, and the observable consequence is a redistribution of margin rather than a general decline in it. Part V argues that this allocation is currently in the wrong place and that the common law already contains the reasoning to move it.

9. The courts as a natural experiment

Theory about unobservable quality is difficult to test, because unobservable quality is unobserved. Litigation provides an unusual exception. When a claim is advanced in court it is subjected to adversarial verification by a party with an incentive to find the flaw, and when it fails the failure is recorded in a public document.

This section aims to establish, from four decided cases across three jurisdictions, that the asymmetry set out at §5 is not a theoretical construction but an observed pattern with a consistent shape: the fabricated authority is cheap to produce, expensive to detect, and detected only where an adversary or the court was funded to look. We take the cases in order of the position the maker occupied, from represented litigant to litigant in person, because the pattern is what survives that variation.

In *Mata v Avianca Inc* the Southern District of New York sanctioned counsel who had filed an affirmation citing six wholly fabricated decisions.[80] What makes the judgment analytically useful is not the fabrication but Castel J's account of *why* it survived. The invented decisions attributed holdings to real, identifiable judges; they carried plausible reporter citations; they contained internal quotations. Each of those features

[76] Patrick Bolton, Xavier Freixas and Joel Shapiro, 'The Credit Ratings Game' (2012) 67 *The Journal of Finance* 85.

[77] Jérôme Mathis, James McAndrews and Jean-Charles Rochet, 'Rating the raters: Are reputation concerns powerful enough to discipline rating agencies?' (2009) 56 *Journal of Monetary Economics* 657.

[78] Mark DeFond and Jieying Zhang, 'A review of archival auditing research' (2014) 58 *Journal of Accounting and Economics* 275.

[79] William J Baumol, 'Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis' (1967) 57 *American Economic Review* 415.

[80] *Mata v Avianca Inc* 678 F Supp 3d 443 (SDNY 2023). Sanctions were imposed under the Federal Rules of Civil Procedure, r 11(b)(2) and (c).

raises g only trivially—it costs no more to fabricate a citation with a judge's name than without—while raising v substantially, because checking now requires consulting the reporter rather than noticing an implausibility on the face of the document.

Park v Kim extends the point to the appellate level, where the Second Circuit referred counsel to its Grievance Panel after a reply brief cited an authority that could not be located.[81] The court's formulation of the duty is instructive: the failure was not the use of the tool but the failure to inquire into the validity of the argument advanced. That is a verification duty, stated as such.

Harber v Commissioners for HMRC is the purest instance, because there was no professional to blame.[82] A litigant in person advanced nine First-tier Tribunal decisions that did not exist. The Tribunal found—and this is the part that matters—that she was unaware they were fabricated and did not know how to check an authority using the Tribunal's own website, BAILII, or any other source. Here g was zero and v was, for this litigant, effectively infinite. The cost of checking fell entirely on the Tribunal, which is to say on the public.

The most authoritative English statement so far is *R (Ayinde) v London Borough of Haringey; Al-Haroun v Qatar National Bank QPSC*, where a Divisional Court presided over by the President of the King's Bench Division exercised the *Hamid* jurisdiction over two matters involving fabricated or unverifiable authority.[83] It is a first-instance decision of a Divisional Court and should not be described as settling the law, though it is likely to be followed; the referral arose from Ritchie J's judgment at first instance, in which the fabricated authorities were identified.[84] The *Hamid* jurisdiction is the court's inherent power to police the duties practitioners owe it, and its invocation here is doctrinally significant: the court treated the failure to verify as a matter going to the administration of justice rather than to the competence of the individual practitioner.[85]

Four observations follow, and they are the evidential basis for the rest of this paper.

First, the failures are not confined to the unrepresented, the junior or the careless. They occur among admitted practitioners in commercial matters before senior courts.

Second, they were caught by *adversarial* verification, not by professional self-checking. In each case an opponent or the court did the work the filing party had not. This is verification supplied by an institution that happens to be structured to supply it, and there is no equivalent institution for an audit memorandum, a compliance opinion, a due diligence report or a peer-reviewed paper.

Third, the visible cases are, necessarily, the ones where verification occurred. The unobserved population is the set of assertions that were never checked because no adversary existed. There is no basis for assuming that population is smaller or cleaner. Every reason suggests the reverse: adversarial process is expensive and rare, and most professional assertion is never subjected to it.

Fourth, the legal response has so far been directed at the practitioner. That is understandable—the practitioner is before the court—but it allocates the entire verification cost to the party least able to reduce it, and none of it to the parties who captured the saving on g. Part V takes up that allocation, and concludes at §20.1A that the common law will not move it: the duty proposed there is statutory for that reason.

Part III—The Same Failure in Two Other Markets

10. Why two more cases are needed

Part II established the mechanism and Part II §9 tested it in litigation, where adversarial process supplies verification and the failures are recorded. That is a strong test and a narrow one, and it invites an obvious

[81] *Park v Kim* 91 F 4th 610 (2d Cir 2024). The referral was made under the Rules of the United States Court of Appeals for the Second Circuit, r 46.2.

[82] *Harber v Commissioners for His Majesty's Revenue and Customs* [2023] UKFTT 1007 (TC) [24]–[36].

[83] *R (Ayinde) v London Borough of Haringey; Al-Haroun v Qatar National Bank QPSC* [2025] EWHC 1383 (Admin) [6]–[9], [58]–[68].

[84] *R (Ayinde) v London Borough of Haringey* [2025] EWHC 1040 (Admin) (Ritchie J).

[85] *R (Hamid) v Secretary of State for the Home Department* [2012] EWHC 3070 (Admin).

objection: courts are unusual, litigants are adversaries, and nothing follows about markets where nobody is paid to find the flaw.

The objection is right that courts are unusual, and wrong about the direction of the inference. Litigation is the setting in which verification is *most* available, because an opponent is funded to supply it. If assertions fail there, the presumption for settings without an adversary is worse rather than better. This Part examines two such settings—scholarly publication and third-party assurance—both of which are credence markets, both of which have long-standing verification institutions, and both of which were exhibiting the predicted failure before generative artificial intelligence arrived. That last point matters: it means the mechanism is not an artefact of the technology, and the technology is an accelerant of something already under way.

11. Scholarly publication

This section examines scholarly publication as the second of the three cases, taking the verification institution first, its condition before 2022 second, and what the technology has changed third. Our purpose is to establish that the failure identified in Part II is not peculiar to litigation, and that its severity varies with who funds the verifier.

11.1 The verification institution and its condition before 2022

Peer review is a verification technology in the exact sense used here: a mechanism for establishing quality that the consumer of the good cannot assess directly. Its condition, before generative models were widely available, was already poor, and the evidence is unusually good because the field has spent two decades measuring itself.

Ioannidis's analysis of the statistical structure of published findings concluded that most are false, on the arithmetic of prior probability, power and researcher degrees of freedom.[86] The empirical replication programmes broadly confirmed the prediction: the Open Science Collaboration's reproduction of one hundred psychological studies recovered a substantially smaller effect than originally reported in the majority of cases, and Camerer and colleagues' replication of social science experiments published in *Nature* and *Science* found a similar attenuation.[87] Economics is not exempt: Brodeur, Cook and Heyes document the distributional signature of p-hacking across causal-inference methods, and Christensen and Miguel's review sets out the credibility problem in the discipline directly.[88] Simmons, Nelson and Simonsohn had already shown that undisclosed flexibility in analysis makes it trivial to obtain a significant result from nothing.[89] Fanelli's meta-analysis of self-report surveys found admitted fabrication and falsification at rates that are small in absolute terms and alarming given the incentive to under-report.[90]

Notice what all of this establishes. It is not that scholars are dishonest. It is that **the verification institution was not verifying**, and had not been for some time, for reasons that are structural: reviewers are unpaid, anonymous, time-constrained, and—critically—do not re-run the analysis, because re-running the analysis costs more than reviewing the paper.

11.2 What the market for retractions reveals

The economics of retraction supplies the price signal, and the two measured effects run in different directions. Lu, Jin, Uzzi and Jones find that a retraction imposes a citation penalty on the retracted authors' *own* prior work, falling by about seven per cent a year per earlier publication against matched controls, and propagating

[86] John P A Ioannidis, 'Why Most Published Research Findings Are False' (2005) 2 *PLoS Medicine* e124.

[87] Open Science Collaboration, 'Estimating the reproducibility of psychological science' (2015) 349 *Science* aac4716; Colin F Camerer and others, 'Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015' (2018) 2 *Nature Human Behaviour* 637.

[88] Abel Brodeur and others, 'Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics' (2020) 110 *American Economic Review* 3634; Garret Christensen and Edward Miguel, 'Transparency, Reproducibility, and the Credibility of Economics Research' (2018) 56 *Journal of Economic Literature* 920.

[89] Joseph P Simmons, Leif D Nelson and Uri Simonsohn, 'False-Positive Psychology' (2011) 22 *Psychological Science* 1359.

[90] Daniele Fanelli, 'How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data' (2009) 4 *PLoS ONE* e5738.

up to four degrees through the authors' citation network—a penalty that disappears where the authors self-report the error.[91] Azoulay, Furman, Krieger and Murray find that the penalty also reaches work by *other* authors: articles intellectually proximate to a retracted paper, published before the retraction, suffer a lasting five to ten per cent decline in citation rate, with the effect larger where the retraction involved misconduct rather than honest error, and with new entry and funding into the field falling as well.[92] Furman, Jensen and Murray document the role of retraction in governing knowledge within scientific communities.[93]

The mechanism these papers describe is precisely Tirole's collective reputation, operating on a group whose individual conduct is not separable.[94] The penalty falls on the field rather than the individual because the field is what the reader can observe. That is efficient only if the field can police itself, and the replication literature is the record of it failing to.

11.3 What has changed since

Onto that structure the collapse of g has been introduced.

Else reported in *Nature* that abstracts generated by a language model were not reliably distinguished from genuine ones by domain scientists.[95] Van Noorden's survey of the fake-paper problem estimates a volume of fabricated literature that is difficult to reconcile with the assumption that peer review filters it.[96] Candal-Pedreira and colleagues' cross-sectional study of retracted paper-mill output documents the industrial character of the production: papers manufactured at scale, sold, and published.[97] The verification response has not kept pace, and the detection tools are unreliable and biased in the ways set out at Part II §8.2.[98]

The prediction of Part II holds exactly. Volume rises. Verification cost per unit is unchanged. The verification institution, already unable to bear the cost, substitutes proxies—journal prestige, author affiliation, formatting, fluency—every one of which the technology imitates at zero marginal cost. The retraction corpus grows, which looks like the system working and is better read as the visible fraction of an unobserved population, on the same reasoning as Part II §9.

11.4 Why this matters beyond the academy

Scholarly literature is an input to law, regulation, medicine and policy. Where a court receives expert evidence resting on published findings, where a regulator adopts a threshold from a literature, or where a compliance function relies on a published methodology, the verification failure propagates into settings that have no capacity to detect it. This is the network-contagion structure applied to claims rather than to balance sheets: a failure in a node with many downstream dependencies does not stay in the node.[99]

12. Third-party assurance

This section turns to assurance, and to the case in which the verifier is paid by the party whose assertions it verifies. We take audit first, because the failure there is documented; certification and supply-chain monitoring follow, because they exhibit the same structure without the statutory backstop.

[91] Susan Feng Lu and others, 'The Retraction Penalty: Evidence from the Web of Science' (2013) 3 *Scientific Reports* 3146.

[92] Pierre Azoulay, Jeffrey L Furman, Joshua L Krieger and Fiona Murray, 'Retractions' (2015) 97 *Review of Economics and Statistics* 1118.

[93] Jeffrey L Furman, Kyle Jensen and Fiona Murray, 'Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine' (2012) 41 *Research Policy* 276.

[94] Tirole (n 22).

[95] Holly Else, 'Abstracts written by ChatGPT fool scientists' (2023) 613 *Nature* 423.

[96] Richard Van Noorden, 'How big is science's fake-paper problem?' (2023) 623 *Nature* 466.

[97] Cristina Candal-Pedreira and others, 'Retracted papers originating from paper mills: cross sectional study' (2022) 379 *BMJ* e071517.

[98] Weber-Wulff and others (n 72); Liang and others (n 73).

[99] The transmission structure is the same as that formalised for production and financial networks: see Daron Acemoglu and others, 'The Network Origins of Aggregate Fluctuations' (2012) 80 *Econometrica* 1977; Matthew Elliott, Benjamin Golub and Matthew O Jackson, 'Financial Networks and Contagion' (2014) 104 *American Economic Review* 3115.

12.1 The audit case

Audit is the purest institutional response to the credence problem: a third party, paid by the audited entity, attesting to the reliability of information for the benefit of parties who cannot verify it themselves. Its economics have been examined for four decades and the findings are not encouraging even before the technology is considered.

DeFond and Zhang's review of the archival literature documents the persistent difficulty of establishing that audit quality is observable to the market at all.[100] Dye's analysis shows the interaction between auditing standards, legal liability and auditor wealth: the incentive to audit well depends on the auditor having assets to lose and on breach being demonstrable.[101] Both conditions are verification conditions.

The structural difficulty is that the assurance provider is selected and paid by the party whose assertions are being assured, and that the quality of the assurance is itself a credence good to the ultimate user. This is the ratings problem in a different market, and the ratings literature is clear that reputational discipline does not resolve it.[102]

12.2 Certification and supply-chain monitoring

The same structure appears wherever a private standard substitutes for direct inspection. Terlaak's analysis of certified management standards asks whether they produce order without law and finds the answer conditional on monitoring quality.[103] Potoski and Prakash's work on ISO 14001 treats certification as a club good whose value depends on the credibility of exclusion.[104] Short, Toffel and Hugill's study of supply-chain monitoring finds that the effectiveness of an audit depends heavily on the auditor's independence, training and incentives—that is, on the cost that has been spent on verification.[105]

Each of these literatures reaches the same place from a different direction: certification transmits information only to the extent that someone has borne the cost of establishing what is certified, and the party bearing that cost is systematically not the party benefiting from the certification.

The standards literature explains why the private response tends to under-supply verification. Farrell and Shapiro's study of high-definition television standard setting, and Besen and Farrell's account of the strategies firms adopt in standardisation, both establish that standards are chosen for competitive advantage and not for informational quality.[106] Locke and colleagues' work on corporate codes of conduct found that monitoring alone changed little; what changed outcomes was reorganisation of the work itself.[107] Kleiner and Krueger's measurement of occupational licensing quantifies the cost of the provider-level alternative—a wage premium accruing to the licensed with contested evidence of quality gain.[108] Each is a different route to the same conclusion: verification is a public good within the market that supplies it, and is supplied accordingly.

12.3 The prediction for assurance markets

If Part II is right, the effect of collapsing g on assurance markets is not that assurance disappears. It is that assurance becomes **more valuable and more concentrated**, because the ability to verify is now the scarce input rather than the ability to produce. Two observable consequences follow, and both are testable.

[100] DeFond and Zhang (n 78).

[101] Ronald A Dye, 'Auditing Standards, Legal Liability, and Auditor Wealth' (1993) 101 *Journal of Political Economy* 887.

[102] Bolton, Freixas and Shapiro (n 76); Mathis, McAndrews and Rochet (n 77).

[103] Ann Terlaak, 'Order without law? The role of certified management standards in shaping socially desired firm behaviors' (2007) 32 *Academy of Management Review* 968.

[104] Matthew Potoski and Aseem Prakash, 'Green Clubs and Voluntary Governance: ISO 14001 and Firms' Regulatory Compliance' (2005) 49 *American Journal of Political Science* 235.

[105] Jodi L Short, Michael W Toffel and Andrea R Hugill, 'Monitoring global supply chains' (2016) 37 *Strategic Management Journal* 1878.

[106] Joseph Farrell and Carl Shapiro, 'Standard Setting in High-Definition Television' (1992) 1992 *Brookings Papers on Economic Activity. Microeconomics* 1; Stanley M Besen and Joseph Farrell, 'Choosing How to Compete: Strategies and Tactics in Standardization' (1994) 8 *Journal of Economic Perspectives* 117.

[107] Richard M Locke, Fei Qin and Alberto Brause, 'Beyond corporate codes of conduct: Work organization and labour standards at Nike's suppliers' (2007) 146 *International Labour Review* 21.

[108] Morris M Kleiner and Alan B Krueger, 'The Prevalence and Effects of Occupational Licensing' (2010) 48 *British Journal of Industrial Relations* 676.

First, the share of professional-services revenue attributable to assurance, review and quality-control functions should rise relative to the share attributable to production. This is verification rent in the sense of Part II §6: a return to the scarce factor, which is now the capacity to check rather than the capacity to draft.

Second, that rent should accrue disproportionately to holders of the underlying *registers* rather than to holders of professional expertise, because verification requires an authoritative corpus and expertise without a corpus cannot verify. This is the more interesting prediction and the more consequential one. It implies that the economic beneficiary of the verification constraint is whoever controls the reference data—the citator, the registry, the standards body, the identifier scheme—and not the profession that has historically captured the rent from scarcity of judgement.

If that is right, the distributional consequence of generative artificial intelligence in the professions is not primarily the displacement of junior labour, which is what the labour-market literature has concentrated on. It is a transfer of rent from expertise to infrastructure. That direction of travel has been reached independently from the growth side, where Catalini, Hui and Wu describe rents migrating to verification-grade reference data, provenance and liability underwriting; it also follows from Teece's complementary-assets argument once the register is identified as the complementary asset.[109] What is offered here is not the claim but the test: the ratio of enterprise value in reference-data businesses to enterprise value in the professional-services firms that consume their data, which is observable and which the claim commits to.

13. What the three cases share

Litigation, scholarship and assurance are institutionally unalike and share a structure.

Each of the three exhibits the same four features. A claim is produced by a party with an interest in its acceptance. An institution exists whose function is to establish whether that claim is sound. That institution was already under strain, because verification was already expensive relative to production. And the technology has widened the gap between the two costs by acting on one side of it only.

The three differ in who funds the verifier, and the differences bear on severity rather than on kind. In litigation an opponent is funded to find the flaw, which is why failures there are comparatively visible and comparatively rare. In scholarship the reviewer is unpaid and time-constrained, and failures are correspondingly less visible and more numerous. In assurance the verifier is paid by the party whose assertions it assures, and failures are ordinarily identified on insolvency rather than on inspection. Verification quality tracks who funds it and what the funder is paid to find.

We turn in Part IV to the one of the three in which that proposition can be tested arithmetically. Only scholarly publishing discloses both the number of failures it caught and the volume of output it was checking. The courts publish neither figure, and the assurance market publishes its opinions and not its errors. Part IV therefore measures the caught rate in scholarly publishing across two decades, and reports what the measurement can and cannot establish.

Part IV—What a Verification Apparatus Cannot Tell You About Itself

14. Why this measurement, and what it can decide

Part II stated a proposition about cost structure and Part III showed the same failure in two further markets. Neither is a measurement, and the objection at Part II §8.1—that all of this is the transitional trough of a productivity J-curve—cannot be met by argument alone. It requires a series.

The requirement is specific. What must be distinguished is whether verification cost behaves like a **stock**, accumulated once and thereafter lowering the per-claim cost of checking, or like a **flow**, a recurring per-unit cost of each new claim. Those readings make different predictions about any domain in which a verification apparatus has been built out over an extended period. On the stock reading, the apparatus accumulates

[109] Catalini, Hui and Wu (n 10); Teece (n 8).

and the proportion of defective work it catches should rise. On the flow reading, each new unit of output requires its own check regardless of what infrastructure exists, so a rising volume of output can outrun a rising apparatus and the caught proportion need not move.

Scholarly publishing is the right domain in which to look, for three reasons. It has a verification apparatus that was demonstrably built out over the period. Crossref's Similarity Check service, built on iThenticate, was launched in 2008 and adopted across the large majority of participating publishers within a decade; systematic image forensics entered mainstream practice after Bik, Casadevall and Fang published their screen in 2016; post-publication commenting platforms opened over the same period and are now a recognised route by which defects are surfaced; data-availability requirements were adopted across the major journal families; and the STM Association's Integrity Hub began operating shared paper-mill detection across publishers from 2022.[110] It has an observable failure event: the retraction. And, uniquely among the credence markets discussed in this paper, both numerator and denominator are public. The professional markets of Part II have neither: nobody publishes the number of negligent opinions given.

One thing must be said before the numbers, because it governs how they can be read. A retraction is not an observation of defective work. It is an observation of defective work **that was caught**. The caught rate is the product of the rate at which defective work is produced and the probability that its production is detected, and no series of retractions can separate those two factors. That is not a defect in the design. It is the quantity the study is about, and it is stated here rather than in the results because it governs how the figures may be read.

15. Method

Numerator. The Retraction Watch database, maintained by the Center for Scientific Integrity and distributed by Crossref under a public-domain dedication, comprising 71,496 records at the date of extraction. Each record carries the date of the original publication, the date of the retraction, the article type, the publisher and a structured list of reasons.

Denominator. Crossref annual counts of works of type `journal-article`, retrieved by API for each publication year from 2005 to 2024. Publication volume over the period roughly triples, from 1.75 million to 6.10 million works a year, and no statement about rates is meaningful without it.

The statistic. For each publication cohort year y and window k:

> $\text{rate}(y,\ k) = 10{,}000 \times$ (articles published in y and retracted within k years of publication) $\div$ (articles published in y)

Three corrections are required, and the first two materially change the answer.

Population. Retraction Watch covers conference proceedings, book chapters, preprints and theses in addition to journal articles; the denominator does not. Mixing them does not merely add noise, it manufactures a finding. Computed without the restriction, the 2010 and 2011 cohorts show three-year rates of 20.0 and 18.2 per ten thousand against a restricted-series baseline near 2 for the adjacent cohorts—an apparent tenfold spike in the middle of the series. Inspection shows that 88 per cent of those events are a single publisher's bulk withdrawal of conference proceedings, recorded overwhelmingly under the reason "Notice—Limited or No Information", and that the top six affected venues are all conference proceedings. It is an administrative purge of material that is not in the denominator, and it is not a detection of anything. Restricting the numerator to article types corresponding to the denominator removes it entirely and leaves a series with no discontinuity at those years.

Censoring. A cohort can exhibit a k-year retraction only if k years have elapsed by the data cut. Cohorts without full exposure are excluded rather than reported at an artificially low rate. There remains a residual *reporting* lag—a retraction issued shortly before the cut may not yet be recorded—which depresses the

[110] On similarity screening see Crossref, 'Similarity Check' https://www.crossref.org/services/similarity-check/ accessed 4 August 2026; on image forensics, Elisabeth M Bik, Arturo Casadevall and Ferric C Fang, 'The Prevalence of Inappropriate Image Duplication in Biomedical Research Publications' (2016) 7 *mBio* e00809; on shared paper-mill detection, STM Association, 'STM Integrity Hub' https://www.stm-assoc.org/stm-integrity-hub/ accessed 4 August 2026.

most recent included cohort. That works against the direction of the result reported below, so the result is conservative.

Reason classification. Retraction reasons are reported both in total and for the subset indicating a failure of research integrity rather than honest error: falsification, fabrication, paper mill, plagiarism, image manipulation, data duplication and manipulation, compromised peer review, computer-generated content, concerns about authorship and affiliation, and forged authorship. The boundary is a judgement and it is drawn deliberately broadly. Both series are reported at every cohort so that a reader who would draw it elsewhere can use the total. The classification also includes concerns about authorship and affiliation, which are the marker of the paper-mill trade; a reader who regards that as too broad should note that removing it *raises* the reported ratio from 11.4 to 11.7, and that removing the forged-authorship term as well raises it to 12.1, and that a narrow core of falsification, fabrication, paper mill, plagiarism, image manipulation, machine-generated content and compromised peer review gives 12.9. Every narrowing therefore raises the ratio, and the boundary drawn here is conservative against the objection that it is drawn too wide. It is not conservative in the other direction: broadening it to take in the general misconduct categories gives 10.6, and taking in duplicate publication as well gives 8.1. A reader who would draw the line further out should use those figures, and the direction of the series is unchanged at any of them.

The derivation is published and runs from the two public sources named.

What this adds to a literature that already exists. Retraction has been counted before and counted well, and the paper positions itself against that work rather than around it. Fang, Steen and Casadevall established that the majority of retractions are attributable to misconduct rather than error; Grieneisen and Zhang surveyed the retracted literature comprehensively; and Steen, Casadevall and Fang asked directly why the number of retractions had risen, and answered in favour of improved detection rather than increased misconduct.[111] Cokol and colleagues went further and modelled the *undetected* residue, concluding that the number of papers which should have been retracted greatly exceeds the number that were.[112] Bik, Casadevall and Fang screened more than twenty thousand biomedical papers for image duplication and found problematic figures in 3.8 per cent, at least half suggestive of deliberate manipulation.[113] More recently, Van Noorden reported that the proportion of papers published in a year which go on to be retracted has more than trebled in a decade and exceeded 0.2 per cent for 2022.[114]

Three things are added here and none of them is the direction of the trend, which is established. First, a strict **fixed-window** construction: every cohort is measured over the same elapsed period, so the comparison across cohorts is not contaminated by the fact that older papers have had longer to be caught, which the cumulative measures do not control. Second, a cohort-level split between **total and integrity-type** retraction, which shows the composition moving and not only the level. Third, and the reason for the exercise, the use of the series to interrogate the stock/flow question of Part II and the identification result at §16.2—which is where this Part's contribution actually lies, and which none of the works above states, Steen and colleagues having answered the underlying question in one direction without treating it as undecidable.

16. Results

This section reports the series, then sets out what it cannot decide and why, and finally states the three results which survive that limitation. We report the figures before the interpretation because the interpretation is contested and the figures are not.

16.1 The caught rate rises, and rises fastest for integrity failures

[111] Ferric C Fang, R Grant Steen and Arturo Casadevall, 'Misconduct accounts for the majority of retracted scientific publications' (2012) 109 *Proceedings of the National Academy of Sciences* 17028; Michael L Grieneisen and Minghua Zhang, 'A Comprehensive Survey of Retracted Articles from the Scholarly Literature' (2012) 7 *PLoS ONE* e44118; R Grant Steen, Arturo Casadevall and Ferric C Fang, 'Why Has the Number of Scientific Retractions Increased?' (2013) 8 *PLoS ONE* e68397.

[112] Murat Cokol and others, 'How many scientific papers should be retracted?' (2007) 8 *EMBO Reports* 422.

[113] Bik, Casadevall and Fang (n 110).

[114] Richard Van Noorden, 'More than 10,000 research papers were retracted in 2023—a new record' (2023) 624 *Nature* 479.

Cohort	Articles published	Retracted 3y	Rate /10,000	Integrity-type	Rate /10,000	Integrity share
2005	1,747,097	141	0.81	39	0.22	28%
2006	1,917,794	221	1.15	75	0.39	34%
2007	2,013,125	322	1.60	152	0.76	47%
2008	2,165,405	300	1.39	117	0.54	39%
2009	2,327,201	471	2.02	173	0.74	37%
2010	2,420,959	483	2.00	223	0.92	46%
2011	2,606,604	589	2.26	237	0.91	40%
2012	2,811,548	780	2.77	345	1.23	44%
2013	3,030,239	677	2.23	315	1.04	47%
2014	3,200,647	947	2.96	460	1.44	49%
2015	3,357,780	1,115	3.32	571	1.70	51%
2016	3,588,167	1,107	3.09	434	1.21	39%
2017	3,798,188	1,098	2.89	413	1.09	38%
2018	4,105,980	1,376	3.35	576	1.40	42%
2019	4,442,981	1,900	4.28	931	2.10	49%
2020	4,903,604	2,630	5.36	1,398	2.85	53%
2021	5,302,152	5,315	10.02	3,914	7.38	74%
2022	5,525,887	10,549	19.09	7,868	14.24	75%
2023	5,644,321	2,707	4.80	2,027	3.59	75%

Aggregating the first five cohorts against the last five, the three-year caught rate rises from **1.43 to 8.95** per ten thousand, a factor of **6.3**. The integrity-type rate rises from **0.55 to 6.25**, a factor of **11.4**. Repeating the exercise at a five-year window, which trades cohorts for exposure, gives 1.90 to 7.08 (a factor of 3.7) and 0.78 to 4.40 (a factor of 5.6). The direction and the ordering are the same at both windows, though the margin is not: integrity-type failures rise about 1.8 times as fast as retraction generally at three years, and about 1.5 times as fast at five.

Three features of the series need comment before it is interpreted.

The **2021 and 2022 cohorts** are extreme and they are real. They record the industrialised paper-mill wave and the publisher responses to it, and they are the reason the aggregate ratio is as large as it is. Excluding them entirely and comparing 2005–2009 against 2016–2020 still gives a rise from 1.43 to 3.89 in the total and from 0.55 to 1.80 in the integrity series, factors of 2.7 and 3.3. The finding does not depend on the wave.

The **2023 cohort** is depressed on two counts and is retained because excluding it would flatter the result. The exposure rule admits a cohort where a paper published at its mid-point has had the full window, so papers published late in 2023 have had less than three years; restricting the cohort to its fully-exposed first half gives about 5.0 rather than the 4.80 reported. And Retraction Watch carries a reporting lag, so retractions issued shortly before the cut are not yet recorded. Both work against the direction of the result.

The **integrity share** rises from 38 per cent of retractions across the first five cohorts to 70 per cent across the last five, a factor of 1.8. Taking the endpoint cohorts alone gives 28 per cent against 75 per cent, which is the larger figure and is the one an earlier draft reported; the five-cohort aggregate is used here because every other headline in this Part is computed on that basis. The composition of what is caught has changed, and not only the quantity.

16.2 What the series cannot decide, and why that is the finding

The caught rate is the product of two unobservables:

$$\text{caught}(y) = \text{produced}(y) \times \text{detected}(y)$$

and the series measures only their product. Two readings sit at the extremes and both are consistent with the data.

Reading one: detection was roughly constant and production rose. On this reading the production of integrity-defective work rose about elevenfold in twenty years, and the scholarly record is deteriorating at a rate that the apparatus is tracking accurately.

Reading two: production was roughly constant and detection improved. On this reading detection in the early cohorts was catching something on the order of a twelfth of what was there, which means the 2005–2015 record contains an undetected residue roughly ten times the size of what was found in it. The apparatus is not tracking a deterioration; it is belatedly revealing a level that was always present.

That ten-fold figure is a **floor and not an estimate**, and stating the two readings as though they bracketed the truth would be wrong. It holds only if detection is now close to complete, which nothing suggests. Bik, Casadevall and Fang's direct screen found problematic figures in 3.8 per cent of the biomedical papers examined, of which at least half were suggestive of deliberate manipulation. That is roughly 380 per ten thousand on the whole measure and roughly 190 per ten thousand on the integrity comparison, against a caught rate in this series of between one and two per ten thousand for the same period.[115] The screen covers one field and one failure mode and cannot simply be transposed, but it establishes the direction of the error in the floor, which is that the floor is very low.

The two readings describe different states of the record. On the first, the corpus was sound and its condition is deteriorating. On the second, the corpus was never sound and we have only recently acquired the means of showing it. The policy responses differ, the appropriate confidence in pre-2015 literature differs, and the implication for whether the problem is self-correcting differs completely.

Nothing in the data chooses between them, and the reason is exactly the paper's thesis. To separate production from detection one needs an independent estimate of the rate at which defective work is produced—which is to say, one needs to know what the verification apparatus missed. An apparatus cannot measure its own coverage, because the measurement requires the thing the apparatus failed to observe. This is the structure described at Part II §5. Scholarly publishing is the only one of the three markets in which both terms of the product are public, which is why it can be exhibited here and not in the other two.

One objection to the strong form of that claim has to be met before the discriminators are set out, because it is the obvious technical one. It is not true that nothing inside the system bears on the residue at all. Cokol and colleagues (n 112) modelled the undetected residue from the internal record and concluded that the number of papers which should have been retracted greatly exceeds the number that were; and where two detection channels are partially independent—journal-side screening and third-party post-publication scrutiny, say—a capture–recapture estimator will in principle recover the size of what neither caught.

Neither route identifies the quantity, and the reason is the same in both cases. A model of the residue returns an estimate conditional on assumptions about the detection process which are themselves the unobservable; change the assumed detection function and the estimate moves with it. Capture–recapture requires the two channels to be independent conditional on the paper, and they are not: a paper with a conspicuous defect is more likely to be caught by both, which biases the estimator downwards by an unknown amount. Both methods therefore convert the identification problem into an assumption problem rather than solving it. That is a weaker claim than the one an earlier draft made, and it is the one that can be sustained: what is unavailable is not information bearing on the residue but a route to it that does not pass through an assumption about the thing being measured.

Three partial discriminators exist and none resolves it.

The **composition shift** is informative but ambiguous. If detection had improved uniformly, the integrity share of retractions should be roughly stable; it nearly doubled. That is consistent with integrity failures growing faster than honest error, and equally consistent with detection improving *selectively*, since similarity detection, image forensics and paper-mill screening target integrity failures specifically and have no analogue for honest error. The tools that were built are precisely the tools that would produce this composition shift at constant underlying rates.

[115] Bik, Casadevall and Fang (n 110).

The **five-year window** is informative about latency but not about coverage. Rates at five years exceed rates at three years in every cohort, which establishes that detection continues after the three-year mark and that all figures here are floors. It says nothing about the residue that is never detected at any horizon.

The **third discriminator is the most useful and it leans against reading two.** If detection improved elevenfold against a residue ten times what was originally found, the improved apparatus turned on the back catalogue should be finding that residue now. Counting retractions of papers published in 2005–2010 by the calendar year in which the retraction was *issued*, and expressing them against that cohort's publication volume, gives 2.11 per ten thousand for retractions issued in 2006–2013 and **0.69** for those issued in 2018–2025. The flow of retractions against the old corpus has fallen by roughly two-thirds at exactly the time the tools improved.

That is evidence against the strong form of reading two, and it should be weighed rather than treated as decisive, because it is confounded in three ways that all run the same way. Similarity screening operates at the point of submission and therefore never touches the back catalogue at all. Post-publication scrutiny concentrates on recent work, because that is what is being cited and built on. And the incentives on a publisher to investigate a twenty-year-old paper are weak: the correction literature records that retraction decisions turn substantially on institutional investigation, which requires an employer with a continuing interest in the author.[116] Each of those would produce a falling flow against an old corpus even if the residue were large. The discriminator therefore narrows the range rather than closing it: it makes the extreme version of reading two—a residue ten times the catch, sitting undetected in a corpus now under much better surveillance—harder to sustain, without establishing where in the range the truth lies.

16.3 What the result does establish

Three results survive the identification problem.

The two readings are not neutral between the accounts, and one of them favours the objection. An earlier draft asserted that on either reading the gap between production and detection failed to close, and that neither was the signature of a transitional cost being amortised. That is wrong on reading two, and the error mattered. If production was constant and detection improved elevenfold, then the gap closed dramatically, and an elevenfold return on two decades of accumulated verification infrastructure is *exactly* the J-curve signature. Reading two is evidence **for** the objection this Part was built to test, not neutral between the accounts.

What survives is therefore conditional and should be stated as such. On reading one the apparatus is running to keep up and the flow account holds. On reading two the stock account holds for the component the tools reach, and the question becomes what the residue looks like—which is where the Bik comparison above and the back-catalogue discriminator do their work, the first suggesting the residue is large and the second that it is not being found. The honest position is that Part IV constrains the possibilities without choosing between them, and that its contribution is the constraint rather than a resolution.

The two components can be separated on this corpus, and they behave differently. The partition at Part II §8.1 is stated as a claim about cost structure and §22 concedes that the relative size of the two components is asserted rather than measured. The corpus permits a sharper test than the integrity-versus-error split reported at §16.1, and it is reported here because it is the closest thing to a direct measurement the data will yield.

Similarity screening and image forensics reach a specific set of the integrity reasons: plagiarism, duplication of data, duplication of images, and image manipulation. They do not reach paper-mill production, compromised peer review, forged or disputed authorship, falsification, fabrication, manipulation of results or of data, or machine-generated content. Splitting the integrity series on that line, and assigning a retraction to the second component wherever it carries any reason the tools do not reach—which is conservative towards the first—gives a partition rather than an overlap, so the two rows sum to the integrity rate reported above.

[116] Fang, Steen and Casadevall (n 111), on the role of institutional investigation in producing a retraction.

Component of the integrity series	First five cohorts	Last five	Factor
Reached by similarity and image screening	0.32 /10,000	0.62 /10,000	×1.9
Carrying a reason not reached by them	0.23 /10,000	5.63 /10,000	×24.9
Integrity series, as reported at §16.1	*0.55*	*6.25*	*×11.4*

Crossref's similarity service launched as CrossCheck on 19 June 2008 and has therefore been in operation across all but the first three cohorts measured.[117] The component that oldest and best-amortised tool reaches roughly doubles. The component it does not reach rises by a factor of twenty-five. The whole of the divergence between the integrity series and the aggregate is carried by the second component.

Three qualifications bound what this shows, and the first is the most serious. At a five-year window the same split gives ×2.5 against ×9.5. The ordering survives the change of window, as every other result in this Part does, but the margin does not: the twenty-five-fold figure is specific to the three-year measure and is inflated by the 2021 and 2022 paper-mill cohorts, which had not fully matured at five years by the cut. The defensible statement is that the unreached component rises between four and thirteen times faster than the reached one depending on the window—9.5 against 2.5 at five years, 24.9 against 1.9 at three—not that it rises thirteen times faster.

Second, these are not the same two components as the partition at §8.1. What similarity screening reaches is a subset of the mechanisable, and what it does not reach includes failures which are mechanisable in principle and for which tools were built later—paper-mill screening being the obvious case, which is why the second row's rise is partly a detection artefact of the same kind §16.2 describes.

Third, the identification problem applies to each series separately. The near-flat line for the reached component is as consistent with a stable underlying rate detected well as with a falling underlying rate detected better. What the split does establish, and it is not nothing, is that the aggregate rise is not distributed across the failure types the long-amortised tool addresses. That is the direction the partition predicts and it is the first time it has been measured rather than asserted.

The mechanisable component did amortise, and it was not enough. Similarity detection is a stock: built once, applied cheaply thereafter, and it demonstrably works—plagiarism and duplication are detected at scale and quickly. That is precisely the component Part II §8.1 concedes behaves like a stock. What the composition shift shows is that the caught population has moved towards categories that mechanisation reaches, while the total continued to rise. A stock was built, it retired the cheap-to-check tail, and the problem did not go away.

The per-unit component of the burden tripled irrespective of which reading is right. Publication volume rose from 1.75 million journal articles in 2005 to 5.64 million in 2023, a factor of 3.2. Checking is performed per article, so the checking burden attributable to volume alone tripled over the period, and it did so whether the rate of defective production rose, fell or stayed where it was. That is the component of verification cost which behaves as a flow, and it is the only component of the rise which the identification problem at §16.2 does not contaminate. The absolute count of integrity-type retractions within three years rose over the same cohorts from 39 to 2,027; that figure is arithmetically the product of the volume rise and the rate rise, so it decomposes but does not discriminate, and it is reported here for completeness rather than as evidence.

17. Limits

Four limits apply.

The population is not the population of interest. Crossref journal-article counts and Retraction Watch coverage do not describe the same universe. Retraction Watch monitors comprehensively but not

[117] Crossref, 'CrossCheck Plagiarism Screening Service Launches Today' (Crossref, 19 June 2008) https://www.crossref.org/news/2008-06-19-crosscheck-plagiarism-screening-service-launches-today/ accessed 3 August 2026. The service was renamed Similarity Check and is powered by iThenticate.

exhaustively; Grieneisen and Zhang's independent survey of the retracted literature found items that no single database then captured, and the coverage of the database used here has itself improved over the period[118]—which is a further channel through which measured rates rise without either production or journal-side detection changing. This cuts in the same direction as reading two and cannot be netted out.

Retraction is a lagging and institutionally mediated event. Whether a defective paper is retracted depends on the publisher, the institution, the jurisdiction and the willingness of somebody to press the point. It is a measure of what the system was willing to act on, not of what it found.

The integrity boundary is a judgement. Reasons are reported in overlapping free-text categories and a paper can carry several. The classification here is deliberately broad and both series are reported so that the sensitivity is visible.

Fixed-window rates are floors. Every cohort will accrue further retractions after its window closes, and the earlier cohorts have had longer to do so—which biases the comparison *against* the reported rise, and is one reason the five-year series shows a smaller factor than the three-year series.

None of these limits touches the identification result at §16.2, which does not depend on the level or the slope but on the structure of the quantity being measured.

18. What this Part contributes to the argument

Part II claimed that verification cost divides into a mechanisable component that amortises and an irreducible component that does not, and that the second is the larger. This Part supplies the first evidence for that claim from a domain where both sides of the ratio are public: the mechanisable component was mechanised, the caught rate rose, and the problem grew rather than closed.

It also supplies something the argument did not have, which is a demonstration rather than an assertion of the central difficulty. Every institution discussed in this paper reports on its own verification performance—courts on sanctions, auditors on opinions, journals on retractions—and every one of those reports is a count of what was caught, presented in a context that invites it to be read as a count of what occurred. The two are separated by exactly the unobservable this paper is about. The retraction series is the case where the arithmetic can be done and the gap made visible. There is no reason to think the professional markets of Part II are better placed; they are worse, because they do not publish the numerator at all.

Part V—What Follows

19. The design problem

A proposal follows from the analysis, and it is constrained by the finding that produced it: it may not assume that verification becomes cheap, because the whole argument is that it does not. A proposal which depends on proof becoming cheaper restates the problem it is offered to address, and is excluded on that ground.

The proposal therefore takes the cost as given and *partitions* it. It does not move a burden from one party to another, and an earlier formulation of this argument said that it did. Part II §8.1 and Part IV establish that verification cost divides into a component a register can discharge—whether a cited authority exists—and a component no register can discharge—whether a real authority supports the proposition it is advanced for. The first is allocated below to the party that captures the saving from cheap production. The second is left, expressly and cumulatively, with the professional, because there is nobody else who can bear it. Limb three is therefore not a saving provision but the operative one, and the difference between dividing a cost and moving it is the whole of what this Part adds to the literature on allocating it.

[118] Grieneisen and Zhang (n 111). On the sources themselves see the Center for Scientific Integrity, *The Retraction Watch Database* (distributed by Crossref under a public-domain dedication) http://retractiondatabase.org/ accessed 4 August 2026, and Crossref, 'REST API' https://api.crossref.org/ accessed 4 August 2026, from which the annual journal-article counts were retrieved on 4 August 2026.

20. A verification duty on the party that captures the saving

This section sets out the present allocation of verification cost and why it is wrong, considers whether the common law would move it, drafts the duty in the form in which it would have to be enacted, positions it against the instruments now in force, and states why the alternatives are worse.

20.1 The present allocation, and why it is wrong

Part II §9 showed the courts allocating the entire verification cost to the practitioner who filed the document. In *Ayinde* the Divisional Court exercised the *Hamid* jurisdiction over the practitioners; in *Park v Kim* the Second Circuit referred counsel to its Grievance Panel; in *Mata* the district court sanctioned the individuals and the firm.[119] That allocation is intelligible—the practitioner is before the court, owes it duties, and is the last party who could have checked—and it is economically wrong in a way that will produce predictable failure.

It is wrong because it places the burden on the party that captured the *smallest* share of the saving on g, and gave the party that captured the largest share no incentive at all. A practitioner who uses a generative tool saves an hour. The vendor of the tool captures a subscription across an entire profession. The court, and through it the public, absorbs the verification cost of every unchecked assertion that reaches it. Under the standard economic analysis of accident cost, liability should rest with the cheapest cost avoider.[120] The cheapest avoider of a fabricated citation is not the practitioner reading the output; it is the party that could have resolved the citation against a register before emitting it, at a marginal cost close to zero and once for every user.

20.1A Where the duty comes from, and where it does not

That is an economic argument for an allocation. It is not, by itself, a legal argument for a duty, and an earlier draft of this paper conflated the two. English law does not impose duties of care on cheapest-cost-avoider grounds. It develops the law of negligence incrementally and by analogy with established categories, and *Robinson* is emphatic that the search for a general principle is not how the question is approached.[121] The economic argument tells a legislature what to do. It does not tell a court that a duty exists.

To whom would the duty be owed? The question must be answered before any other, and the candidates are not equivalent. To the *user*, there is a contract, and the duty would in practice be a matter of its terms and of consumer protection rather than of tort. To the *court*, no duty of care is owed by a stranger to proceedings, and the *Hamid* jurisdiction operates on officers of the court and not on their suppliers. To the *opposing party and the public*, the loss is pure economic loss caused by a statement, which is the most restrictively policed category in English law. The duty proposed here is owed to the user and, through the mechanism in Limb two, is intended to protect third parties by making the defect visible rather than by giving them a cause of action.

Could it be reached at common law? The route would be *Hedley Byrne*: an assumption of responsibility for the accuracy of a statement, with reliance and knowledge of reliance.[122] The obstacles are serious and must be stated rather than glossed. Under *Caparo* the maker of a statement owes no duty to an indeterminate class using it for indeterminate purposes.[123] *Playboy Club* refused a duty where the representee was not the person to whom responsibility had been assumed—the reference was addressed to another, and the Club was an undisclosed principal. *NRAM* is if anything the stronger authority against the present proposal, and for a different reason: there the representee *was* the addressee, and the duty still failed, because reliance without independent verification was not reasonable where the recipient had the means and the professional

[119] *Ayinde* (n 83); *Park v Kim* (n 81); *Mata v Avianca Inc* (n 80).

[120] Guido Calabresi, *The Costs of Accidents: A Legal and Economic Analysis* (Yale University Press 1970) ch 7, 135, 143–144. The cheapest-cost-avoider concept is developed there; the later article with Melamed, 'Property Rules, Liability Rules, and Inalienability: One View of the Cathedral' (1972) 85 *Harvard Law Review* 1089, addresses the different question of entitlement protection and is not authority for it. On the choice between liability and regulation as instruments once the allocation is settled, see Steven Shavell, 'Liability for Harm versus Regulation of Safety' (1984) 13 *The Journal of Legal Studies* 357.

[121] *Robinson v Chief Constable of West Yorkshire Police* [2018] UKSC 4, [2018] AC 736 [21]–[29] (Lord Reed).

[122] *Hedley Byrne & Co Ltd v Heller & Partners Ltd* [1964] AC 465 (HL).

[123] *Caparo Industries plc v Dickman* [1990] 2 AC 605 (HL) 620–621 (Lord Bridge).

obligation to check.[124] Applied here, that is the objection in its sharpest form—a practitioner who does not check an authority before citing it is the party whose reliance was unreasonable, and a vendor would rely on that finding. A general-purpose system emitting text to millions of users for unknown purposes is close to the paradigm case in which English law declines a duty, and an intermediate professional under an independent verification duty of his own—Limb three—is precisely the factor that has historically broken the chain.

There is one countervailing feature, and it is the strongest material available. Where these tools are marketed to the profession on the express representation that the fabrication problem has been solved, and are found on measurement still to fabricate, the representation is specific, is directed at an identified class using the product for the very purpose contemplated, and is relied upon.[125] That is an assumption of responsibility in the *Hedley Byrne* sense and it is capable of founding a duty in respect of the represented attribute. But it is a duty arising from what a particular vendor said, not a duty on vendors generally, and it therefore cannot do the work the duty proposed at Part V requires.

So the duty must be imposed, not found. That is the honest position. After *Barton v Morris*, a term of this kind is not to be arrived at by construction or implication, and a duty of this kind is not to be arrived at by incremental extension.[126] The duty proposed at Part V is therefore framed as a statutory duty. *Manchester Building Society v Grant Thornton* has a role, but a different and smaller one than an earlier draft claimed: it is a scope-of-duty principle limiting the *extent* of recoverable loss once a duty exists, by asking what risk the service was undertaken to protect against.[127] It does not generate duties and has never been used to do so. Applied to the statutory duty proposed below it answers the question of quantum—the vendor is answerable for losses flowing from the assertion of a non-existent authority, not for the consequences of a bad argument built on a real one—and that is a useful limit on an obligation which would otherwise be open-ended. *Khan v Meadows*, handed down the same day, applies the same six-question analysis in the clinical-negligence context and confirms its generality across professional settings.[128]

20.2 The duty

This section sets out the proposed duty in the form in which it would have to be enacted. It is drafted rather than described, because a proposal which stops at description leaves the questions that decide whether it can operate—who owes it, to whom, on what standard, and with what defence—for somebody else to answer.

Verification of Emitted Authority Act (proposed short title: VEA)

Preamble. An Act to require persons who supply automated systems capable of emitting citations to identified authorities to take reasonable steps to establish that those authorities exist; to require disclosure of the verification state of such citations; to preserve existing professional obligations; and for connected purposes.

1. Definitions. For the purposes of this Act—

(a) "relevant system" means a system supplied in the course of a business which, in response to a request, generates text capable of containing a citation, and includes a system supplied by way of a licence to operate it on the licensee's own infrastructure;

(b) "citation" means a reference which purports to identify a judicial decision, statutory provision, published article, book or dataset, whether or not it is capable of being resolved and whether or not the authority referred to exists;

(c) "recognised register" means a publicly available index of the class of authority in question, whether maintained by a public body, a publisher or a third party;

[124] *Steel v NRAM Ltd* [2018] UKSC 13, [2018] 1 WLR 1190 [24]–[35]; *Banca Nazionale del Lavoro SpA v Playboy Club London Ltd* [2018] UKSC 43, [2018] 1 WLR 4041 [7]–[16].

[125] See Varun Magesh and others, 'Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools' (2025) 22 *Journal of Empirical Legal Studies* 216. See also Dahl and others (n 54). The marketing representations are described in both.

[126] *Barton v Morris* [2023] UKSC 3, [2023] AC 684 [22]–[28]; and see *Robinson* (n 121).

[127] *Manchester Building Society v Grant Thornton UK LLP* [2021] UKSC 20, [2022] AC 783 [10]–[22], [41] (Lord Hodge and Lord Sales).

[128] *Khan v Meadows* [2021] UKSC 21, [2022] AC 852 [28].

(d) "resolution" means the operation of matching a citation against a recognised register and returning a determination that the authority exists and bears the identifier given, does not so exist, or falls outside the coverage of every register consulted;

(e) "supplier" means the person who makes a relevant system available to a user, whether or not that person trained or developed it;

(f) "user" means the person to whom the output of a relevant system is made available, whether or not for consideration.

2. Limb one: resolution before assertion. A supplier shall take reasonable steps to resolve each citation emitted by a relevant system before it is emitted. The standard is reasonableness and not accuracy, and a citation which cannot be resolved need not be suppressed. In determining whether the steps taken were reasonable, regard shall be had to the registers available for the class of authority concerned, the cost of resolution relative to the value of the service, and the state of the art at the time of emission.

3. Limb two: disclosure of the verification state. Where a citation has not been resolved, the supplier shall ensure that fact accompanies the citation, and shall take such steps as are technically feasible at the time of emission to ensure that it persists when the output is copied, exported or reproduced. The disclosure shall distinguish a citation which is absent from the registers consulted, one which falls outside their coverage, and one in respect of which no register was consulted.

4. Limb three: preservation of professional duties. Nothing in this Act reduces any duty owed by a professional person to a court, a client, a regulator or a third party. The duties are cumulative.

5. Burden. Where it is shown that a citation was emitted unresolved and unmarked, it is for the supplier to show that the steps taken satisfied section 2.

6. Territorial application. This Act applies to a supplier which makes a relevant system available to a user in the United Kingdom, whether or not the supplier is established here, and whether or not the output is generated here.

7. Enforcement. The duties in sections 2 and 3 are enforceable by the regulator designated under section 8, on the balance of probabilities, by compliance notice and by civil penalty. A penalty under this section shall not exceed the greater of £10 million and 2 per cent of the supplier's worldwide turnover in the preceding financial year; within that limit the Secretary of State may by regulations prescribe the basis of calculation, which shall have regard to the number of outputs emitted in breach. No action for damages lies at the suit of a user in respect of a breach of section 2 or section 3 alone.

8. Designated regulator and appeals. The Secretary of State shall by regulations designate a regulator for the purposes of section 7. A person on whom a compliance notice is served or a penalty imposed may appeal to the First-tier Tribunal, which may quash, vary or confirm the notice or penalty. The regulator shall publish annually the number of compliance notices issued, the number of penalties imposed, and the number of relevant systems assessed.

Three points of construction should be made explicit, because they are where the drafting does its work.

Limb one is a **conduct** standard and not a **status** test. It asks what the supplier did before emission, not whether the supplier is of a particular size, sophistication or market position. That is deliberate. A status test invites arbitrage through corporate structure, which is the standard objection to threshold-based regulatory design and the reason the Product Liability Directive attaches to the product rather than to the size of the producer; the conduct in question, running a lookup, is observable and cheap to evidence.[129]

Limb two is the limb on which the economic argument turns. Part II establishes that unverified output is indistinguishable from verified output at the point of receipt. A duty which separates them restores the missing information and converts a credence attribute into an experience attribute for the single characteristic that can be mechanically established. The requirement that the mark persist through copying is

[129] Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products [2024] OJ L2024/2853, arts 4(1) and 8, which attach liability to the product and to the economic operator's role in placing it on the market rather than to the operator's size.

not incidental; a disclosure which is lost on export discloses nothing. It is qualified by technical feasibility because it has to be: no drafting can require a supplier to control what happens to plain text once a user has selected and pasted it, and a duty stated in absolute terms would be either unenforceable or a prohibition on emitting text at all. What the qualifier requires is that the supplier adopt the means available—structured output formats carry the field, and a document export controlled by the supplier can carry it—and it is the regulator's supervisory function under section 7 rather than a court's construction of the section that will fix where the line falls in a given year.

Limb three is not a saving provision. It is the operative allocation. The practitioner remains the only party who can assess whether a genuine authority supports the proposition advanced, which is a different question from whether the authority exists and one no register can answer. Part II §8.1 argues, and Part IV supplies the first evidence for the view, that this second component is the larger of the two; §22 states the width of that claim, which is short of establishing it. The Act therefore allocates the smaller and mechanisable component to the supplier and leaves the larger and irreducible component where it already sits.

20.2A Why the existing AI governance instruments do not reach it

A reasonable objection to the duty proposed here is that the field is already regulated, and that a further duty is redundant. Two enacted instruments now occupy ground adjacent to it, and an earlier draft of this section engaged neither, relying instead on a commentary written about a legislative proposal that was substantially amended before enactment. That was an error of exactly the kind this paper is about, and the corrected analysis narrows the proposal's claim considerably while making what survives of it much more precise.

The Artificial Intelligence Act, art 50. Regulation (EU) 2024/1689 is the developed example of risk-tiered product safety, and Veale and Zuiderveen Borgesius's analysis of the Commission's proposal identified its structure as a product-safety regime adapted from four decades of European practice.[130] But the provision that matters here is not in the risk-tiered part of the Regulation at all. Article 50 imposes *horizontal* transparency obligations independent of risk classification. Providers of systems generating synthetic audio, image, video or text must ensure that outputs are marked in a machine-readable format and detectable as artificially generated or manipulated, with solutions that are effective, interoperable, robust and reliable as far as is technically feasible; deployers publishing AI-generated text on matters of public interest must disclose that it is AI-generated, subject to an editorial-responsibility carve-out; and the AI Office is to facilitate codes of practice on detection and labelling.[131] Those obligations became applicable on 2 August 2026.[132] Penalties reach EUR 15 million or 3 per cent of worldwide turnover.[133]

That is a provenance-marking obligation, and Limb two of this proposal is a provenance-disclosure duty. The overlap must be conceded. **What art 50 requires is that output be marked as machine-generated. What it does not require is that an assertion within the output be resolved against a register, or that its resolution state travel with it.** Those are different facts and only the second is decision-relevant to a reader. A document marked "AI-generated" tells a court that a tool was used, which §20.3 argues is the wrong disclosure: it produces stigma without information, and it is uninformative precisely because the marking is uniform across output whose citations resolve and output whose citations do not. Limb two survives art 50 as a narrower and more useful obligation, and the residual claim is now stated at that width rather than as a claim that the field is unregulated.

The Product Liability Directive. Directive (EU) 2024/2853 does three things that bear directly on the proposal, and it does them on the defendant side.[134] It brings software—embedded, standalone, or supplied

[130] Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act' (2021) 22 *Computer Law Review International* 97. The article analyses the Commission's 2021 proposal, which was substantially amended before enactment; it is cited here for its account of the regime's product-safety architecture, which survived, and not for the content of any enacted provision.

[131] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L2024/1689, art 50(1), (2), (4) and (7), and recitals 132–135.

[132] ibid art 113 (application of Chapter IV from 2 August 2026).

[133] ibid art 99(4)(g).

[134] Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective

as a service, expressly including AI systems—within the definition of "product", closing the gap that made the vendor's position uncertain. It empowers a court to order disclosure of evidence against a defendant, with a presumption of defectiveness where the order is not complied with. And it creates rebuttable presumptions of defectiveness and of causation where the claimant faces excessive difficulties of proof owing to technical or scientific complexity, with the opacity of machine-learning systems expressly in contemplation. Member States must transpose by 9 December 2026.

The significance for this paper is larger than a point about redundancy. The allocation proposed here, and the related technique of shifting a burden to the party holding the information where the other party's proof is structurally obstructed, are not speculative devices. The European legislature has enacted both, in a domestic-law setting, within the last two years. The proposal below should therefore not be presented as an invention, and it is not. What the duty proposed at Part V adds to the Directive is the *ex ante* obligation: the Directive allocates loss after harm has occurred and requires a claimant, an injury and a forum, whereas the failure described in Part II is one whose cost falls on a court and a public who are not claimants and will not sue. A liability rule that operates only through litigation cannot reach a harm whose defining feature is that nobody with standing notices it.

The rights-based strand fares no better for this purpose. The debate over whether the General Data Protection Regulation confers a right to explanation is now well developed, and the sceptical readings have largely prevailed: Wachter, Mittelstadt and Floridi argue that no such right exists in the operative provisions, and Edwards and Veale conclude that even a robust version would not remedy the harms it is invoked against.[135] Kaminski's binary-governance account is the most constructive response, arguing that individual rights and systemic accountability are complementary rather than alternative.[136] But explanation and verification are different goods. Selbst and Barocas separate the properties precisely—inscrutability and non-intuitiveness are distinct problems requiring distinct responses—and neither response establishes whether the output is *true*.[137] Burrell's taxonomy of opacity makes the same separation from the technical side.[138] One can be told exactly how a system produced a citation and be no closer to knowing whether the case exists.

The accountability literature identifies the structural obstacle. Raji and colleagues' internal audit framework and Cobbe, Veale and Singh's analysis of algorithmic supply chains both show that responsibility diffuses across a chain of providers, integrators and deployers such that no single party is positioned to answer for the output.[139] Pasquale's account of the black box anticipated the same difficulty from the accountability side a decade earlier.[140] The proposal is deliberately narrow in response: it attaches at the point of emission, to a specific and mechanically checkable class of assertion, precisely because broader duties dissolve in the diffusion these authors describe.

Two further considerations argue for acting now rather than waiting.

The empirical position is not improving. Magesh and colleagues' assessment of commercial legal research tools found hallucination persisting in systems marketed as having eliminated it.[141] Dahl and colleagues' earlier measurement established the base rate and its concentration in exactly the queries hardest

products [2024] OJ L2024/2853, arts 4(1) (definition of "product", expressly including software), 9 (disclosure of evidence), 10(2)(a) (presumption on non-disclosure), 10(4) (presumptions where proof is excessively difficult owing to technical or scientific complexity) and 22 (transposition by 9 December 2026).

[135]Sandra Wachter, Brent Mittelstadt and Luciano Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7 *International Data Privacy Law* 76; Lilian Edwards and Michael Veale, 'Slave to the Algorithm? Why a "right to an explanation" is probably not the remedy you are looking for' (2017) 16 *Duke Law & Technology Review* 18.

[136]Margot E Kaminski, 'Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability' (2019) 92 *Southern California Law Review* 1529.

[137]Andrew D Selbst and Solon Barocas, 'The Intuitive Appeal of Explainable Machines' (2018) 87 *Fordham Law Review* 1085.

[138]Jenna Burrell, 'How the machine "thinks": Understanding opacity in machine learning algorithms' (2016) 3 *Big Data & Society* 1.

[139]Inioluwa Deborah Raji and others, 'Closing the AI accountability gap' (2020) *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 33; Jennifer Cobbe, Michael Veale and Jatinder Singh, 'Understanding accountability in algorithmic supply chains' (2023) *2023 ACM Conference on Fairness, Accountability, and Transparency* 1186.

[140]Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015).

[141]Magesh and others (n 125).

to check.[142] These are products sold to the profession on the representation that the problem has been solved.

The corpus itself may degrade. Shumailov and colleagues demonstrate model collapse where systems are trained on recursively generated data: the distribution narrows and tail information is lost.[143] Bender and Koller's argument that form is not meaning identifies why the failure is not a defect to be engineered away.[144] If a growing share of the textual record is machine-generated and unmarked, the register against which any verification must be performed is itself contaminated—which makes the provenance obligation in Limb two more urgent, not less. Grimmelmann's analysis of how copyright treats machine reading, and Calo's account of what robotics teaches about cyberlaw's transferable lessons, both anticipate the general difficulty of regulating a technology whose outputs are indistinguishable in form from the human artefacts the law was written for.[145]

20.3 Why this and not the alternatives

Why not prohibition. A rule against the use of such tools in professional work is unenforceable—use is unobservable, which is the premise of the whole analysis—and would in any event forgo real gains. A rule whose breach cannot be observed does not alter conduct, and the enforcement literature treats detectability as a condition of efficacy rather than as an implementation detail.[146]

Why not disclosure alone. Requiring practitioners to disclose that a tool was used discloses the wrong fact. Whether a tool was used is uninformative; whether the assertions were resolved is the fact a reader needs. Disclosure of the former without the latter produces stigma without information.

Why not a standard rather than a rule. Kaplow's analysis of rules and standards turns on where the cost of specifying content falls, and favours rules where conduct is frequent and the specification cost can be spread.[147] Citation resolution is the paradigm case: it is enormously frequent, mechanically specifiable, and identical across users. Diver's treatment of optimal precision supports the same conclusion.[148]

Why not leave it to the market. Because Part II §8.3 showed that the reputational mechanism cannot operate on an attribute that is unobservable at the point of purchase. That is not a market that will resolve itself; it is Akerlof's market, and its equilibrium is the withdrawal of the high-quality product.[149]

20.4 Feasibility

The duty is only worth proposing if it can be discharged, and the objection that no adequate register exists deserves a concrete answer rather than an assurance.

The author has built one. A citation index of 75,837 English judgments spanning 2001 to 2026 across thirteen court identifiers was assembled from Find Case Law under a computational-analysis permission and combined with the Retraction Watch corpus described at §15. It runs offline, resolves a citation in milliseconds, and—critically—distinguishes a citation that is *absent from* the corpus from one that falls *outside* its coverage. That distinction is the difference between a usable instrument and an unusable one: the index contains no House of Lords judgments, so it reports *R (Bancoult) v Secretary of State for Foreign and Commonwealth Affairs (No 2)* as outside its reach rather than as unverified.[150] A verifier that called that decision fabricated would teach its users to ignore every warning it gave.

---

[142] Dahl and others (n 54).

[143] Ilia Shumailov and others, 'AI models collapse when trained on recursively generated data' (2024) 631 *Nature* 755.

[144] Emily M Bender and Alexander Koller, 'Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data' (2020) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 5185.

[145] James Grimmelmann, 'Copyright for Literate Robots' (2016) 101 *Iowa Law Review* 657; Ryan Calo, 'Robotics and the New Cyberlaw' (2015) 103 *California Law Review* 513.

[146] Gary S Becker, 'Crime and Punishment: An Economic Approach' (1968) 76 *Journal of Political Economy* 169, 176–179 (expected sanction as the product of detection probability and penalty); A Mitchell Polinsky and Steven Shavell, 'The Economic Theory of Public Enforcement of Law' (2000) 38 *Journal of Economic Literature* 45.

[147] Louis Kaplow, 'Rules versus Standards: An Economic Analysis' (1992) 42 *Duke Law Journal* 557.

[148] Colin S Diver, 'The Optimal Precision of Administrative Rules' (1983) 93 *The Yale Law Journal* 65.

[149] Akerlof (n 5).

[150] *R (Bancoult) v Secretary of State for Foreign and Commonwealth Affairs (No 2)* [2008] UKHL 61, [2009] 1 AC 453.

The technical components have existed for decades. Cryptographic time-stamping and hash-linked structures were solved problems by the early 1990s.[151] What has been missing is not capability but obligation, and Limb one supplies it.

21. Enforcement and the probability of detection

The proposal imposes a duty, and a duty is only as good as the probability of its being enforced. Where that probability is low the enforcement literature is consistent: it is the expected sanction and not the nominal rule which governs conduct.

Becker's framework and Stigler's refinement establish the arithmetic: what governs behaviour is the expected sanction, and a rule carrying a negligible probability of enforcement does not alter conduct.[152] The institutional literature adds that the *form* of enforcement matters as much as its intensity. Milgrom, North and Weingast's account of the Law Merchant shows how a private institution can sustain contracting without a state, provided it maintains a *reliable record* of who did what.[153] Dixit's model of trade expansion under weak enforcement reaches the same conclusion from the theoretical side: the constraint on the extent of the market is the reach of whatever mechanism verifies conduct.[154]

The same structure appears in the literature on medieval trade, and in two distinct forms. Milgrom, North and Weingast show that the Law Merchant sustained exchange by centralising record-keeping and admitting to the fairs only merchants in good standing. Greif, Milgrom and Weingast address a different problem—how a ruler could commit to the security of alien merchants—and their answer is a coordinated collective embargo rather than a register. Only the first is the mechanism relied on here, and it is the reach of the record, not the reach of the sanction, that sets the extent of the market.[155] A market grows to the extent that conduct can be verified and recorded; where the record fails, exchange retreats to parties who can monitor each other directly. The proposal here is, in that light, an attempt to maintain a record adequate to the extent of the market: the register against which an assertion resolves, and the mark that travels with it when it does not.

The political-economy literature supplies the caution. Shleifer and Vishny's analysis of corruption shows how enforcement discretion becomes a rent, and Olson's account of stationary banditry explains why an enforcing authority's incentives are not automatically aligned with the rule it enforces.[156] Applied here the risk is specific and was not conceded in an earlier draft: whoever maintains the register against which citations resolve acquires a rent, and the holder of a register has an incentive to disclose the minimum consistent with capturing it—which is Lizzeri's result, discussed at §6.1A and repeated here because it is the objection most likely to defeat the proposal in operation. The mitigations are that the resolution standard is *reasonableness* rather than accuracy, so no single register is made the gatekeeper, and that Limb two requires the unresolved state to be disclosed rather than the resolution to be certified. Neither is sufficient on its own, and a duty of this kind should not be enacted without a rule on who may hold the register and on what terms.

Why the duty is enforced by a regulator and not by an action in damages. Three reasons, and they follow from the analysis rather than from a preference for public enforcement.

First, the harm is diffuse and the loss is not the user's. A user who receives an unresolved citation marked as unresolved has suffered nothing; the cost falls on the court, the opposing party, and the reader who relies on the assertion downstream. A remedy available only to the person who has not been harmed will not be exercised, and §20.2A makes the same objection to the Product Liability Directive.

[151] Stuart Haber and W Scott Stornetta, 'How to time-stamp a digital document' (1991) 3 *Journal of Cryptology* 99; Ralph C Merkle, 'A Digital Signature Based on a Conventional Encryption Function' (1988) *Lecture Notes in Computer Science* 369; Arvind Narayanan and Jeremy Clark, 'Bitcoin's academic pedigree' (2017) 60 *Communications of the ACM* 36.

[152] Becker (n 146); George J Stigler, 'The Optimum Enforcement of Laws' (1970) 78 *Journal of Political Economy* 526.

[153] Paul R Milgrom, Douglass C North and Barry R Weingast, 'The role of institutions in the revival of trade: the Law Merchant, private judges, and the Champagne fairs' (1990) 2 *Economics & Politics* 1.

[154] Avinash Dixit, 'Trade Expansion and Contract Enforcement' (2003) 111 *Journal of Political Economy* 1293.

[155] Avner Greif, Paul Milgrom and Barry R Weingast, 'Coordination, Commitment, and Enforcement: The Case of the Merchant Guild' (1994) 102 *Journal of Political Economy* 745.

[156] Andrei Shleifer and Robert W Vishny, 'Corruption' (1993) 108 *Quarterly Journal of Economics* 599; Mancur Olson, 'Dictatorship, Democracy, and Development' (1993) 87 *American Political Science Review* 567.

Second, breach is observable in aggregate and not in the individual case. Whether a supplier's resolution process was reasonable is answered by examining the process across many outputs, which is a supervisory exercise. A single claimant cannot obtain that evidence and a court cannot economically construct it from one instance.

Third, private enforcement would price the duty by the size of the loss in the cases that happen to be litigated, and the analysis at §20.1 is that the loss in the litigated case is a poor proxy for the social cost. A civil penalty set by reference to the scale of emission is closer to the quantity the duty is meant to internalise.

Two further observations bear on feasibility.

What is a rule here and what is a standard. The duty is not uniformly one or the other, and the distinction should be drawn precisely, because the objection turns on it. Limb one's *conduct* is a bright line: either the supplier resolved the citation against a register before emitting it or it did not, and that is observable after the fact. The disclosure required by limb two is likewise a rule—the fact either accompanies the citation or it does not. Two things are standards, and both are so because the alternative is to legislate a technical specification that will be obsolete before the Act is in force: the question of which registers must be consulted and at what cost, which section 2 answers by reasonableness with three stated factors; and the persistence of the mark through copying and export, which section 3 qualifies by technical feasibility at the time of emission. Section 20.3 argues for a rule on the resolution *conduct* because that conduct is frequent and identical across users; neither of these two questions has that character, which is why neither is drafted as one.

That division takes account of Weisbach's argument in the tax context, which runs the other way from the direction in which it is usually invoked. His point is that *rules* are what sophisticated parties plan around, that specifying a rule for every transaction is what drives a tax code's complexity, and that a standard overriding the literal words is the cheaper answer where the alternative is indefinite specification.[157] Applied here, that is an argument against attempting to enumerate in advance which registers count for which class of authority, which is what section 2 declines to do. It is not an argument against the bright line in limb one, because the conduct there requires no enumeration: the register consulted may vary, the requirement to have consulted one does not.

Measurement of legal texts is now feasible enough to monitor compliance. The empirical legal literature has demonstrated that large bodies of legal text can be analysed computationally at scale: Livermore and colleagues on judicial writing, Ash and Hansen's survey of text algorithms in economics, and Bybee and colleagues' extraction of macroeconomic signal from business news all establish the capability.[158] Whether a jurisdiction's filings show a rising share of unresolved citations is, on that evidence, a measurable quantity—which matters, because a duty whose breach cannot be observed in aggregate cannot be shown to be operating.

Part VI—Conclusion

22. Conclusions on each element

This paper began from an asymmetry that is easy to state: the cost of producing a plausible claim has collapsed and the cost of establishing whether it is true has not. Four conclusions follow and they are of unequal strength.

The re-specification is secure, at the width §5 states it. In markets for goods whose quality the buyer cannot assess, what governs the economic effect of a technology acting on production cost is the *level* of verification cost, not the share of tasks the technology can perform. That follows from the structure of credence-goods markets as they have been understood since Darby and Karni, and requires no new assumption. Its consequence—that measured productivity can fall while output volume rises, and that this

[157] David A Weisbach, 'Formalism in the Tax Law' (1999) 66 *The University of Chicago Law Review* 860.

[158] Michael A Livermore, Allen B Riddell and Daniel N Rockmore, 'The Supreme Court and the Judicial Genre' (2017) 59 *Arizona Law Review* 837; Elliott Ash and Stephen Hansen, 'Text Algorithms in Economics' (2023) 15 *Annual Review of Economics* 659; Leland Bybee and others, 'Business News and Business Cycles' (2024) 79 *The Journal of Finance* 3105.

is arithmetic rather than mismeasurement—is stated in falsifiable form at §7 and the data cannot yet resolve it.

The stock–flow partition is the paper's strongest theoretical claim and it is a claim about cost structure rather than about magnitude. Verification cost divides. Establishing that an authority exists is mechanisable, amortises, and is properly described by the intangible-capital account the productivity literature offers. Establishing that a real authority supports the proposition it is advanced for is a fresh judgement about a fresh pairing every time, and no accumulated infrastructure retires it. The J-curve describes the first component and has been offered as an account of the whole. Its relative magnitudes are asserted at §8.1 on the strength of the mechanism at §7, and §16.3 supplies the first direct measurement of them on a real corpus: split on what similarity and image screening reaches, the component the amortised tool addresses roughly doubles over two decades while the component it does not reach rises between nine and twenty-five times, depending on the window. That is the direction the partition predicts. It is one corpus, one failure vocabulary, and a proxy for the division rather than the division itself, so it corroborates the partition without establishing its general magnitude.

The cohort study is the most secure element and the most limited. The measurement is reproducible from two public sources by a published method, and it produces a result that was not obvious: over two decades of sustained investment in verification infrastructure, the rate at which defective work is caught within a fixed window rose sixfold, and for integrity-type failures elevenfold. What it cannot do is say whether that is a deteriorating record or an improving apparatus, and the reason it cannot is the paper's own thesis. Every figure in Part IV is a floor and should be read as one.

The proposal is modest and is correctly sized to the analysis. It does not restore the world in which verification was cheap. What it does is stop the mechanisable component of the cost falling on whoever happens to be standing closest to the harm, and require that the state of verification travel with the assertion. The irreducible component it leaves exactly where it was, because the analysis gives no ground for moving it.

The finding at §16.2 governs the others and is restated here. A verification apparatus reports the defects it caught, and its reporting is read as a statement about the defects that occurred. The two quantities are separated by the rate of undetected production, which is the quantity the apparatus exists because nobody can observe. Investment in the apparatus does not close that gap: it raises the catch rate, which raises the reported number, which is then read as a change in the world rather than a change in the instrument. Twenty years of retraction data cannot distinguish an elevenfold deterioration in the record from an elevenfold improvement in detection. The courts, the audit profession and the compliance function are worse placed than scholarly publishing on this point, because they do not publish the numerator at all.

23. What would change the analysis

Three developments would require this paper to be substantially revised.

A verification technology that generalises beyond existence. If resolution against authoritative registers became universal, embedded and default, the mechanisable component of verification cost would collapse. That would not defeat the argument, because §8.1 confines it to the component a register cannot reach. It would defeat the argument only if the *residue* also became cheap, which requires a technology that evaluates the fit between an authority and a claim—a different thing from a register, and not one that presently exists.

An independent estimate of the production rate of defective work. This is what the identification result at §16.2 lacks, and it is not obviously unobtainable. A systematic re-examination of a random sample of published work by a method independent of the retraction apparatus would supply it, at considerable cost, once. It would resolve the central ambiguity in this paper's empirical contribution, and the fact that nobody has done it is itself an instance of the constraint being described.

Evidence that the drag falls with cumulative experience. The discriminating test at §4.3 is stated so that it can be run. If the measured verification burden in a sector declines as the technology is used more rather than scaling with the volume of output, the flow reading fails and the J-curve account is right. That

is the prediction the author would most expect to be wrong, and it is stated in the form most likely to make the error visible.

24. Summative remarks

Legal and market consequence attaches to what can be checked. Generative systems have reduced the cost of producing a claim without reducing the cost of checking one, and the sectors in which that matters most are those where the buyer cannot assess quality even after consumption.

The failures which follow are not visible as failures. At the point of receipt an unverified claim cannot be distinguished from a verified one; an absence of pricing for verification risk cannot be distinguished from a judgement that the risk is small; and a rising rate of caught defects cannot be distinguished from a rising rate of produced defects. In each case the institution's own reporting is consistent with two states of the world, and the state it reports is the one its instruments happen to record.

We have therefore insisted throughout on measurement, and on stating what a measurement cannot reach. The cohort table in Part IV does not solve the problem. It measures one credence market in which both terms of the ratio are public, and establishes that even there the two readings cannot be separated. That is a narrower conclusion than the one this paper set out to reach, and it is the one the data supports.

A companion paper puts the same question to a body of law rather than to a market, and arrives at a structurally similar result by a different route. Neither result depends on the other.

The derivation code for Part IV, the registers, and the verification apparatus described in the Appendix accompany this paper as supplementary material, itemised at Appendix §A.6. Every figure reported is reproducible from the stated sources by the stated method.

Bibliography

Books

Calabresi G, *The Costs of Accidents: A Legal and Economic Analysis* (Yale University Press 1970)
Gordon R, *The Rise and Fall of American Growth* (Princeton University Press 2016)
Pasquale F, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015)

Contributions to edited collections

Agrawal A, Gans J and Goldfarb A, 'Prediction, Judgment, and Complexity' in *Economics of Artificial Intelligence* (University of Chicago Press 2019) 89
Brynjolfsson E, Rock D and Syverson C, 'Artificial Intelligence and the Modern Productivity Paradox' in *Economics of Artificial Intelligence* (University of Chicago Press 2019) 23
Merkle RC, 'A Digital Signature Based on a Conventional Encryption Function' in *Lecture Notes in Computer Science* (Springer Berlin Heidelberg 1988) 369

Journal articles

Acemoglu D, 'The Simple Macroeconomics of AI' (2025) 40 Economic Policy 13
Acemoglu D and others, 'The Network Origins of Aggregate Fluctuations' (2012) 80 Econometrica 1977
Acemoglu D and Restrepo P, 'Automation and New Tasks: How Technology Displaces and Reinstates Labor' (2019) 33 Journal of Economic Perspectives 3
Acemoglu D and Restrepo P, 'The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment' (2018) 108 American Economic Review 1488
Aghion P and Howitt P, 'A Model of Growth Through Creative Destruction' (1992) 60 Econometrica 323

Akerlof GA, 'The Market for "Lemons": Quality Uncertainty and the Market Mechanism' (1970) 84 Quarterly Journal of Economics 488
Alchian AA and Demsetz H, 'Production, Information Costs, and Economic Organization' (1972) 62 American Economic Review 777
Arntz M, Gregory T and Zierahn U, 'Revisiting the risk of automation' (2017) 159 Economics Letters 157
Ash E and Hansen S, 'Text Algorithms in Economics' (2023) 15 Annual Review of Economics 659
Autor DH, Levy F and Murnane RJ, 'The Skill Content of Recent Technological Change: An Empirical Exploration' (2003) 118 Quarterly Journal of Economics 1279
Azoulay P and others, 'Retractions' (2015) 97 Review of Economics and Statistics 1118
Babina T and others, 'Artificial intelligence, firm growth, and product innovation' (2024) 151 Journal of Financial Economics 103745
Bajari P and Tadelis S, 'Incentives versus Transaction Costs: A Theory of Procurement Contracts' (2001) 32 RAND Journal of Economics 387
Baker GP and Hubbard TN, 'Contractibility and Asset Ownership: On-Board Computers and Governance in U. S. Trucking' (2004) 119 Quarterly Journal of Economics 1443
Barzel Y, 'Measurement Cost and the Organization of Markets' (1982) 25 Journal of Law and Economics 27
Baumol WJ, 'Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis' (1967) 57 American Economic Review 415
Becker GS, 'Crime and Punishment: An Economic Approach' (1968) 76 Journal of Political Economy 169
Bender EM and Koller A, 'Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data' (2020) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics 5185
Bender EM and others, 'On the Dangers of Stochastic Parrots' (2021) Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 610
Besen SM and Farrell J, 'Choosing How to Compete: Strategies and Tactics in Standardization' (1994) 8 Journal of Economic Perspectives 117
Bik EM, Casadevall A and Fang FC, 'The Prevalence of Inappropriate Image Duplication in Biomedical Research Publications' (2016) 7 mBio e00809
Bloom N and others, 'Are Ideas Getting Harder to Find?' (2020) 110 American Economic Review 1104
Bolton P, Freixas X and Shapiro J, 'The Credit Ratings Game' (2012) 67 Journal of Finance 85
Brodeur A, Cook N and Heyes A, 'Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics' (2020) 110 American Economic Review 3634
Brynjolfsson E, Li D and Raymond L, 'Generative AI at Work' (2025) 140 Quarterly Journal of Economics 889
Brynjolfsson E, Rock D and Syverson C, 'The Productivity J-Curve: How Intangibles Complement General Purpose Technologies' (2021) 13 American Economic Journal: Macroeconomics 333
Burrell J, 'How the machine "thinks": Understanding opacity in machine learning algorithms' (2016) 3 Big Data & Society
Bybee L and others, 'Business News and Business Cycles' (2024) 79 Journal of Finance 3105
Calabresi G and Melamed AD, 'Property Rules, Liability Rules, and Inalienability: One View of the Cathedral' (1972) 85 Harvard Law Review 1089
Calo R, 'Robotics and the New Cyberlaw' (2015) 103 California Law Review 513
Camerer CF and others, 'Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015' (2018) 2 Nature Human Behaviour 637
Candal-Pedreira C and others, 'Retracted papers originating from paper mills: cross sectional study' (2022) 379 BMJ e071517
Catalini C, Hui X and Wu J, 'Some Simple Economics of AGI' (2026) SSRN Electronic Journal
Christensen G and Miguel E, 'Transparency, Reproducibility, and the Credibility of Economics Research' (2018) 56 Journal of Economic Literature 920
Cobbe J, Veale M and Singh J, 'Understanding accountability in algorithmic supply chains' (2023) 2023 ACM Conference on Fairness Accountability and Transparency 1186
Cokol M and others, 'How many scientific papers should be retracted?' (2007) 8 EMBO Reports 422
Crafts N, 'Artificial intelligence as a general-purpose technology: an historical perspective' (2021) 37 Oxford Review of Economic Policy 521
Dahl M and others, 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models' (2024)

16 Journal of Legal Analysis 64
Darby MR and Karni E, 'Free Competition and the Optimal Amount of Fraud' (1973) 16 Journal of Law and Economics 67
DeFond M and Zhang J, 'A review of archival auditing research' (2014) 58 Journal of Accounting and Economics 275
Diver CS, 'The Optimal Precision of Administrative Rules' (1983) 93 Yale Law Journal 65
Dixit A, 'Trade Expansion and Contract Enforcement' (2003) 111 Journal of Political Economy 1293
Dulleck U and Kerschbamer R, 'On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods' (2006) 44 Journal of Economic Literature 5
Dulleck U, Kerschbamer R and Sutter M, 'The Economics of Credence Goods: An Experiment on the Role of Liability, Verifiability, Reputation, and Competition' (2011) 101 American Economic Review 526
Dye RA, 'Auditing Standards, Legal Liability, and Auditor Wealth' (1993) 101 Journal of Political Economy 887
Edwards L and Veale M, 'Slave to the Algorithm? Why a "right to an explanation" is probably not the remedy you are looking for' (2017) 16 Duke Law & Technology Review 18
Elliott M, Golub B and Jackson MO, 'Financial Networks and Contagion' (2014) 104 American Economic Review 3115
Eloundou T and others, 'GPTs are GPTs: Labor market impact potential of LLMs' (2024) 384 Science 1306
Else H, 'Abstracts written by ChatGPT fool scientists' (2023) 613 Nature 423
Emons W, 'Credence Goods and Fraudulent Experts' (1997) 28 RAND Journal of Economics 107
Fanelli D, 'How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data' (2009) 4 PLoS ONE e5738
Fang FC, Steen RG and Casadevall A, 'Misconduct accounts for the majority of retracted scientific publications' (2012) 109 Proceedings of the National Academy of Sciences 17028
Farrell J and others, 'Standard Setting in High-Definition Television' (1992) 1992 Brookings Papers on Economic Activity. Microeconomics 1
Felten E, Raj M and Seamans R, 'Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses' (2021) 42 Strategic Management Journal 2195
Frey CB and Osborne MA, 'The future of employment: How susceptible are jobs to computerisation?' (2017) 114 Technological Forecasting and Social Change 254
Furman JL, Jensen K and Murray F, 'Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine' (2012) 41 Research Policy 276
Greif A, Milgrom P and Weingast BR, 'Coordination, Commitment, and Enforcement: The Case of the Merchant Guild' (1994) 102 Journal of Political Economy 745
Grieneisen ML and Zhang M, 'A Comprehensive Survey of Retracted Articles from the Scholarly Literature' (2012) 7 PLoS ONE e44118
Griliches Z, 'Productivity, R&D, and the Data Constraint' (1994) 84 American Economic Review 1
Grimmelmann J, 'Copyright for Literate Robots' (2016) 101 Iowa Law Review 657
Grossman SJ, 'The Informational Role of Warranties and Private Disclosure about Product Quality' (1981) 24 Journal of Law and Economics 461
Grossman SJ and Hart OD, 'An Analysis of the Principal-Agent Problem' (1983) 51 Econometrica 7
Haber S and Stornetta WS, 'How to time-stamp a digital document' (1991) 3 Journal of Cryptology 99
Holmström B, 'Moral Hazard and Observability' (1979) 10 Bell Journal of Economics 74
Holmström B and Milgrom P, 'Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design' (1991) 7 Journal of Law, Economics, and Organization 24
Hsieh CT and Klenow PJ, 'Misallocation and Manufacturing TFP in China and India' (2009) 124 Quarterly Journal of Economics 1403
Ioannidis JPA, 'Why Most Published Research Findings Are False' (2005) 2 PLoS Medicine e124
Jensen MC and Meckling WH, 'Theory of the firm: Managerial behavior, agency costs and ownership structure' (1976) 3 Journal of Financial Economics 305
Ji Z and others, 'Survey of Hallucination in Natural Language Generation' (2023) 55 ACM Computing Surveys 1
Kabir S and others, 'Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions' (2024) Proceedings of the CHI Conference on Human Factors in Com-

puting Systems 1
Kalai AT and Vempala SS, 'Calibrated Language Models Must Hallucinate' (2024) Proceedings of the 56th Annual ACM Symposium on Theory of Computing 160
Kaminski ME, 'Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability' (2019) 92 Southern California Law Review 1529
Kaplow L, 'Rules versus Standards: An Economic Analysis' (1992) 42 Duke Law Journal 557
Klein B and Leffler KB, 'The Role of Market Forces in Assuring Contractual Performance' (1981) 89 Journal of Political Economy 615
Kleiner MM and Krueger AB, 'The Prevalence and Effects of Occupational Licensing' (2010) 48 British Journal of Industrial Relations 676
Leland HE, 'Quacks, Lemons, and Licensing: A Theory of Minimum Quality Standards' (1979) 87 Journal of Political Economy 1328
Liang W and others, 'GPT detectors are biased against non-native English writers' (2023) 4 Patterns 100779
Livermore MA, Riddell AB and Rockmore DN, 'The Supreme Court and the Judicial Genre' (2017) 59 Arizona Law Review 837
Lizzeri A, 'Information Revelation and Certification Intermediaries' (1999) 30 RAND Journal of Economics 214
Locke R and others, 'Beyond corporate codes of conduct: Work organization and labour standards at Nike's suppliers' (2007) 146 International Labour Review 21
Lu SF and others, 'The Retraction Penalty: Evidence from the Web of Science' (2013) 3 Scientific Reports
Magesh V and others, 'Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools' (2025) 22 Journal of Empirical Legal Studies 216
Mathis J, McAndrews J and Rochet JC, 'Rating the raters: Are reputation concerns powerful enough to discipline rating agencies?' (2009) 56 Journal of Monetary Economics 657
Milgrom PR, 'Good News and Bad News: Representation Theorems and Applications' (1981) 12 Bell Journal of Economics 380
Milgrom PR, North DC and Weingast BR, 'The role of institutions in the revival of trade: the Law Merchant, private judges, and the Champagne fairs' (1990) 2 Economics & Politics 1
Narayanan A and Clark J, 'Bitcoin's academic pedigree' (2017) 60 Communications of the ACM 36
Nelson P, 'Information and Consumer Behavior' (1970) 78 Journal of Political Economy 311
Nordhaus WD, 'Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth' (2021) 13 American Economic Journal: Macroeconomics 299
Noy S and Zhang W, 'Experimental evidence on the productivity effects of generative artificial intelligence' (2023) 381 Science 187
Olson M, 'Dictatorship, Democracy, and Development' (1993) 87 American Political Science Review 567
Open Science Collaboration, 'Estimating the reproducibility of psychological science' (2015) 349 Science
Otis N and others, 'The Uneven Impact of Generative AI on Entrepreneurial Performance' (2024) 2024 Academy of Management Proceedings
Polinsky AM and Shavell S, 'The Economic Theory of Public Enforcement of Law' (2000) 38 Journal of Economic Literature 45
Potoski M and Prakash A, 'Green Clubs and Voluntary Governance: ISO 14001 and Firms' Regulatory Compliance' (2005) 49 American Journal of Political Science 235
Raji ID and others, 'Closing the AI accountability gap' (2020) Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency 33
Rothschild M and Stiglitz J, 'Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information' (1976) 90 Quarterly Journal of Economics 629
Sanchirico CW, 'Character Evidence and the Object of Trial' (2001) 101 Columbia Law Review 1227
Selbst AD and Barocas S, 'The Intuitive Appeal of Explainable Machines' (2018) 87 Fordham Law Review 1085
Shapiro C, 'Premiums for High Quality Products as Returns to Reputations' (1983) 98 Quarterly Journal of Economics 659
Shavell S, 'Liability for Harm versus Regulation of Safety' (1984) 13 Journal of Legal Studies 357
Shleifer A and Vishny RW, 'Corruption' (1993) 108 Quarterly Journal of Economics 599
Short JL, Toffel MW and Hugill AR, 'Monitoring global supply chains' (2016) 37 Strategic Management

Journal 1878
Shumailov I and others, 'AI models collapse when trained on recursively generated data' (2024) 631 Nature 755
Simmons JP, Nelson LD and Simonsohn U, 'False-Positive Psychology' (2011) 22 Psychological Science 1359
Solow RM, 'Technical Change and the Aggregate Production Function' (1957) 39 Review of Economics and Statistics 312
Spence M, 'Job Market Signaling' (1973) 87 Quarterly Journal of Economics 355
Steen RG, Casadevall A and Fang FC, 'Why Has the Number of Scientific Retractions Increased?' (2013) 8 PLoS ONE e68397
Stigler GJ, 'The Optimum Enforcement of Laws' (1970) 78 Journal of Political Economy 526
Stiglitz JE, 'The Contributions of the Economics of Information to Twentieth Century Economics' (2000) 115 Quarterly Journal of Economics 1441
Syverson C, 'Challenges to Mismeasurement Explanations for the US Productivity Slowdown' (2017) 31 Journal of Economic Perspectives 165
Teece DJ, 'Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy' (1986) 15 Research Policy 285
Terlaak A, 'Order without law? The role of certified management standards in shaping socially desired firm behaviors' (2007) 32 Academy of Management Review 968
Tirole J, 'A Theory of Collective Reputations (with Applications to the Persistence of Corruption and to Firm Quality)' (1996) 63 Review of Economic Studies 1
Townsend RM, 'Optimal contracts and competitive markets with costly state verification' (1979) 21 Journal of Economic Theory 265
Van Noorden R, 'How big is science's fake-paper problem?' (2023) 623 Nature 466
Van Noorden R, 'More than 10,000 research papers were retracted in 2023 — a new record' (2023) 624 Nature 479
Veale M and Zuiderveen Borgesius F, 'Demystifying the Draft EU Artificial Intelligence Act' (2021) 22 Computer Law Review International 97
Wachter S, Mittelstadt B and Floridi L, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7 International Data Privacy Law 76
Weber-Wulff D and others, 'Testing of detection tools for AI-generated text' (2023) 19 International Journal for Educational Integrity
Weisbach DA, 'Formalism in the Tax Law' (1999) 66 University of Chicago Law Review 860
Williamson OE, 'Transaction-Cost Economics: The Governance of Contractual Relations' (1979) 22 Journal of Law and Economics 233

Reports, working papers and official documents

Autor DH, 'Applying AI to Rebuild Middle Class Jobs' (2024) NBER Working Paper 32140
Bessen JE, 'AI and Jobs: The Role of Demand' (2018) NBER Working Paper 24235
Best Law Firms, 'Law Firms Embrace AFAs, But Clients Want More Flexibility' (2025) Best Law Firms
Center for Scientific Integrity, 'The Retraction Watch Database' (2026) Center for Scientific Integrity
Crossref, 'Similarity Check' (2026) Crossref
Crossref, 'REST API' (2026) Crossref
Crossref, 'CrossCheck Plagiarism Screening Service Launches Today' (2008) Crossref
Hydari MZ and Muzaffar F, 'Redrawing the AI Map: A Theory of Accountability Boundaries in Agentic Ecosystems' (2026) 2605.23179 arXiv
STM Association, 'STM Integrity Hub' (2026) STM Association
The American Lawyer, 'With Law Firm AI Use on the Rise, Expect Alternative Fee Arrangements to Pick Up Steam in 2026' (2025) American Lawyer
Webb M, 'The Impact of Artificial Intelligence on the Labor Market' (2019) SSRN 3482150

Appendix—Verification apparatus

Every authority in this paper was checked against a primary or authoritative record before citation, by the following method. It is set out in full because the paper's argument makes the author's own practice material, and because the most instructive failures it records are the author's own.

Channel 1—Crossref. Each candidate reference was resolved against the Crossref REST API and accepted only where the returned title, first author and year agreed with the claim. Where a working paper and a journal article shared a title, candidates were ranked to prefer the version of record. Twelve entries initially accepted against a preprint were re-resolved to the published version.

Channel 2—Find Case Law. English authority was checked against a citation index of 75,837 judgments spanning 2001 to 2026 across thirteen court identifiers, harvested under a computational-analysis permission granted by The National Archives. The English authorities within the corpus were confirmed directly. *R (Bancoult) v Secretary of State for Foreign and Commonwealth Affairs (No 2)* was reported as falling outside the corpus, correctly: the index carries no House of Lords judgments.

Channel 3—manual. Law reviews, monographs, working papers whose working-paper form is the canonical citation, and official instruments were verified individually against the publisher, the reporter or the issuing authority, and the evidence recorded. Two categories mattered here and neither is reachable by an automated channel: the enacted European instruments discussed at §20.2A, which were read provision by provision against the official text, and the two public datasets underlying Part IV, whose coverage and field definitions were checked against their own documentation before either was used.

Continuous validation. Every citation in the manuscript is re-checked against the registers on each revision for title, year, volume, journal and first page. The build refuses to produce a manuscript in which a footnote is referenced but undefined, defined but unreferenced, defined in two parts at once, or pointed at by a cross-reference that resolves to nothing.

What the apparatus cannot do, established by nine classes of failure

The apparatus answers one question: does this authority exist, and is it what it is called. Everything below survived it. They are set out because they measure the limit of mechanical verification more precisely than any description of the method could, and because a paper making this argument that reported only its successes would be worth less than one that reported none of them.

Class one—correctly cited, wrongly used. Two load-bearing sources were cited for propositions they do not support. Dulleck, Kerschbamer and Sutter were cited for the proposition that removing verifiability degrades outcomes where liability is intact; their finding is that liability has a crucial effect and verifiability at best a minor one. Kalai and Vempala were cited to establish that fabricated citation is structurally irreducible; their result expressly excludes facts appearing more than once in training data, and gives references to publications as its example. Both were the hinges of their respective sections. Neither error was detectable by the apparatus, because both citations were perfectly formed and pointed at the works named.

A third instance is more embarrassing because it was a straight inversion of two findings that sit in adjacent sentences. Azoulay, Furman, Krieger and Murray measure the citation penalty to work by *other* authors intellectually proximate to a retracted paper; Lu, Jin, Uzzi and Jones measure the penalty to the retracted authors' *own* prior work. The paper had them the wrong way round, and had also given the first of them a co-author it does not have.

Class two—mechanically corrected into error. Acemoglu's 'The Simple Macroeconomics of AI' was first cited as (2025). Resolution against Crossref returned an issued date of August 2024—the advance-access date—and the citation was "corrected" to (2024). The version of record is volume 40, issue 121, of January 2025, and the original citation was right. The register was consulted, the register answered accurately, and the answer produced a worse citation than the one it replaced. The checker now carries both dates where they differ and accepts either, which is the only correct behaviour and was not the behaviour of any tool then available.

Class three—corrected in the register, uncorrected in the text. Two references were found to carry a digital object identifier belonging to a different article, and a third was attributed to the wrong three authors. All three were caught by the automated channel. But the correction propagated to the generated bibliography, which is built from the register, and not to the footnote, which is written by hand. For a period the paper's own bibliography contradicted its own footnote, and an earlier version of this Appendix reported an error as caught while the text still carried it. The instances are set out in the companion paper, which cites the works concerned; what belongs here is the failure mode, which is that a checked layer and an unchecked layer coexist in every scholarly apparatus in use.

Class four—the source was read, twice, and misread both times in opposite directions. This occurred on an instrument analysed in the companion paper, and is recorded here because the apparatus is shared and the lesson is not domain-specific. A short, free, publicly available regulation was described first as dispensing with a requirement it expressly imports, and then, on correction, as imposing a characterisation it merely permits. Nothing about the failure was a resource problem: the text was open on the screen each time. What was expensive was not obtaining it but attending to the difference between a permission and a characterisation across four provisions and a recital.

Class five—the statistic that confirms whatever it is pointed at. In the companion paper an empirical section was drafted three times. Two successive positive findings were computed, published internally, and then discarded: the first defeated by a base rate, the second by a degree distribution, each discovered only because the previous one had been. The published result is negative. It is recorded here because the same hazard applies to the cohort study in Part IV of this paper, and is the reason its identification problem is stated at §16.2 before its results rather than after them.

Class six—the tool that was checking the wrong thing. Two defects in the apparatus itself were found by examination rather than by the apparatus. The assembler renumbers footnotes into reading order and rewrites OSCOLA cross-references to match; it aborted on a dangling note but keyed its map on the source number alone, so where two parts of the manuscript happened to use the same number, every reference in the earlier part was silently redirected to the later part's authority. That is a citation which resolves, to the wrong thing—precisely the failure the script exists to prevent. Separately, its regular expressions matched digits only, so two footnote markers written with a letter suffix were invisible to the dangling-note check and reached the assembled text unresolved. Both are now checked and each aborts the build. The reproduction of each was performed by injecting the defect into a copy and confirming the abort; that is a manual step and is recorded as one.

Class seven—the check passed and the artefact was wrong. Two footnote markers were each used twice in the assembled manuscript. That is valid markdown, and the assembler's checks were satisfied: every note was defined, none was orphaned, no number was defined twice, and every cross-reference resolved. But the renderer emits a *fresh* note for each occurrence of a marker, so the rendered document carried two more notes than the manuscript, and every OSCOLA cross-reference pointing above the first duplicate landed one or two notes short. Thirty-two references in the published PDF and Word file named one authority and pointed at another. The markdown was correct throughout, which is why nothing caught it: the apparatus was validating the source and the reader was holding the output. The assembler now refuses to build where a marker is used more than once, which removes the cause. The consequence—whether the rendered file carries the same notes as the manuscript—is checked by reading the note table out of the Word file and the PDF and comparing it against the source, and that check is performed by hand at each render rather than by the build. It is recorded here as a manual step because describing it as automated would be the same species of overstatement this Appendix exists to catalogue.

This is the failure this paper could least afford, and it is the one that came nearest to publication. It is also the sharpest available statement of the argument: a verification system that checks the artefact it can see rather than the artefact the reader receives will report success indefinitely.

Class eight—an error introduced by the correction of another. In the course of correcting a misattributed author list, a citation was added whose author list the author invented: a 2021 physics paper was given a first author who did not write it and three authors where there are six. The register recorded the entry as verified against a digital object identifier whose own record contradicted it. The reason it passed

is precise and worth stating: the checker compared *surnames* only, so an invented given name attached to a correct surname was never tested. The check has been extended to compare the given name where the register holds one. This error was introduced in the round of revision devoted to eliminating errors of exactly this kind, on a source cited to justify withdrawing an overclaim—which is the strongest evidence in this document that verification cost is not eliminated by attention, discipline or the knowledge that one is being examined.

Class eight A—the instrument's own collision. Examination of the citation index built for this paper found that its normalisation function strips alphanumeric tier suffixes from a neutral citation, so that *[2024] EWCOP 53 (T1)* and *[2024] EWCOP 53 (T3)*—two different judgments—fold to a single key. The index therefore holds 75,837 entries against 75,838 distinct neutral citations in the harvest. The consequence is small in scale and exact in kind: a query for one of those two judgments resolves, and resolves to the other. That is the failure this paper is about, occurring inside the instrument the paper offers as a partial answer to it, and it was not found by the instrument.

What this establishes. A register can answer whether an authority exists and whether it bears the identifier given. It cannot answer whether the authority supports the proposition it is cited for, which of its dates is the one a reader needs, whether a correction made in one place has reached the other, whether a provision permits or characterises, whether a statistic is measuring anything, or whether the checking tool is checking what it claims. It cannot answer whether the document a reader holds says what the document the author checked says. Those are the questions on which every failure recorded here turned, and not one of them is a register question. That is the irreducible verification cost the paper is about, measured on the paper itself. The apparatus reduced the cost of checking by a large factor; it did not remove it, and what it left behind was not the easy part. This is offered as the paper's most direct piece of evidence, and it is evidence against the author rather than for him.

A.6 Supplementary material

The claim that the work is reproducible is only worth making if a reader is told what to reproduce it from. The following accompany this paper.

File	Lines	Contents
`build_cohort.py`	208	Derivation of every figure in Part IV from the Retraction Watch corpus and the Crossref annual counts, including the population restriction, the censoring rule and the integrity-reason vocabulary.
`cohort-table.json`	—	The output of the above: the full cohort table at k = 3 and k = 5 with the summary aggregates.
`refs_register.json`	251 entries	Every non-case authority, with the identifier it was resolved against and the fields on which agreement was required.
`cases_register.json`	42 entries	Every English authority, with its neutral citation as held in the harvested index.

File	Lines	Contents
`verify_refs.py`, `check_paper.py`, `check_structure.py`	460	The checking tools described above, including the checks added in consequence of classes seven, eight and eight A.
`assemble.py`, `build_apparatus.py`, `make_final.py`, `build.py`	594	The assembly pipeline, including the guards whose absence produced class seven.

The Retraction Watch corpus is distributed CC0 through Crossref and the Crossref annual counts are retrievable from its public API; neither is redistributed here, and `build_cohort.py` names the exact query used for each. The harvested judgment index is not redistributed, because the computational-analysis permission under which it was obtained does not extend to republication; `cases_register.json` carries the neutral citations, which is what a reader needs to repeat any individual check.