The Manufacture of Participation:

Revenue-Sharing Publishing, Search Governance, and the Substitution of Generated Answers for Published Documents

Flyxion
Independent Researcher

November 2026

Abstract

This essay reconstructs the transformation of the participatory web from a set of heterogeneous experiments in collective documentation into an infrastructure in which voluntary intellectual contribution is converted into speculative labor, governed by search and recommendation intermediaries, and increasingly displaced by generated answers. It begins with the publishing environments of the late 1990s and the first decade of the 2000s, including Everything2, h2g2, Squidoo, and HubPages, and distinguishes document-oriented participation, in which a contribution retains its usefulness independently of the author's continued activity, from engagement-oriented participation, in which reward depends on repeated performance under changing platform rules. It then analyzes the revenue-sharing model, the search-engine optimization checklist as an informal labor-management interface, and the formal structure of speculative earnings under heavy-tailed payoffs and selective publicity of winners. The essay treats Google's Panda and Penguin updates as a case study in private classification as infrastructural authority, traces the migration of conditional-reward structures to creator dashboards on social platforms and video services, and separates exploitative compensation from deceptive earnings representation and from legally actionable fraud. Finally, it argues that answer engines alter the epistemic role of search by replacing documents with generated summaries, and it formalizes the resulting loss of documentary specificity through a distinction among semantic equivalence, inferential equivalence, and documentary identity. A technical appendix gives a surplus accounting for platform intermediation and shows that questions about exact expression and provenance cannot in general be answered from semantic summaries alone. The essay does not claim a coordinated design. Its claim is that distinct firms with distinct objectives repeatedly produced the same structural asymmetry between those who supply the documentary record and those who govern its visibility, value, and presentation.

Keywords: user-generated content; speculative labor; search governance; Goodhart's law; creator economy; answer engines; documentary identity; provenance

Introduction

Two quite different pictures of the web coexist in common memory. In the first, the web is a vast and loosely organized archive to which anyone may add a document, a page, a correction, or an odd piece of local knowledge, and in which that addition remains findable and useful for years after its author has lost interest in it. In the second, the web is a set of engagement-oriented surfaces, markets for attention in which participants are told to post constantly, to track dashboards, and to chase thresholds that unlock a share of advertising revenue that most of them will not receive. The thesis developed here is that these pictures are not merely successive eras but stages of a single institutional process, and that the process can be described without recourse to any hypothesis of coordination. The organizing question is who determines what is worth producing, who receives the value that production creates, and whether the public retains independent access to the documentary record on which both depend.

The argument proceeds in five parts. The first part recovers the early participatory web as a body of organizational experiments, each of which was an answer to the problem of how unspecialized contributors might add to a common body of knowledge, and introduces a distinction between publishing a document and performing for an audience. The second part examines the commercial conversion of that participation into labor through revenue sharing, search-engine optimization guidance, and the statistical structure of creator earnings. The third part treats search engines as governors of publishing: not owners of the publications they rank, but arbiters of whether those publications are encountered at all. The fourth part follows the same structure of conditional reward into social-media and video platforms and asks when the promise of opportunity becomes a misleading inducement. The fifth part considers the transition from search engines that direct readers to documents to answer engines that substitute generated text for them, and argues that this transition has epistemic costs that are separable from its economic ones.

A formal appendix collects a surplus accounting for platform intermediation, a model of selective disclosure and entry, a model of dependence on a search intermediary, and a treatment of documentary sufficiency and evidentiary support. Three results carry the narrative. The first shows that selective publicity of winners can enlarge entry beyond what full information would support, given that prospective contributors treat the displayed winners as representative. The second shows that a summary which identifies certain pairs of corpora cannot answer questions about exact expression; it does not show that every generated summary loses such information. The third separates answer recoverability from evidentiary support, and distinguishes the local witness that supports a positive claim from the corpus-wide certificate that a negative claim requires. The historical sections point forward to these results where they bear on the argument, and the narrative can be read without them.

The word manufacture in the title is used in the sense of production by a structural process, in which participation is produced as an output of incentives, interfaces, and thresholds, and

not in the sense of fabrication by design. Likewise, the essay uses intermediary for a functional role, that of standing between producers and readers, and platform for an organizational form that may occupy that role; a search engine is an intermediary, and a revenue-sharing publisher is both a platform and an intermediary between its contributors and the search engine.

Four methodological commitments govern the discussion. The first is that structural asymmetry is not conspiracy. Firms that compete with one another, pursue different revenue strategies, and sometimes fail in each other's presence can nonetheless converge on arrangements in which risk, labor, and maintenance fall on contributors while discretion over eligibility, distribution, and reuse remains with the intermediary. Explaining the convergence requires only attention to incentives. The second commitment is to separate three categories that are often run together in polemical writing: conduct that is exploitative in an economic sense, conduct that is misleading in a way that would support a consumer-protection claim, and conduct that would satisfy the elements of fraud in a particular jurisdiction. These are different questions with different evidentiary requirements, and a conclusion on one does not settle the others. The third commitment is to separate historical observation from empirical proposition. Where the argument depends on distributions of earnings, on the scale of third-party schemes, or on the traffic consequences of generated answers, the essay states what is documented, what is plausible, and what would be required to settle the matter, and it supplies figures only where the evidence supports them.

A fourth commitment concerns the function of citations. Sources are used for three distinct purposes: to establish historical events, as with the Search Engine Land retrospectives and company announcements; to supply empirical measurements, as with the Pew study and the published eligibility thresholds; and to supply theoretical interpretation, as with the literature on aspirational labor, algorithmic management, and platform governance. A citation of the third kind supports a way of reading the evidence and not the evidence itself, and the propositions in the appendix are the essay's own contributions, not claims made by the works cited alongside them.

The word exploitation, where it appears, covers three analytically independent senses: an unequal division of economic surplus, uncompensated labor undertaken under distorted expectations, and the reuse of accumulated contributions beyond their original commercial context. An unequal division need not involve deception, deception can occur where a platform is unprofitable, and reuse can be contested even where contributors were once well compensated. The essay specifies which sense is at issue when it matters.

The essay also declines two simplifications. It does not treat low or zero compensation as the wrong in question, since people publish for many reasons and the open-source, public-domain, and encyclopedic traditions show that voluntary contribution without payment can be a durable and admirable form of collective work. And it does not treat every improvement in search quality or every convenience of generated synthesis as illegitimate. The difficulty lies in the conjunction of three features: a promise of economic opportunity that recruits effort, a concentration of

control over discovery that determines which effort is rewarded, and a reuse of the accumulated record that proceeds without corresponding changes in consent, compensation, or independent access. The essay calls this conjunction the intermediation triad. The sections that follow develop each feature in turn and then ask what arrangement would preserve the benefits of contribution without the conversion of contributors into inventory.

Part I
Before the Creator Economy

1 The Internet as an Accumulating Public Record

The original attraction of publishing on the web was not the prospect of an audience but the prospect of a persistent address. A person who knew something specific, whether the wiring of a particular model of radio, the history of a village, the meaning of a rarely used word, or a workaround for an obscure software defect, could write it down, place it at an address, and leave it. The document did not need an author who maintained a following or posted at intervals. It needed only to be reachable, which in turn depended on three properties of the early web that are easy to overlook because they have become so ordinary: hyperlinks that allowed one document to point at another, addresses that were meant to persist, and search that indexed documents by what they said rather than by who was currently speaking.

These properties together made the web an accumulating record. A contribution made in one year could answer a question asked in another, and the usefulness of the contribution was largely independent of its performance at the moment of publication. A page about a discontinued component might receive three visits a year and still be the only account of that component anywhere. Its value was not captured by any measure of immediate engagement, and for that reason it could not be justified within a model that treats attention as the unit of account. The early web was tolerant of contributions whose value was real, small, and delayed, because nothing in its basic architecture demanded that every contribution justify itself by current traffic.

A distinction follows that will recur throughout this essay: the distinction between a contribution's enduring informational value and its immediate commercial performance. The two can coincide, and when they do the case for commercial arrangements is strongest. They can also diverge sharply. The documentary record is full of items with high informational value and negligible commercial performance, such as a careful transcription of a local survey, a glossary of dialect terms, or a thorough negative result. Conversely, an enormous volume of material with high commercial performance, such as repetitive listings and search-targeted filler, has little informational value for anyone beyond the moment of the query. A publishing environment that rewards the second kind at the expense of the first changes the composition of the record, and

it does so without any person deciding that the record should be degraded. The mechanism is incentive alignment: contributors respond to what is rewarded, and over time what is produced reflects what is rewarded.

The significance of persistent addressing deserves particular emphasis because it underwrites later arguments about provenance. A document at a stable address can be cited, compared, and revisited, and its existence at a given time can be established by independent copies. An address is an anchor for accountability, since a claim made at an address can later be shown to have been made there. When addresses decay, or when content migrates from documents to feeds that are reordered by the platform, the capacity to establish what was said and when is weakened. The point is not nostalgic. It concerns the infrastructural conditions under which a public record can be audited by people other than those who maintain it.

2 Nodes, Lenses, Hubs, and Guides

The late 1990s and early 2000s produced a remarkable diversity of collaborative publishing experiments, and it is the diversity rather than any single success that merits attention. Everything2, launched in 1999, organized writing around nodes, each node being a titled topic to which individual contributors added writeups, so that the same title could accumulate multiple independent treatments. The structure encouraged a kind of plural documentation in which no writeup was authoritative and the node functioned as a meeting place for differing accounts. It also embedded a modest reputation system and a culture of idiosyncratic voice, which meant that the encyclopedic form was shaped by literary and humorous ambition as well as by informational aims.

Douglas Adams's h2g2, also launched in 1999, proposed a different organizing principle. The conceit of a collectively authored guide to practically everything was inherited from fiction, and its stated ambition included eventual access through portable devices and contextualization by geographic location, an anticipation of services that would not become ordinary for another decade. The project is instructive for what it assumed about contributors: that people would write because they enjoyed knowing and telling, that a shared editorial culture could emerge from conversation, and that the guide's value would lie in its breadth and its voice. Its institutional history, including its later stewardship under the BBC, shows how difficult it was to translate that assumption into a durable financial model, which is a difficulty that the commercial successors addressed by changing the assumption.

Squidoo, launched in 2005 by Seth Godin and colleagues, offered modular pages called lenses, each assembled from content modules that a contributor, called a lensmaster, could arrange without knowledge of markup. The metaphor was explicit: a lens was supposed to focus attention on a topic and show readers what they needed to see. Godin's own writing from the period frames the project as an experiment in amplifying the expertise of ordinary people, and the accompanying

booklet title, which asserts that everyone is an expert on something, states the founding premise directly [12]. In January 2006 the company announced a profit-sharing arrangement under which lensmasters would receive a portion of advertising and affiliate income generated by their lenses [30], a change that altered the project's character by attaching a monetary expectation to participation. HubPages, launched in 2006, developed its own system of hubs composed of capsules, again modular and again tied to a revenue share, and it would later acquire Squidoo in August 2014 in a transaction in which only a portion of Squidoo's content, chiefly accounts with featured lenses, was slated to migrate [38, 16]. Founding dates for these services are taken from secondary compilations and should be read accordingly.

Around these larger names stood an ecosystem of related services that differed in details but shared a family resemblance. Suite101, Associated Content, Helium, Mahalo, InfoBarrel, and others recruited contributors under combinations of upfront payment, per-view payment, and revenue share, and they differed in whether they curated by editorial review, by community voting, or by algorithmic filter. Mahalo, launched in 2007, presented itself as a human-curated search engine in which writers produced pages that matched popular search terms, a model that anticipated the later reversal in which the commissioning logic runs from a query list to a page rather than from a body of knowledge to a query [20]. Some of these services paid contributors modestly and immediately, others promised participation in future advertising income, and the variety is itself a record of the search for a sustainable arrangement.

The historical significance of this period lies in the diversity of organizational experiments rather than in their commercial outcomes. Each represented a different answer to a standing question: how might ordinary people contribute to a common body of knowledge, with what structure, under what incentives, and with what relation between contributor and platform? Nodes preserved plurality, guides preserved voice, lenses and hubs preserved modular composition, and each choice implied a different theory of what a contribution is. That the field later narrowed, with many of these answers disappearing as companies closed, merged, or pivoted, is a consequence that should not be read back as evidence that the surviving forms were the best answers. The narrowing was driven substantially by the dependence of the commercial forms on a single class of intermediary, namely the search engine, whose behavior shaped which experiments could survive.

3 The Difference Between Publishing and Performing

A useful analytic distinction can be drawn between document-oriented and engagement-oriented participation. In a document-oriented system, a contribution retains its usefulness independently of the author's continued activity. The author writes, the document persists, and the document's value to readers depends on its content and its findability rather than on the author's subsequent behavior. In an engagement-oriented system, by contrast, rewards accrue to repeated publication, interaction, audience maintenance, and responsiveness to changing platform incentives. The unit

of value is not the document but the stream, and the contributor's standing depends on sustained performance. A contributor who stops posting does not merely stop adding; in many such systems the accumulated standing decays, so that past work provides diminishing protection.

The distinction should not be treated as absolute. Early websites employed rankings, ratings, badges, and promotional mechanisms, and many contributors to document-oriented platforms cared a great deal about visibility. Conversely, some engagement-oriented environments, such as long-lived video archives, retain durable documents that continue to be discovered. The distinction is better understood as describing a dominant tendency in the reward structure rather than a partition of platforms. What matters is whether the design makes the persistence of a contribution independent of the contributor's ongoing labor, because that independence determines whether the contributor can stop without loss, whether contribution is compatible with a life that has other demands, and whether the record accumulates or churns.

Nevertheless the distinction supplies a framework for understanding the later transition toward feeds and creator dashboards. A feed is by construction an engagement-oriented surface: it orders items by recency and predicted response, it makes older items invisible by default, and it renders the contributor's recent activity as the primary signal of relevance. A dashboard translates this orientation into management information, presenting the contributor with measures of reach, retention, and follower growth, and thereby inviting self-management against metrics that the platform defines. The shift from document to stream is also a shift in the contributor's relation to time, because a document can be finished, whereas a stream imposes a continuing obligation. Historical accounts of the commercial reorganization of social media describe a comparable movement from loosely connected publishing toward platforms that engineer and monetize connectivity, and they stress that this development was contingent rather than inevitable [36].

The historical emergence of the idea that ordinary participation in a publishing community ought to become a source of personal income also bears on the argument. This idea has respectable antecedents in the tradition of professional authorship and in the observation that unpaid labor is socially undervalued, and the claim that contributors should share in the revenue they help to generate is not unreasonable on its face. The scholarly literature on digital labor has long noted that online participation can be simultaneously voluntarily pleasurable and structurally appropriated, and that these two descriptions are compatible [34]. What changed with revenue-sharing platforms was that the expectation of income became the principal organizing promise, so that participation was framed as the first step in a path toward earnings rather than as an end in itself. The consequences of that reframing for the quality of effort, the allocation of risk, and the relation between contributor and platform are the subject of the second part.

Part II
The Conversion of Participation into Labor

4 The Promise of Revenue Sharing

The economic model used by Squidoo, HubPages, and comparable services can be described in a few sentences, and its structure is more significant than its details. Contributors perform research, writing, editing, formatting, tagging, promotion, and maintenance. Platforms provide hosting, publishing tools, advertising infrastructure, and access to audiences, the last usually through search referrals. Advertising and affiliate revenue generated by pages is divided between the platform and the contributor according to a published formula, subject to eligibility conditions and minimum payout thresholds. The contributor thus receives a variable payment tied to the performance of pages, while the platform receives the remainder of the revenue and retains the pages themselves as an asset.

The allocation of risk in this arrangement is asymmetric. A contributor can invest substantial time without receiving meaningful compensation, since the revenue generated by an individual page depends on factors that the contributor does not control, including the platform's overall standing with search engines, the advertising market, and the contributor's topical choices. The platform, by contrast, bears marginal costs that are small relative to its potential gain, because each additional page adds to an inventory of potentially monetizable content at the cost of hosting and moderation. A platform with a large number of pages benefits from the right tail of the distribution of page performance, while the contributor who wrote a page in the left or middle of the distribution bears the full cost of the labor. Because the platform is paid in proportion to the sum across pages while each contributor is paid in proportion to the performance of their own pages, the platform diversifies and the contributor does not. This is the elementary logic of intermediation, and it is not itself objectionable: a bookseller, a record label, or a stock exchange all diversify across producers. What distinguishes the revenue-sharing platform is that the producers are also the sources of the inventory, that they supply labor rather than capital, and that they are typically unable to diversify across platforms because the work of producing a page is specific to the platform's format. Accounts of platform capitalism describe this general pattern, in which a platform owns the infrastructure on which others' activity is conducted and extracts data and revenue from it [31].

The analysis requires that several quantities be kept distinct. Gross advertising revenue is the amount the platform receives from advertisers and ad networks in connection with a page. The contributor revenue share is the fraction of some defined portion of that revenue that is allocated to the contributor, and the defined portion may be net of costs, may exclude certain revenue categories, and may be calculated from a base that the contributor cannot observe. Eligibility thresholds determine whether the share is paid at all, and minimum payout thresholds determine

when accumulated earnings are transferred, so that a contributor whose pages earn small amounts may never reach the payout level. Actual earnings, finally, are what the contributor receives over a period, and they may be substantially lower than the nominal share would suggest. A discussion that treats the headline percentage as the contributor's return will systematically overstate it, and an adequate account must follow each quantity through to the contributor's net position.

An illustration from the early period is the removal of aggregate revenue figures from a platform's public pages. When Squidoo launched, its home page reportedly displayed aggregate royalty figures; by February 2006 these had disappeared from the main page after a redesign, a change noted by an observer who remarked that opening with transparency and then trending toward closure is worse than never having been transparent [9]. The platform responded that such statistics would surface elsewhere. The episode is minor and rests on a single contemporary observation, so it is offered as an illustration of how disclosure of aggregate earnings can be withdrawn or relocated, and not as evidence of a general pattern, which would require further documented cases. The evidentiary significance of such disclosures is taken up in the discussion of deception below. The chronology alone does not show that contributors' motivations changed in the manner described; establishing that would require recruitment materials, archived contributor guides, and participant testimony, which this essay does not assemble.

5 The SEO Checklist as a Labor-Management System

The ordinary contributor's experience of search-engine optimization was a distinctive feature of this period and may be the most revealing, because it shows how a platform could direct labor without employing anyone. Contributors were encouraged to select keywords, modify titles, adjust headings, add tags, generate backlinks, monitor rankings, and revise content according to search performance. Guidance came from platform documentation, from forums, from third-party tutorials, and from a thriving trade in advice books. The instructions varied in sophistication, but they shared an imperative form: do this, and your page may rank, and if it ranks it may earn.

The important argument is that these instructions constituted a form of work organization. A firm that employs workers can assign tasks, supervise their execution, and compensate them for time. A platform that publishes the work of unpaid contributors cannot assign tasks in this way, but it can publish a checklist, and the checklist can have much the same effect on behavior. The contributor who follows the checklist is performing tasks that serve the platform's commercial objectives, namely the production of search-friendly inventory, without being employed, directly supervised, or guaranteed payment. The checklist translates the platform's objectives into discrete actions that individual contributors are expected to perform, and it does so under conditions in which noncompliance is penalized not by dismissal but by invisibility.

The notion of an informal management interface captures the point. Management in the ordinary sense involves setting objectives, specifying methods, monitoring performance, and applying consequences. The SEO checklist set objectives through the definition of what counts as a well-optimized page, specified methods through its enumerated steps, monitored performance through rank trackers and platform statistics, and applied consequences through the allocation of traffic. All of the functions of management were present, but they were distributed across the platform, the search engine, and the advice industry, and none of them was formally the responsibility of any party. The contributor experienced the system as a set of constraints that were not exactly commands, which is characteristic of governance by incentive rather than by order. The comparison is with research on algorithmic management, which identifies mechanisms by which algorithmic systems direct, evaluate, and discipline workers [17]. That literature concerns settings with conventional employment relations. Extending its categories to unpaid contributors governed by published guidance and ranking outcomes is the argument of this essay and is not a claim made by those authors. A related body of work shows how cultural producers become dependent on platform governance, metrics, and infrastructure, so that their products are contingent on arrangements the platform may change [24].

One consequence was the displacement of the author's original informational purpose. A contributor who begins with a desire to explain something learns that the page will be rewarded to the extent that it matches the language of searches, includes the keywords that searchers use, and is surrounded by links from other pages. These requirements are not intrinsically opposed to informing readers, and sometimes they coincide with it, since a title that matches what people ask can help them find an answer. But the optimizing logic tends to detach the page's structure from the author's understanding of the topic and to attach it instead to a model of the search engine. The result is a systematic pressure toward pages that are written for a classifier, and the downstream effect on readers is a body of text whose organization reflects the ranking function rather than the subject. The effect on the contributor is a change in self-conception, from someone who writes about a topic to someone who produces entries for an index, and the effect on the platform is the accumulation of inventory whose value depends on a ranking function that it does not control. The search engine itself later described, as a problem to be reduced, sites created primarily to match very specific search queries and content that feels made for search engines rather than for people [35]. That description acknowledges the behavior at issue without conceding that ranking incentives encouraged it.

6 Speculative Labor and the Visibility of Winners

The economics of contributor earnings can be given a formal account that clarifies what is and is not being claimed. Let contributor i invest h_i hours of labor and incur additional costs c_i, such as tools, subscriptions, or purchased materials, and receive a random payment R_i. Let w_i

denote the opportunity cost of an hour of the contributor's labor, that is, the value of the best alternative use of that hour. The expected private return from participation in monetary terms is

$$\mathbb{E}[\Pi_i] = \mathbb{E}[R_i] - w_i h_i - c_i. \tag{1}$$

The formula is a bookkeeping identity and has no content until assumptions are placed on the distribution of R_i and on how contributors form beliefs about it. Two observations nevertheless follow at once. A platform can recruit large numbers of contributors even when $\mathbb{E}[\Pi_i]$ is negative, provided that participants overestimate $\mathbb{E}[R_i]$, that they place value on nonmonetary benefits, or that their opportunity cost w_i is sufficiently low. And a platform's profitability depends on the aggregate across contributors rather than on any individual's return, so that it can profit from a population in which the typical participant loses.

The average earnings of successful creators must be distinguished from the expected earnings of a new entrant. The two are different quantities for a reason that is often obscured in publicity: the earnings that become visible are drawn from a selected subset. Earnings distributions in creative and attention markets are typically heavy-tailed, in the sense that a small fraction of participants receives a large fraction of the total, a pattern that has been analyzed in the economics of superstars, where small differences in talent or luck are magnified into large differences in reward when markets are scaled by cheap reproduction [28]. In a heavy-tailed distribution the mean is dominated by rare large values and can lie far above the median, so a statement about the mean can be true while the typical participant earns a small fraction of it. An entrant who infers from published success stories that the mean is attainable is making an error that is structural rather than careless.

Survivorship bias compounds the error. The classical illustration is the analysis of aircraft that returned from missions, in which the damage visible on survivors was misleading as a guide to where armor was needed, since the aircraft hit in the most critical places had not returned [18]. The creator economy presents a related selection problem. The creators whose results are seen are those who persisted and succeeded, those whose results are published are those whom the platform chooses to promote, and those who tried and stopped are by construction unobserved. Selective publicity of winners therefore operates twice, once through survival and once through curation, and the combination can make the apparent odds of success far better than the actual odds without any single false statement having been made. The difference between possible and probable success is the central issue, and Proposition 1 in the appendix states one mechanism by which selective publicity of winners enlarges entry beyond the informed level. That a person can succeed is established by a single example; whether a person should expect to succeed depends on the distribution. The sociological literature on aspirational labor documents the same structure among aspiring social-media professionals, who perform extensive unpaid work in pursuit of careers that comparatively few achieve and who are encouraged to regard that work as an investment [5]. The present argument overlaps with that account at the level of the

individual participant and extends it in three directions: historically, to the revenue-sharing publishers that preceded social media; institutionally, to the governance of discovery by search intermediaries; and epistemically, to the substitution of generated answers for the documents on which the arrangement rests.

Importantly, the mathematical model must allow contributors to receive genuine nonmonetary benefits, and the exploitative mechanism cannot be inferred simply from low or negative monetary returns. Contributors may value the practice of writing, the satisfaction of explaining, the fellowship of a community, the accumulation of skill, the creation of a portfolio, or the chance to be useful to strangers. Let b_i denote the subjective value of these benefits in monetary equivalents. A rational contributor participates when $\mathbb{E}[R_i] + b_i - w_i h_i - c_i \geq 0$. A population in which most members have negative monetary return but positive total return is not, by that fact alone, being exploited; they may be buying something they want. The analysis of exploitation must therefore turn on other features: whether contributors' beliefs about $\mathbb{E}[R_i]$ were induced by information that the platform knew to be unrepresentative, whether the nonmonetary benefits were offered as part of the pitch while the monetary promise remained the principal inducement, whether participants were free to leave without losing accumulated work, and whether the platform captured value from the accumulated work in ways contributors had not anticipated. The model treats b_i as given, but it may be endogenous to platform framing. A contributor who writes for the pleasure of explaining and one who writes because a dashboard has presented the activity as building a personal brand may both report nonmonetary benefit, and in the second case the benefit may be partly produced by the platform's own presentation. Evidence that separates the cases, such as whether the reported benefit persists when the platform's framing is withdrawn, for instance in comparisons of contributors before and after a platform removes its earnings dashboard, or between contributors on otherwise similar services that do and do not present such framing, is what the normative assessment requires, and it is at this point that the normative question arises of whether preferences produced by the platform's own presentation should count as the contributor's own. The formal appendix returns to this problem and gives a proposition showing how selective disclosure can enlarge entry beyond the informed level.

Part III

Search Engines as Governors of Publishing

7 Google as a Research Instrument

Before discussing the economic influence of search engines, it is necessary to establish what conventional search made intellectually possible, because the later argument about substitution depends on what is lost. Google's historical value extended well beyond retrieving answers

to questions. It allowed users to investigate the documentary record: to discover unusual expressions, to trace quotations to their sources, to compare terminology across communities, to locate prior uses of a phrase, and to encounter documents that did not conform to the user's initial expectations. The search results page was a map of what had been written, and the user's skill lay in reading that map, noticing what was present and what was absent, and following links into documents whose existence the user had not suspected.

A distinction is essential here: searching for a proposition differs from searching for its expression. A person who asks whether a claim is true seeks a proposition and may be satisfied by any adequate statement of it. A person who asks whether a particular sentence, formula, definition, or term has appeared before seeks an expression, and no paraphrase will do. The first is a question about the world; the second is a question about the record. A paraphrase of *Moby-Dick* is not *Moby-Dick*, and a passage that is semantically similar to a sentence in some earlier document is not evidence that the earlier document contains that sentence. Many research practices turn on the second kind of question: establishing priority of a scientific term, verifying a quotation attributed to a historical figure, checking whether a theorem was stated in a given form, determining whether a phrase was in use before some date, or confirming that a passage was not copied. In each case the evidentiary value lies in the specificity of the match.

Exact-match retrieval, provenance, and independent document inspection are therefore epistemic capabilities rather than interface preferences. Exact-match retrieval allows negative results to be meaningful: if a distinctive phrase returns no documents, that fact carries information about the indexed record, bounded by the coverage of the index. Provenance allows a retrieved item to be traced to a source that can be examined. Independent document inspection allows the reader to see the context, the surrounding argument, the date, the author, and the apparatus of citation, and thereby to assess the item rather than accept it. A system that supplies answers without these capabilities may be accurate and useful for some purposes while removing the means by which its accuracy could be checked for others. The point is not that earlier search was infallible, since index coverage was partial, ranking was opaque, and results were polluted by manipulation. The point is that the form of the interaction preserved the user's ability to read the primary documents, and that ability is what makes research possible.

8 Panda, Penguin, and the Reorganization of the Web

Google's major search updates during 2011 and 2012 reorganized the commercial web, and they provide a case study in how a change to a private ranking function can restructure an industry. Panda, introduced in February 2011, targeted low-quality content and substantially affected a number of user-generated publishing platforms. Penguin, introduced in April 2012, targeted manipulative linking practices and thereby addressed a different component of the optimization ecosystem.

A convenient historical case study is the retrospective published by Search Engine Land in February 2013, which examined the search visibility of twenty-two sites identified as early Panda losers, using data from the visibility index of Searchmetrics. The retrospective compared visibility on three dates: shortly before Panda, shortly after it, and the time of writing. Its principal finding was that none of the twenty-two sites had returned to its pre-Panda visibility, and only two had improved relative to their post-Panda position [19]. The companion article reported that HubPages visibility was approximately 62 percent lower than before Panda, that Mahalo's was approximately 92 percent lower, and that Suite101's had fallen by approximately 96 percent [20, 19]. The same coverage records that Mahalo's founder announced layoffs of roughly a tenth of its staff within a week of Panda's launch, citing a significant dip in traffic and revenue, and that HubPages recovered some visibility in 2011 by moving content to subdomains, a recovery that had largely disappeared by 2013 [20].

These figures require careful interpretation. A visibility index is a modeled measure of how prominently a site's pages appear in search results for a sample of queries; it is not a count of visits, and it is not revenue. A site could lose visibility without losing a proportional amount of traffic if its traffic had a significant direct or referral component, and it could lose traffic without losing a proportional amount of revenue if its monetization improved. Conversely, a loss of visibility could understate a loss of revenue if the lost queries were the most valuable ones. The essay therefore treats the reported declines as documented changes in modeled search visibility and not as direct measures of business outcomes, while noting that the business consequences for several of these firms were severe, and that HubPages' own later history, including its acquisition of its largest competitor and its eventual sale, is consistent with a prolonged period of constraint.

Panda is treated here as an unusually visible demonstration of dependency and not as the single cause of the decline or consolidation of revenue-sharing publishers. Changes in advertising markets, editorial quality, mobile use, social discovery, and management also bore on those firms, and the argument requires only that their economic viability was exposed to an external classification decision.

Three questions deserve investigation, and the existing evidence answers them unevenly. The first is whether these changes disproportionately affected independent publishers relative to large institutional ones. Reporting at the time noted both winners and losers, with large video and reference properties among those who gained visibility, and the structure of a quality classifier that evaluates whole sites may favor sites with stronger institutional signals (an inference of this essay and not a finding of the reporting cited), but a firm conclusion would require a study of matched sites that has not been supplied here. The second is whether the updates reduced genuinely manipulative practices. There is strong reason to think that some targeted practices, such as large-scale generation of thin pages, were real and harmful to searchers, and some platforms had indeed tolerated them; a contemporaneous trade report stated that Google had downgraded Squidoo's rankings by about thirty percent for tolerating search spam, with the

company then removing offending pages [23]. The third is whether the absence of transparent ranking criteria complicated adaptation. It plainly did. A publisher cannot adapt to a rule it cannot read, and the guidance supplied by the search engine, being general, left sites to infer requirements from outcomes. That inference is itself the feedback loop with which the next chapter is concerned.

Penguin, introduced in April 2012, addressed a different component of the optimization ecosystem. Backlink generation was among the tasks that the optimization checklist asked of contributors, and a ranking system that began to penalize manipulative linking turned a recommended practice into a potential liability. This essay does not document the specific effects of Penguin on revenue-sharing publishers, and the observation that guidance followed in good faith could later be penalized is an inference from the structure of the arrangement and not an empirical finding. Its relevance is to the adaptation problem discussed in the next section: a contributor following published advice cannot know which elements of it the classifier will later treat as manipulation.

9 Private Classification as Infrastructural Authority

The historical cases support a general institutional argument. A search engine need not own a publication to influence its economic survival. It can govern discoverability through ranking, indexing, eligibility, and classification, and for a publisher whose traffic depends predominantly on search referrals, control over discoverability is control over the conditions of existence. The relation is one of dependency without ownership: the publisher is formally independent, bearing its own costs and making its own editorial decisions, while its viability is determined by a classification that another party makes according to criteria that the publisher cannot fully observe.

The distinction between formal authority and effective control is central. Google could not generally prohibit someone from publishing an article, and it did not claim to. But it could substantially influence whether readers encountered that article, and in a web where the overwhelming majority of discovery passed through a small number of intermediaries, influence over discovery was in practice influence over publication. The language of platform studies helps here: platforms present themselves as neutral conduits while exercising substantial discretion over what flows through them [10]. The more general literature on opaque algorithmic classification has argued that when private actors make consequential decisions by means of undisclosed procedures, the affected parties lose the ability to contest or even to understand those decisions [25]. A ranking is such a decision. It is made at scale, continuously, without explanation to the affected party, and with consequences for livelihoods. Work on the relevance of algorithms emphasizes that deciding what counts as relevant is itself an exercise of institutional power over what becomes visible [11], and work on the platformization of cultural production shows how

producers must continually adapt to platform requirements that they neither set nor control [24].

Publishers adapted their editorial practices to inferred ranking requirements, and this adaptation created a feedback loop between the classification and the material being classified. If a classifier rewards pages with certain observable features, publishers produce pages with those features, which alters the distribution of pages on which the classifier operates, which in turn requires revision of the classifier, and so on. In general the loop need not have a stable endpoint, and it is unstable when publishers adapt faster than the classifier is revised, and each iteration imposes adaptation costs on publishers who are not party to the process. A publisher who invests in a format that was favored at one moment may find the investment penalized at the next. This is a cost of governance by opaque rule, and it is borne disproportionately by those least able to absorb it.

The mechanism connects to Goodhart's law and its relatives. In its original formulation as applied to monetary policy, the observation is that a statistical regularity tends to collapse once pressure is placed upon it for control purposes [13], and a widely cited paraphrase holds that when a measure becomes a target it ceases to be a good measure [32]. The social-science version, associated with Campbell, states that the more a quantitative indicator is used for decisions, the more subject it is to corruption pressures and the more apt it is to distort the processes it is intended to monitor [3]. Search ranking is a case in point: the features that correlate with quality become targets once ranking depends on them, and the correlation weakens. Appealing to this law does not imply that every attempt to improve search quality is illegitimate. A search engine that failed to respond to manipulation would degrade, and users have a real interest in not being served filler. The argument is rather that the response to manipulation, when it takes the form of an undisclosed reclassification with sudden effects, places the cost of the adjustment on publishers in ways that a more transparent process would not, and that this distribution of cost is a design choice rather than a necessity.

The structural argument does not require that a single intermediary dominate, but it is sensitive to market structure. Competition among search engines, and among platforms for creators, bears on how much discretion any one of them can exercise. A publisher that depends on one search engine has few alternatives if that engine reclassifies it, and a publisher that depends on a smaller engine may have fewer still, since the traffic available from alternatives is itself limited. Where several intermediaries are close substitutes for readers and for publishers, and where moving between them is cheap, the discretion described here is disciplined by the possibility of exit. Where one intermediary supplies most discovery, or where accumulated standing such as followers, history, and ranking cannot be moved, exit is costly and discretion is correspondingly less constrained. The degree of concentration in each setting is an empirical matter that this essay does not measure, and the claims made here should be read as conditional on it.

Part IV
The Industrialization of Aspiration

The fourth part follows the structure of conditional reward into a different institutional setting, and two things change. The unit of production shifts from the page held as inventory to the stream sustained as performance, so that the contributor's standing depends on continuing activity and not on a persisting document. And the gatekeeper changes: where revenue-sharing publishers depended on a single external search engine, social and video platforms combine discovery, recommendation, and monetization thresholds within one institution, so that the party that decides what is seen is also the party that decides what is paid. In the vocabulary of Section 3, this is the passage from document-oriented to engagement-oriented participation, carried from the publishing platform to the feed.

10 From SEO Checklists to Creator Dashboards

The optimization practices of article-publishing services migrated to social-media platforms, and the migration changed their form without changing their structure. Squidoo's search guidance told contributors which words to place in titles and how to organize modules; Facebook's Professional Mode, creator dashboards on video and short-form services, and similar interfaces present performance metrics, posting recommendations, follower thresholds, and monetization eligibility requirements. The format has moved from checklist to dashboard and from instruction to indicator, though the checklist has not disappeared, and the underlying relation is the same: the platform communicates what behaviors it will reward, and participants adjust. The aspirational character of this adjustment has been documented in studies of social-media work [5], and the movement from open publishing to engineered connectivity forms its historical background [36].

The common structure is conditional reward. Participants are encouraged to perform repeated tasks that may improve their prospects of compensation, while the platform retains discretion over eligibility, distribution, and program rules. The elements of this structure can be listed in prose. A participant must reach a defined threshold, which may be a number of followers, hours of viewing, or posts within a window, before earning anything. The threshold is set by the platform and may change. Once the threshold is reached, the participant becomes eligible for a share of revenue, but distribution of the participant's work to audiences remains governed by recommendation systems that the platform controls. Program rules can be revised, content can be demonetized, and accounts can be restricted, with limited explanation. At each stage the participant bears the cost of effort while the platform holds discretion, and the dashboard gives the experience of agency by exposing quantities that the participant can attempt to raise.

The geographical distribution of aspiring creators is a component of the story that requires care. Participation in countries such as India and Bangladesh is sometimes asserted, on the

reasoning that the relative value of potential platform income is higher and that informal communities have formed around the objective of reaching monetization thresholds, but this essay cites no study of its scale or character and treats it as a research hypothesis on which the general argument does not depend. Within any such communities it is important to separate official monetization programs run by platforms from third-party engagement clubs, which exchange reciprocal interaction to inflate metrics, and from potentially deceptive earnings schemes that sell courses or services promising access to income. The categories differ in who is the intermediary, in what is promised, and in who is responsible for any misrepresentation. A platform's official program is the platform's responsibility; a third-party club is a secondary market that exists because the platform's thresholds create demand for engagement; an earnings scheme that sells guarantees may itself be fraudulent while depending on the platform's real thresholds for its plausibility.

The essay avoids unsupported numerical claims about the total population of unpaid aspiring creators. Estimates of this population circulate widely, but they typically rely on extrapolations from surveys of limited coverage or on counts of accounts that do not distinguish active participants from dormant ones. What can be established with present evidence is structural: the thresholds exist, the conditional-reward form is common across services, and the effort required to reach them falls on participants. The empirical question of how many participants fall below the thresholds, how long they persist, and what they spend remains an open research problem that would require platform data or carefully designed independent sampling.

11 YouTube and the Economics of the Unpaid Archive

YouTube is a particularly important example because it combines an enormous user-generated archive, concentrated creator earnings, formal monetization thresholds, and a highly developed recommendation system. It is also the service in which the relation among uploads, viewing hours, advertising inventory, and platform revenue is most visible in the public rules. Eligibility for revenue sharing through the Partner Program has been tied since 2018 to a combination of subscriber count and accumulated viewing time, and the current published requirements are one thousand subscribers together with either four thousand valid public watch hours in the preceding twelve months or ten million valid public Shorts views in the preceding ninety days, with an earlier tier of limited features available at lower thresholds [37]. YouTube announced on August 10, 2026 that, from February 1, 2027, new applicants for ad and Premium revenue sharing will need eight thousand qualified watch hours in the preceding 365 days or twenty million qualified Shorts views in the preceding ninety days, that the subscriber requirement is unchanged, and that existing Partner Program members retain their entry status [33]. According to coverage of the announcement, existing members must also maintain ten million qualified Shorts views in each ninety-day window to continue earning from Shorts, while channels below

that level remain in the program and can continue to earn from long-form content. Announced changes may be delayed or revised before taking effect, and the figures given here are those announced as of November 2026. The significance of the change does not depend on its exact details. Because the program's entry requirements had last changed significantly in 2018 [22], the revision is an unusual event, but together with the earlier change it illustrates the platform's discretion over eligibility: a creator who is building toward a published target may find that new applicants face a higher one.

An inquiry into the economics of the unpaid archive requires attention to what the available statistics can and cannot support. The relevant questions concern the distribution of channel earnings, the proportion of active creators receiving compensation, and the relationship between uploads, viewing hours, advertising inventory, and platform revenue. Each is a question about a different population, and the populations must not be conflated. A video is a unit of content, a channel is a unit of account and eligibility, and an active creator is a person who has published recently. A claim about the percentage of videos that have never earned revenue cannot be inferred directly from statistics about monetized channels, because a monetized channel may hold many videos that earn nothing, while an unmonetized channel's videos earn nothing by definition. Likewise, a statement about the share of channels that meet the thresholds does not give the share of people who have tried and stopped, since abandoned channels and one-time uploaders may dominate the count of channels without representing aspiring professionals. Careless inference across these units is the main source of dubious figures in popular accounts.

The structure of the platform nevertheless makes certain claims secure. Because eligibility is threshold-based, a creator below the threshold does not receive a share of advertising proceeds. YouTube's terms of service have since late 2020 granted it the right to monetize content on the service, and the company stated that ads might appear on videos from channels not in the Partner Program, with no revenue share for those uploaders; the terms as quoted in trade coverage state that the agreement does not entitle the user to any payments [21]. The claim rests on trade coverage of the terms and not on the terms themselves. The public record therefore establishes that advertising can be served against the content of non-partner channels without payment to the uploader. What it does not establish is how often this occurs, how much revenue it generates, or how it is distributed across channels. The documented practice is that the platform can retain and distribute contributions before the creator becomes eligible for a share of advertising proceeds, and those contributions can add to the archive that attracts viewers and informs recommendations. Whether that situation is a defect is a question of design and expectation. A viewer who watches an unmonetized video may contribute to platform engagement as a viewer of a monetized video does, though whether it generates revenue depends on whether advertising is served, and the platform's decision to share revenue only above a threshold is a choice about how to divide the proceeds of the aggregate. A defender of the arrangement can say that the threshold excludes fraud and low-quality material and that unmonetized creators receive hosting,

distribution, and tools without charge. A critic can say that the unmonetized archive may be an asset of the platform created by participants who were not compensated and who may not have anticipated its later uses. Both statements are partly true, and the weight given to each depends on what contributors were told.

The closing question is whether the system is better understood as a publishing platform, a competitive entertainment market, or a speculative labor arrangement. The answer is plausibly that it is all three, and that different participants experience it as different things. For a hobbyist who posts occasionally with no thought of income, it is a publishing platform. For a professional production company, it is a competitive market with particular rules. For an aspiring creator who invests months of effort in the hope of crossing the threshold, it has the features of speculative labor: effort is advanced against an uncertain, heavy-tailed return whose parameters the participant cannot observe. The platform's rhetoric tends to address all three audiences at once, and the ambiguity is itself an economically significant feature, because it allows a promise made to the third audience to be defended by reference to the experience of the first.

12 Opportunity, Exploitation, and Deception

The essay now turns to a stronger allegation: that the promise of economic opportunity has been used to recruit labor under misleading conditions, and that this amounts to fraud. The allegation must be examined without assuming the legal conclusion, because several different claims are often bundled under the word.

Five categories should be distinguished. Exploitative compensation concerns the division of value between platform and contributor, and it can exist without any misrepresentation if contributors accept terms that are accurately described but unfavorable. Misleading advertising concerns the content of the platform's promotional statements and whether a reasonable audience would draw a false impression from them. Deceptive earnings representation is a narrower category in which statements about what participants earn, or are likely to earn, are false or unrepresentative, and it is a recognized subject of consumer-protection enforcement; the United States Federal Trade Commission, for example, has authority over unfair or deceptive acts under its governing statute and maintains rules governing earnings claims in the marketing of business opportunities [6, 7]. Whether the Business Opportunity Rule reaches arrangements of the kind discussed here depends on whether a given program meets its definition of a business opportunity, and the essay does not resolve that question. Fraudulent inducement requires, in most jurisdictions, a material misrepresentation or concealment, knowledge of its falsity or reckless disregard for the truth, intent that the other party rely, and actual and reasonable reliance resulting in harm. Ordinary commercial risk is the baseline against which all of these are measured: people frequently invest effort in ventures that fail, and failure is not itself wrongful. Studies of aspirational labor bear on this baseline, since they show that participants often

understand the risk and nonetheless treat unpaid effort as an investment, which is a reason to hesitate before regarding every unsuccessful participant as deceived [5].

A platform's failure to compensate most participants is not sufficient by itself to establish fraud. If the platform accurately described the conditions, including the thresholds, the uncertainty, and the distribution of outcomes, then participants who nonetheless chose to invest accepted a risk, and the case for wrongdoing is weak even if the outcomes are poor. The consideration of information asymmetry in markets supplies the relevant lens: when one party knows much more than the other about the quality or distribution of what is being exchanged, exchange can fail or can proceed on terms that the informed party exploits [1]. The platform knows the distribution of creator earnings and the participant does not. A stronger case would therefore require evidence concerning what the platform represented, what it knew about actual earnings distributions, and whether material information was deliberately concealed or misrepresented. A platform that published only success stories while holding internal data showing that the typical participant earned nothing, and that encouraged participants to invest additional effort on the strength of those stories, would present a more serious case than one that published the data.

The evidentiary significance of several categories of material follows from this. Internal earnings data would show what the platform knew. Promotional success stories would show what it chose to represent and how typical those representations were relative to the data. Eligibility disclosures would show whether participants were told of thresholds and whether changes to those thresholds were communicated before effort was invested. Payout records would show what was actually paid, in what amounts, to what proportion of participants. Communications encouraging creators to invest additional effort, including automated prompts, would show whether the platform acted to increase participants' commitments at moments when it had reason to think returns were unlikely. None of this material is publicly available for the major platforms, and the essay therefore does not claim that the conditions for fraud are met. What it claims is narrower: that the structural features of the arrangement, namely information asymmetry, selective publicity, and conditional reward governed by discretionary rules, are the features that a regulator or court would examine, and that the governing question is whether the promise of economic opportunity is being used to recruit labor under materially misleading conditions.

A second observation is that the misleading character of an inducement does not depend on any single false statement. A pattern of accurate statements, selected and arranged to convey an impression that the full data would contradict, can mislead. The legal systems of different jurisdictions treat this issue differently, with some recognizing misleading half-truths as actionable and others requiring a more explicit falsehood. For present purposes it is enough to note that the analysis of selective disclosure in Section 6 and the formal appendix identifies a mechanism by which a platform could induce overentry without making any literally false statement, and that this mechanism is a plausible object of investigation whether or not it is ultimately held to

be unlawful.

Part V
From Finding Documents to Replacing Them

The Intermediation Triad

The preceding parts have described two of the features whose conjunction is the essay's central claim, and the fifth part adds the third. The first feature is a promise of economic opportunity that recruits effort, the subject of the second and fourth parts. The second is concentrated control over discovery that determines which effort is rewarded, the subject of the third part. The third is reuse of the accumulated record, in the form of answers and models, without corresponding changes in consent, compensation, or independent access, the subject of the fifth part. Any one of these is familiar and has defenders. The claim is that their conjunction is what distinguishes the present arrangement, because each feature amplifies the others: the promise recruits the effort that fills the archive, control over discovery determines what the archive contains and who is paid, and reuse extracts further value from the archive at the point where the first two have already fixed its composition. The conjunction is named here the intermediation triad so that it can be examined as a unit, and a reader who wishes to dispute the essay's conclusion can do so by showing that one of the three features is absent from a particular setting or that the features do not reinforce one another. The triad is proposed as a mechanism of cumulative institutional dependency and not as a definition of exploitation, and its features do not map one to one onto the three senses of exploitation distinguished in the introduction. The promise of opportunity can recruit effort even when it is accurate, so it is not the same as labor undertaken under distorted expectations, and control over discovery is a mechanism that may or may not produce an unequal division of surplus. The triad describes how the arrangement operates; whether any of the three senses of exploitation is present in a given case is a separate question. Cases in which only one or two features are present can still be analyzed, and may involve real wrongs of other kinds, but the essay's central diagnosis concerns the conjunction.

13 The Search Engine Becomes an Answer Engine

The transition from traditional search results to generative answer products can be dated by a short sequence of announcements. Google announced the conversational service Bard in February 2023 and renamed it Gemini in February 2024, and it introduced AI Overviews, generated summaries placed at the top of many search results pages, to broad deployment in the United States in 2024 [26, 15, 27]. The rollout drew public reports of errors and subsequent adjustments, which this essay does not assess, since its concern is with the information a summary preserves

and not with the correctness of any particular answer. A related but institutionally distinct development is source-grounded tools such as NotebookLM, which work primarily with sources that the user selects rather than inserting generated responses into public search results. The distinction matters, because a tool that operates on user-selected documents leaves the choice of documents with the user, whereas a summary placed at the head of a general search results page substitutes for the choice.

The relationship among search queries, retrieved documents, generated answers, and publisher traffic has changed accordingly. In a traditional results page, the query produced a list of documents, and the value of the search engine to the user was in directing the user to one or more of them. The engine's value to publishers was in the traffic it sent. In an answer interface, the query produces a text that may draw on documents without sending the user to them. The user's need may be satisfied at the page, and the document's role is that of an input rather than a destination. Observational evidence on the traffic consequence is beginning to accumulate. A study published by the Pew Research Center analyzed the browsing data of 900 adults in the United States for March 2025, covering 68,879 unique Google searches, and found that users who encountered an AI summary clicked on a traditional result link in 8 percent of visits, compared with 15 percent for visits without a summary, and that clicks on sources cited within the summary occurred in about 1 percent of visits [4]. The study is a snapshot of a particular month and a particular panel, it compares visits with and without a summary and not the same queries before and after deployment, so its figures are not a causal estimate. It measures behavior at the search page rather than the traffic or revenue of any publisher, but it is directly relevant to the question of whether generated answers reduce visits to the underlying documents.

The central question is whether a system designed to direct users toward independently published material is increasingly becoming a substitute for consulting that material. In the vocabulary of Section 3, an answer interface is an engagement-oriented surface, organized around the user's session on the page, that absorbs the contents of document-oriented sources and returns them without the persistence and addressability that made them documents. There is no reason to think the answer is uniformly yes. For some queries, such as the conversion of a unit or a reminder of a date, a generated answer is the appropriate end of the interaction and no document is needed. For other queries, such as the evaluation of a medical claim or the determination of what an author wrote, the document is the point and a summary is at best a pointer. The structural concern is that the same interface treats these cases alike. A user cannot easily tell from the form of an answer whether it is the kind of answer that suffices, and the incentives of the interface favor retention of the user on the page. The publisher's difficulty is that the economic arrangement under which documents were produced presupposed that referral was the compensation for being indexed; if the index is used to produce substitutes, the implicit exchange of content for traffic is altered without a negotiation.

14 The Loss of Documentary Specificity

The essay's epistemological argument concerns the properties of a document that are not exhausted by its semantic content. These include exact wording, authorship, publication history, citation context, argumentative structure, and its relationship to other documents. A summary may preserve what the document says about a topic while discarding all of these. Three distinctions allow the loss to be stated precisely.

Semantic equivalence concerns whether two expressions communicate substantially similar meanings. Two sentences that a competent reader would paraphrase into each other are semantically equivalent in this sense, and the relation is vague, graded, and context-dependent. Inferential equivalence concerns whether two expressions license the same conclusions under specified assumptions. It is stricter than semantic equivalence, because two passages may convey a similar gist while differing in a qualification that blocks an inference, such as the scope of a quantifier, the conditions of a theorem, or the population to which a finding applies. Consider the statements that every continuous real function on a closed bounded interval attains a maximum, and that every continuous real function attains a maximum. A quick summary might treat them as saying the same thing about continuous functions and maxima, but the second is false, since the identity function on the real line has no maximum, so the two do not license the same conclusions. Documentary identity concerns whether two items constitute the same recorded expression or the same historical artifact. It is the strictest relation and the one at issue in questions of quotation, priority, and attribution: a document is identical to another only if it is the same recorded text, and a perfect paraphrase is not.

These distinctions matter in literature, philosophy, law, mathematics, and scientific research, and each field has its own cases. In literature, the difference between a work and a summary of it is the difference between the object of criticism and a report on it. In philosophy, the precise formulation of an argument often carries its force, and a paraphrase that smooths a premise may transform a valid argument into an invalid one. In law, the wording of a statute or a contract is operative, and a reading that approximates it may not bind. In mathematics, a theorem's statement and proof are exactly specified objects, and a result that is described as similar to a theorem is not that theorem. In scientific research, the claim that a finding was reported in a particular paper depends on what the paper said, and a plausible reconstruction is not a citation.

A generated answer may preserve semantic relevance while losing inferential qualifications or documentary evidence. Even an accurate summary cannot substitute for every purpose served by access to the original text. The summary is a many-to-one image of the document, and information about which document, in which wording, produced the image is not recoverable from the image. Where the question asked of the system depends only on semantic content, this loss is harmless. Where it depends on inferential detail or documentary identity, the loss is material. The appendix states this claim as a proposition about sufficiency and shows that no semantic summary can in general answer exact-occurrence queries (Proposition 2); the proposition

concerns the summary alone, and the provenance annotations discussed after it are the additional structure that would be needed.

Negative searches, exact-phrase investigations, and the establishment of intellectual precedence deserve separate attention. A negative result in exact-match search, that is, the absence of a phrase from an index, supports an inference that a given expression was not published in the indexed record, where the index is defined and adequately searched, so that the inference concerns absence from that index and not from the historical record as a whole, and this inference has value in disputes about priority, plagiarism, and the novelty of terminology. A generated answer cannot supply a negative of this kind, because it is not a report on the index but a synthesis conditioned on the query, and its failure to mention a source is not evidence of the source's absence. An interface that replaces exact search with synthesis therefore removes a method of establishing precedence that independent researchers, who lack institutional archives, depend on. The capacity to show that a term or an argument was in circulation at a given time, with a link to the document, has historically been a means by which unaffiliated authors could defend the record of their work, and its loss shifts the burden of proof toward those with institutional standing.

15 The Reuse of Accumulated Human Contributions

The relationship between user-generated archives and contemporary generative systems raises the question of reuse. Video platforms and other services have accumulated enormous quantities of material contributed by users under historical arrangements that did not necessarily anticipate modern generative applications, and the terms under which contributors uploaded their work, which typically granted the platform a broad license for operating and promoting the service, were written for a different technical environment. Whether those terms permit the training of generative models, the construction of retrieval indexes, or the production of derivative summaries is a question that depends on the text of the terms and the law of the relevant jurisdiction, and it is the subject of active dispute.

The essay therefore distinguishes several processes that are often conflated. Model training uses a corpus to set the parameters of a statistical model, after which the corpus need not be consulted to generate outputs, though the model may reproduce portions of the corpus. Retrieval-augmented generation retrieves documents at the time of the query and conditions the output on them, so that the documents are used at the moment of answering. Indexing records features of documents to support search without producing new expression. Summarization produces text that conveys the content of a document in fewer words. Licensing is a contractual arrangement under which rights are granted, possibly with compensation. User-directed analysis occurs when a user supplies documents to a tool and asks questions about them. These processes have different technical profiles and different legal implications, and a discussion that treats them

as a single act of using data will mislead.

The central economic question is whether new forms of value can be extracted from accumulated contributions without corresponding changes in contributor compensation, consent, or control. Contributors who produced material for a revenue-sharing arrangement anticipated a particular flow of value, namely a share of advertising revenue attached to views of their pages. A reuse that generates value without views, such as the use of the pages to construct a model or to answer questions that formerly led to the pages, falls outside that flow. The platform may argue that the license covers the use, that the contributors received hosting and tools, and that the new value arises from investment in models rather than from the contributions. Contributors may respond that the aggregate of their work supplied the raw material, that they were not asked, and that the arrangement concentrates the gain. The question admits no answer without attention to the terms and to the counterfactual: what would the model have been without the corpus, and how should the value of an input that is indispensable in aggregate and negligible individually be divided.

A further consideration is the possibility that generated answers reduce visits to original publishers while drawing on the informational value of their work. The observational evidence cited in the preceding chapter is suggestive of a reduction in clicks but does not establish the effect on a given publisher, and the direction of causation for particular sites would require analysis of traffic data over time with controls for other changes. It should be considered as a hypothesis with some support and not as an established finding. The conceptual point holds regardless of magnitude: if the answer is built from documents and delivered in a way that removes the reader's reason to visit them, the producers of the documents bear a cost that the producer of the answer does not.

The countervailing benefits of improved accessibility, synthesis, translation, and research assistance are substantial. A generated summary can make dense material readable to people who would not otherwise engage with it, translation tools extend the reach of documents across languages, and synthesis across many sources can answer questions that no single document addresses. For people with disabilities, for non-native readers, and for those new to a field, these benefits are real. An assessment that ignored them would be one-sided. The relevant policy question is not whether such tools should exist but under what conditions they should draw on the record, and whether the benefits can be provided in a way that preserves the sources, credits the authors, and allows independent verification. Answer systems that cite documents with exact links, that quote only within clear limits, and that permit the reader to inspect the source retain many of the benefits while preserving documentary specificity, and the design space for such systems is wide.

Conclusion: The Right to Contribute Without Becoming Inventory

The original ambition of h2g2, Everything2, Squidoo, and the participatory web was that people with something to say could say it where others could find it. The historical problem was never that ordinary people lacked knowledge to contribute, nor that every contribution required payment. Voluntary publication, public-domain writing, open-source software, and collaborative encyclopedias show that a body of knowledge can be built largely from unpaid effort, provided the effort is not conscripted into a scheme that misdescribes what it is for.

The problem arises when institutions convert willingness to contribute into a renewable supply of commercially valuable labor while retaining control over discovery, compensation, and reuse. The conversion proceeds through steps that each have a plausible justification: a revenue share, a checklist, a classification, a dashboard, and a generative layer. Together they describe an arrangement in which the contributor bears risk and supplies inventory while the intermediary holds discretion. No part requires bad faith, which is why the arrangement is hard to correct: each actor can truthfully say that it did not intend the aggregate outcome.

Any such proposal must face the counterfactual. A web without revenue sharing, without ranking, and without generative synthesis would not be obviously better. Revenue sharing funded some publishing infrastructure and paid some contributors real money. Ranking, however opaque, made an enormous archive navigable, and the alternative of no ranking is not a neutral baseline but an unsearchable corpus. Generative synthesis offers accessibility and cross-source answers that no single document provides. The relevant comparison is therefore not between the present arrangement and an absence of intermediaries, but between the present arrangement and alternatives that retain these functions under different terms: disclosure of outcome distributions, contestable and explained classification, portability of accumulated work, and provenance-preserving answer interfaces. Whether such alternatives can be sustained financially is a further empirical question. Public funding, cooperative hosting, subscription models, and licensing arrangements have each been tried in some domain, and none is established as a general solution.

A better publishing environment would preserve durable addresses, independent hosting, portable records, transparent compensation terms, accessible original documents, and the ability to search without compulsory generative mediation. Durable addresses protect the possibility of citation and the establishment of precedence. Independent hosting reduces the dependence of publishers on any one intermediary. Portable records allow a contributor to leave without surrendering accumulated work. Transparent compensation terms, including disclosure of the distribution of outcomes and not only of exemplary successes, allow prospective contributors to estimate their expected return and to decide whether the nonmonetary benefits justify participation. Accessible original documents, delivered by exact links and not replaced by summaries, preserve the reader's ability to verify. And the ability to search without compulsory

generative mediation ensures that exact-match retrieval remains available as an instrument of research. None of these proposals requires the abolition of generative tools. They require that such tools be an option layered over a record that remains independently accessible, rather than a replacement that is imposed.

The vocabulary of a right is borrowed loosely from political theory, where an interests or capabilities framing might serve equally well, and it is used here for its clarity and not because it is uniquely apt. The right invoked in the title of this conclusion should be understood as a normative and political objective, a standard against which institutional arrangements and public policy can be assessed, and not as an existing legal entitlement or a claim about what any current law guarantees. It is a claim about what conditions a publishing environment ought to secure, and the proposals above are accordingly policy proposals. They differ in the instrument they suggest. Disclosure of outcome distributions and portability of accumulated work are suited to regulation or to contract terms; durable addresses and independent hosting to voluntary adoption and to cooperative or public provision; provenance-preserving answer interfaces to design standards that competition or copyright policy might encourage. The essay is diagnostic and does not argue that any single lever suffices. The conception of public knowledge worth defending is one in which the author is neither merely a content supplier nor an aspiring platform employee, and the reader is neither merely an advertising impression nor a passive recipient of generated answers. The possibility worth recovering is that people can contribute to a common intellectual record because they have something worth saying, and that others can encounter those contributions without every interaction being reorganized around monetization. The measure of a publishing environment is whether a person can write something true and useful, put it where it can be found, and leave it, without being required to turn the act into a career or to perform it indefinitely under rules that someone else may change.

A Formal and Empirical Development

This appendix collects the formal material that the historical chapters presuppose. It is deliberately separate from the narrative so that the argument can be read without it. The models are intentionally simple, and their purpose is to identify mechanisms and the empirical quantities on which the conclusions depend, not to estimate those quantities.

A.1 Platform Surplus Accounting

Model the platform as an intermediary receiving a set $I = \{1, \ldots, N\}$ of contributions from a population of authors. For a subset $A \subseteq I$, let $V(A) = V^{(0)}(A) + V^{(f)}(A)$ denote the platform's revenue attributable to an archive consisting of the contributions in A, where $V^{(0)}$ is the value realized through ordinary operation, such as advertising revenue from views, and $V^{(f)}$ is the value realized through later uses not contemplated at the time of contribution, such as the construction

of models or the sale of data access. Here $V^{(f)}$ is read as the imputed economic value of such later uses, meaning the value the platform would forgo if the corpus were removed or would have to pay for an equivalent corpus in a competitive market, and not as realized revenue; attributing it to a corpus requires a counterfactual model of what the use would have yielded without that corpus, which the essay does not supply. Let r_i be the payment to the author of contribution i and let C_P represent hosting, moderation, distribution, and other platform costs. The platform's aggregate contribution surplus is

$$S_P = V(I) - \sum_{i=1}^{N} r_i - C_P. \tag{2}$$

The formulation does not assign a separate value v_i to each contribution, because values cannot in general be assigned independently. If the worth of an archive arises from the combination of its documents, then V includes interaction effects, and in general $V(A \cup B) \neq V(A) + V(B)$ for disjoint A and B, so that $V(I) \neq \sum_i V(\{i\})$. The accounting identity also requires qualification of another kind. Not all of $V(I)$ is profit attributable to contributors, since part of it reflects investments by the platform that are independent of any contribution, and contributors are not thereby entitled to every dollar of platform revenue. Questions of attribution are reserved for a separate analysis. The marginal contribution of i to a sub-archive $A \not\ni i$ is $V(A \cup \{i\}) - V(A)$, and a rule such as the Shapley value,

$$\varphi_i(V) = \sum_{A \subseteq I \setminus \{i\}} \frac{|A|!\,(N - |A| - 1)!}{N!} \big[V(A \cup \{i\}) - V(A)\big], \tag{3}$$

averages these marginal contributions over orderings and distributes exactly $V(I) - V(\emptyset)$ among contributors [29]. The Shapley value is a natural benchmark because it is the unique rule that satisfies efficiency, symmetry, the dummy-player property, and additivity across games. The essay introduces the Shapley value not as a proposed rule of compensation but to show that the marginal contribution of a document depends on the coalition against which it is measured, so that an attribution convention is needed to aggregate it: a contribution may be indispensable in aggregate and negligible individually, which is the formal expression of the problem noted in Section 15, and any claim about what contributors are owed must specify an attribution rule.

More interesting than the definition are propositions about participation incentives. Three questions organize the analysis. Under what conditions can a platform increase the total quantity of contributed labor while decreasing the expected compensation per contributor? When does selective disclosure of successful earnings induce systematic overinvestment? And how does dependence on an external search intermediary affect the viability of the publishing platform? The first and second are addressed below, and the third is addressed in the discussion that follows them.

A.2 Heavy Tails, Selective Disclosure, and Entry

Let the payment R to a new entrant have distribution F on $[0, \infty)$ with finite mean $\mu = \mathbb{E}[R]$. Suppose that prospective contributors observe only those payments that exceed a publicity threshold τ, so that the examples they see are drawn from the conditional distribution of R given $R \geq \tau$. A naive prospective contributor who treats the observed examples as representative forms the belief

$$\hat{\mu}(\tau) = \mathbb{E}[R \mid R \geq \tau]. \tag{4}$$

Proposition 1. *For every τ with $\Pr(R \geq \tau) > 0$, $\hat{\mu}(\tau) \geq \mu$, and the inequality is strict whenever both $\Pr(R < \tau) > 0$ and $\Pr(R \geq \tau) > 0$. Consequently, a contributor with labor cost $wh + c$ and nonmonetary benefit b enters under naive beliefs whenever $\hat{\mu}(\tau) + b \geq wh + c$, and the set of contributor types who enter under naive beliefs contains the set who enter under correct beliefs ($\mu + b \geq wh + c$), strictly so for types with $\mu + b < wh + c \leq \hat{\mu}(\tau) + b$.*

Proof. Write $\mu = \Pr(R \geq \tau)\,\mathbb{E}[R \mid R \geq \tau] + \Pr(R < \tau)\,\mathbb{E}[R \mid R < \tau]$. Since every value of R in the event $\{R < \tau\}$ is smaller than every value in $\{R \geq \tau\}$, we have $\mathbb{E}[R \mid R < \tau] < \tau \leq \mathbb{E}[R \mid R \geq \tau]$ whenever both events have positive probability. Hence μ is a convex combination of a smaller and a larger quantity with positive weight on the smaller, and so $\mu < \hat{\mu}(\tau)$. If $\Pr(R < \tau) = 0$ then $\hat{\mu}(\tau) = \mu$. The entry statement is immediate from comparing the two inequalities. □

Remark 1. For a Pareto distribution with scale x_m and tail index $\alpha > 1$, the mean is $\alpha x_m/(\alpha - 1)$, the median is $x_m 2^{1/\alpha}$, and the conditional mean above any threshold $\tau \geq x_m$ is $\alpha\tau/(\alpha - 1)$. The ratio of the conditional mean to the unconditional mean is therefore τ/x_m, which does not depend on α: within this family the inflation of beliefs is governed by the publicity threshold relative to the scale, not by the tail index. The ratio of the conditional mean to the median is $\alpha\tau/((\alpha - 1)x_m 2^{1/\alpha})$, which grows without bound as τ rises. A prospective contributor who sees only outcomes above a high threshold can thus form a belief far above both the mean and the typical outcome. The proposition does not require any false statement by the platform. It requires only that what prospective contributors see is selected by performance. It also does not show that any platform induces such beliefs: the proposition establishes the consequences of the belief if it is held, and whether the belief is in fact induced is an empirical question.

The preceding proposition holds the distribution of payments fixed while entry varies. The dilution argument that follows is a different, endogenous mechanism, in which entry changes the distribution itself, and the two should not be read as a single derivation. The entry set is enlarged under naive beliefs, but total contributed labor is only one margin. A platform that increases the number of entrants N while leaving the total prize pool unchanged reduces the expected payment per entrant, and the same selective disclosure that induces entry is more effective the more the pool is diluted. The dilution argument can be stated without a specific functional form. If total contributor payments are $P = \sum_i r_i$ and the pool is divided among N

entrants, then the average payment is P/N, and an increase in N with P held fixed or growing more slowly than N reduces it. Meanwhile the platform's surplus S_P can rise with N through the growth of $V(I)$. A platform can therefore face an incentive to recruit beyond the point at which the average contributor's monetary return is positive, provided that selective disclosure sustains entry, and the aggregate labor supplied is then larger and its average reward smaller than under full disclosure. Whether any particular platform behaves in this way is an empirical matter. The model identifies the data required to test it: the full distribution of payments, the information shown to prospective entrants, and the dependence of N on that information.

A.3 Dependence on an Intermediary

A third question is how dependence on an external search intermediary affects a publishing platform's viability. Let the traffic of platform j be $T_j = \phi_j T$, where T is total search referral volume available to the class of platforms and ϕ_j is the share that the search engine's classification assigns to platform j. If revenue is proportional to traffic, then revenue is proportional to ϕ_j, and a reclassification that changes ϕ_j to ϕ'_j changes revenue by the ratio ϕ'_j/ϕ_j with no change in the platform's own conduct. The platform cannot observe the function that maps its pages to ϕ_j and can only estimate it from outcomes. Since contributor revenue shares are paid out of platform revenue, the change is transmitted to contributors, who bear it proportionally, and since the platform cannot diversify across search intermediaries when one dominates, the variance of ϕ_j is a risk that falls on the platform and on its contributors. The reported declines in modeled search visibility after Panda are changes of this kind, subject to the interpretive limits noted in Section 8. The calculation is a counterfactual sensitivity illustration and not an identification of the historical events: it assumes that total referral volume, monetization per visit, and the share of revenue attributable to search traffic remain constant, and that the ranking change is exogenous to the publisher's conduct, whereas observed visibility changes may also reflect characteristics of the affected publishers.

A.4 Documentary Sufficiency

The epistemic question can also be formalized. Let D denote a documentary corpus, q a query, $E(q) \subseteq D$ the retrieved evidence, and a a generated answer. The critical issue is identifying which classes of questions can be answered using a without losing information that remains available through $E(q)$. Let $\mathcal{D}$ be a class of corpora and let $\sigma : \mathcal{D} \to \mathcal{S}$ be a summary map that sends a corpus to its summary. A query Q is a function on corpora to some set of answers.

Definition 1. *A summary map σ is* sufficient *for a query Q on a class $\mathcal{D}$ if $\sigma(D) = \sigma(D')$ implies $Q(D) = Q(D')$ for all $D, D' \in \mathcal{D}$. Equivalently, there is a function f with $Q = f \circ \sigma$ on $\mathcal{D}$.*

Sufficiency is the formal statement that nothing the query depends on has been lost in the

summary. The terminology is borrowed from the classical notion of a sufficient statistic [8], which preserves the information in a sample relevant to estimating a parameter. The present definition differs in that it concerns the preservation of the answer to an arbitrary query on a corpus, and no probabilistic model or parameter is assumed. A semantic summary is one that identifies corpora up to a relation of meaning-preserving paraphrase. The following proposition states that such summaries cannot answer exact-occurrence queries.

Proposition 2. *Let s be a string and let $Q_s(D) = 1$ if s occurs verbatim in some document of D and $Q_s(D) = 0$ otherwise. Suppose $\mathcal{D}$ contains corpora D and D' such that s occurs in exactly one document d of D, and D' is obtained from D by replacing d with a document d' that does not contain s, for example because the occurrence of s has been replaced by a paraphrase. If $\sigma(D) = \sigma(D')$, then σ is not sufficient for Q_s on $\mathcal{D}$.*

Proof. By hypothesis $Q_s(D) = 1$. Every document of D' other than d' is a document of D other than d, and s occurs in none of those, while d' does not contain s. Hence $Q_s(D') = 0$. Since $\sigma(D) = \sigma(D')$ while $Q_s(D) \neq Q_s(D')$, the implication in the definition of sufficiency fails. □

Any summary that treats a passage and its paraphrase as the same cannot answer questions about which of them occurs. The same argument applies to questions of attribution, since the author of a passage is not recoverable from a semantic image of its content, and to questions of publication date. The proposition is elementary, and its significance is practical: it identifies a class of questions, namely those about the occurrence of exact strings, historical attribution, and textual identity, that cannot be answered reliably from semantic summaries without additional provenance and retrieval guarantees. These guarantees can be supplied by systems that retain exact-match indexes and return links to source documents, and the proposition thereby locates the requirement. It is a requirement on the retained structure and not on the quality of the generative component.

Remark 2. Proposition 2 concerns the summary map σ alone. Augmenting the summary changes the result. Let $\sigma^+(D) = (\sigma(D), \iota(D))$, where $\iota(D)$ is an exact-match index of D from which occurrences of any string can be read off. Then σ^+ is trivially sufficient for every Q_s, since $\iota(D)$ determines $Q_s(D)$. Sufficiency for a single query is cheap, since the one bit $Q_s(D)$ is sufficient for Q_s and retains nothing else, whereas sufficiency for the whole family $\{Q_s\}$ of exact-occurrence queries requires retaining enough to reconstruct the set of strings that occur, which is a far stronger requirement. A generated answer that merely links to a source document is an intermediate case: it is sufficient for a given Q_s only if the linked document set is complete for occurrences of s, which a link to one or a few documents does not establish, and a negative answer in particular requires exhaustive exact-match retrieval over D. The proposition therefore does not say that generated answers cannot support exact-occurrence claims. It says that the support must come from structure beyond the semantic summary.

Remark 3. The preceding is a statement about information preservation and should be distinguished from a claim about accuracy. A generative system may answer many exact-occurrence queries correctly by consulting an auxiliary index, and a system that cannot consult one may nonetheless guess correctly. The proposition says that correctness cannot be guaranteed by the summary alone, and that where the evidential value of the answer depends on its being grounded in the record, the grounding must be supplied by something other than the semantic content. A negative answer is subject to a further requirement. To conclude from the absence of a string from retrieved evidence that the string does not occur in D requires that the retrieval be exhaustive over D for exact matches, a property that semantic retrieval does not have.

The proposition concerns whether an answer can be determined by a summary. A different question is whether the answer can be traced to the records that support it. Research on database provenance distinguishes knowing a query result from knowing which source records it depends on: why-provenance identifies the source data that influenced the existence of an output, and where-provenance identifies the locations in the source from which a value was taken [2]. A later algebraic treatment shows that several such notions, together with bag semantics and incomplete and probabilistic databases, arise as instances of a single construction in which tuples are annotated with elements of a commutative semiring [14]. In the present setting the point is that an answer accompanied by annotations identifying its supporting documents carries information that the answer alone does not, and that the loss described by Proposition 2 is partly a loss of provenance. A generative interface that retained document-level provenance for each claim and exposed it to the reader would restore part of what a bare summary discards, though it would not by itself guarantee that the annotated documents support the claim as stated.

Answer recoverability and evidentiary provenance are separate properties, and neither implies the other. Two notions of support must be distinguished, corresponding to open-world and closed-world reasoning. Say that a set S of documents *witnesses* the positive answer $Q_s = 1$ if s occurs in some document of S; the witness then supports the answer in every corpus containing S. A negative answer cannot be supported in this way against arbitrary additions to an open corpus, so support for it is relative to a stated corpus: say that a set S is a *completeness certificate* for $Q_s = 0$ relative to a fixed, enumerated corpus D if $S = D$ and every document in S has been examined by an exact-match procedure that is complete for occurrences of s.

Proposition 3. *Let Q_s be as in Proposition 2. (i) A summary can be sufficient for Q_s without identifying any witness: the map $\sigma_0(D) = Q_s(D)$ is sufficient for Q_s, yet when $Q_s(D) = 1$ it does not determine which document contains s. (ii) In the open world, if the class $\mathcal{D}$ is closed under adding documents and some document containing s exists, then for no corpus D with $Q_s(D) = 0$ does any $S \subseteq D$, including $S = D$, fix the answer in every corpus containing S unchanged. (iii) In the closed world, relative to a fixed corpus D with $Q_s(D) = 0$, a set $S \subsetneq D$ of examined documents does not establish the answer, since replacing a document of $D \setminus S$ by one containing s yields a corpus that agrees with D on S and has $Q_s = 1$. Only examination of all of D does.*

Proof. (i) Sufficiency of σ_0 is immediate, since $Q_s = \mathrm{id} \circ \sigma_0$. Two corpora of equal size in which the single occurrence of s lies in different documents both have $\sigma_0 = 1$, and neither value reveals the location. (ii) Let d^* be a document containing s with $d^* \notin D$. The corpus $D \cup \{d^*\}$ contains S unchanged and has $Q_s = 1$, while D contains S unchanged and has $Q_s = 0$. (iii) Take $d \in D \setminus S$ and let D' be D with d replaced by a document containing s, assuming as in (ii) that such a document exists. Then D' and D agree on every document of S, and $Q_s(D') = 1 \neq Q_s(D)$, so an examination confined to S cannot distinguish them. □

The proposition has a practical reading. A system may return the correct answer while retaining no record of where it came from, and a system may return links to documents that are genuinely relevant while the links do not suffice to establish what is claimed. Documentary provenance has accordingly two forms. For a positive exact-occurrence claim a single witnessing document is a complete support, which is why a citation to the passage is strong evidence. For a negative claim there is no local witness: the support is a completeness guarantee over a stated corpus and a stated search procedure, which is why a handful of cited documents cannot establish absence, and why absence claims should always be stated relative to a defined and adequately searched index and not relative to the historical record. Whether a claim follows from the documents cited for it is a further matter again, since it depends on the content of those documents and not only on their identity.

This connects the essay's historical argument to a more general theory of information-preserving and information-losing intermediaries. A search engine in the traditional form is also an intermediary that filters, ranks, excerpts, and sometimes rewrites queries, and it is not lossless. The relevant distinction is whether the interface preserves access to independently inspectable source documents. Traditional results pages return pointers to documents and leave the documents intact for the reader to examine. An answer interface returns a function of the corpus, and unless it retains and exposes document-level provenance it conceals the preimage. Each is useful for some queries and inadequate for others, and the policy concern is with the replacement of the first by the second in contexts where the queries of interest are of the kind that compression cannot answer.

A.5 Empirical Requirements

The foregoing propositions identify what must be measured. For the platform-economics claims, the required data are the full distribution of contributor payments by platform and year, the information displayed to prospective contributors, the retention of contributors over time, and the thresholds and their changes. For the search-governance claims, the required data are traffic and revenue time series for matched publishers around ranking updates, together with a method for distinguishing the effect of the update from contemporaneous changes in the advertising market. For the answer-engine claims, the required data are click-through and referral series for publishers before and after the deployment of generated summaries, stratified by query type,

together with the terms governing the reuse of contributed material. Public sources currently supply only fragments of these: reported visibility indices for a cohort of sites, published eligibility thresholds, and a single-panel behavioral study. A carefully established finding about how many people contribute substantial labor for negligible compensation, and how that population is distributed across platforms, would be more consequential than any unsupported assertion about the proportion of contributors who earn nothing. The scale of the population is itself an empirical question, and the discipline of the research program is to measure it before characterizing it.

References

[1] Akerlof, George A. 1970. "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism." *The Quarterly Journal of Economics* 84(3): 488–500.

[2] Buneman, Peter, Sanjeev Khanna, and Wang-Chiew Tan. 2001. "Why and Where: A Characterization of Data Provenance." In *Database Theory: ICDT 2001*, Lecture Notes in Computer Science 1973, 316–330. Berlin: Springer.

[3] Campbell, Donald T. 1979. "Assessing the Impact of Planned Social Change." *Evaluation and Program Planning* 2(1): 67–90.

[4] Chapekis, Athena, and Anna Lieb. 2025. "Google Users Are Less Likely to Click on Links When an AI Summary Appears in the Results." Pew Research Center, July 22. `https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/`

[5] Duffy, Brooke Erin. 2017. *(Not) Getting Paid to Do What You Love: Gender, Social Media, and Aspirational Work*. New Haven: Yale University Press.

[6] Federal Trade Commission Act, Section 5. 15 U.S.C. §45.

[7] Federal Trade Commission. Business Opportunity Rule. 16 C.F.R. Part 437.

[8] Fisher, Ronald A. 1922. "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society of London, Series A* 222: 309–368.

[9] Fried, Jason. 2006. "Transparency's Dirty Little Secret." *Signal v. Noise*, February 7. `https://signalvnoise.com/archives2/transparencys_dirty_little_secret`

[10] Gillespie, Tarleton. 2010. "The Politics of 'Platforms'." *New Media & Society* 12(3): 347–364.

[11] Gillespie, Tarleton. 2014. "The Relevance of Algorithms." In *Media Technologies: Essays on Communication, Materiality, and Society*, edited by Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot, 167–194. Cambridge, MA: MIT Press.

[12] Godin, Seth. 2005. *Everyone's an Expert (On Something)*. Squidoo LLC. Licensed under Creative Commons Attribution 2.5. https://seths.blog/wp-content/uploads/2005/10/everyoneisanexpert2.pdf

[13] Goodhart, Charles A. E. 1975. "Problems of Monetary Management: The U.K. Experience." In *Papers in Monetary Economics*, Volume I. Sydney: Reserve Bank of Australia.

[14] Green, Todd J., Grigoris Karvounarakis, and Val Tannen. 2007. "Provenance Semirings." In *Proceedings of the 26th ACM Symposium on Principles of Database Systems*, 31–40. New York: ACM.

[15] Hsiao, Sissie. 2024. "Bard Becomes Gemini: Try Ultra 1.0 and a New Mobile App Today." *Google Blog*, February 8. https://blog.google/products-and-platforms/products/gemini/bard-gemini-advanced-app/.

[16] "HubPages." *Wikipedia, The Free Encyclopedia*. Accessed November 2026. Used for HubPages' founding year and later sale.

[17] Kellogg, Katherine C., Melissa A. Valentine, and Angèle Christin. 2020. "Algorithms at Work: The New Contested Terrain of Control." *Academy of Management Annals* 14(1): 366–410.

[18] Mangel, Marc, and Francisco J. Samaniego. 1984. "Abraham Wald's Work on Aircraft Survivability." *Journal of the American Statistical Association* 79(386): 259–267.

[19] McGee, Matt. 2013. "Google Panda Two Years Later: Losers Still Losing and One Real Recovery." *Search Engine Land*, February. https://searchengineland.com/google-panda-two-years-later-losers-still-losing-one-real-recovery-149491

[20] McGee, Matt. 2013. "Google Panda Two Years Later: The Real Impact." *Search Engine Land*, February. https://searchengineland.com/google-panda-two-years-later-real-impact-149519

[21] MediaPost. 2020. "YouTube Puts Ads on Non-Partner Videos, Sans Payment to Creators." November. Trade coverage of a terms-of-service update; no byline displayed in the material consulted. https://www.mediapost.com/publications/article/357996/youtube-puts-ads-on-non-partner-videos-sans-payme.html.

[22] Mediaweek. 2026. "YouTube Doubles the Hurdle for Creators Chasing Ad Revenue." August. https://mediaweek.com.au/social-media/youtube-doubles-the-hurdle-for-creators-chasing-ad-revenue.

[23] Nelson, Matthew G. 2007. "Downgraded by Google for Search Spam, Squidoo Cleans House." *ClickZ*, July 13. https://clickz.com/downgraded-by-google-for-search-spam-squidoo-cleans-house/64247/.

[24] Nieborg, David B., and Thomas Poell. 2018. "The Platformization of Cultural Production: Theorizing the Contingent Cultural Commodity." *New Media & Society* 20(11): 4275–4292.

[25] Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

[26] Pichai, Sundar. 2023. "An Important Next Step on Our AI Journey." *Google Blog*, February 6. https://blog.google/intl/en-mena/company-news/technology/our-ai-journey/.

[27] Reid, Elizabeth. 2024. "Generative AI in Search: Let Google Do the Searching for You." *Google Blog*, May 14. https://blog.google/products-and-platforms/products/search/generative-ai-google-search-may-2024/.

[28] Rosen, Sherwin. 1981. "The Economics of Superstars." *The American Economic Review* 71(5): 845–858.

[29] Shapley, Lloyd S. 1953. "A Value for n-Person Games." In *Contributions to the Theory of Games*, Volume II, edited by Harold W. Kuhn and Albert W. Tucker, 307–317. Princeton: Princeton University Press.

[30] "Squidoo." *Wikipedia, The Free Encyclopedia*. Accessed November 2026. Used for the January 2006 profit-sharing announcement and launch details; the acquisition is cited to Zeckman (2014).

[31] Srnicek, Nick. 2016. *Platform Capitalism*. Cambridge: Polity.

[32] Strathern, Marilyn. 1997. "'Improving Ratings': Audit in the British University System." *European Review* 5(3): 305–321.

[33] Team YouTube. 2026. "New Opportunities to Earn and Changes to the YouTube Partner Program." Announced August 10, 2026; text consulted on the Google Blog mirror dated August 11, 2026. https://blog.google/intl/en-mena/product-updates/connect-communicate/new-opportunities-to-earn-and-changes-to-the-youtube-partner-program/

[34] Terranova, Tiziana. 2000. "Free Labor: Producing Culture for the Digital Economy." *Social Text* 18(2): 33–58.

[35] Tucker, Elizabeth. 2024. "New Ways We're Tackling Spammy, Low-Quality Content on Search." Google Blog, March 5. https://blog.google/products/search/google-search-update-march-2024/

[36] van Dijck, José. 2013. *The Culture of Connectivity: A Critical History of Social Media*. New York: Oxford University Press.

[37] YouTube Help. "YouTube Partner Program Overview and Eligibility." Google. Accessed November 2026.

[38] Zeckman, Ashley. 2014. "Seth Godin's Squidoo Acquired by HubPages." *Search Engine Watch*, August 19.