

Are We Data? (Do We Matter?)

PARTICIPATION THROUGH THE EYES OF OTHER ACTORS

Brian Ballsun-Stanton

INTS1302 Navigating Digital Worlds, Macquarie University

2026-08-18

We Want Information

COERCION AND THE DEMAND FOR INFORMATION

- Contemporary platforms obtain information through ordinary participation.



Displayed clip: [The 3:02 opening titles of *The Prisoner* \(1967\)](#)
([*The Prisoner* \(1967\), 2024](#))

Displayed clip: [The 3:02 opening titles of *The Prisoner* \(1967\)](#)

- Number Six’s resignation, abduction and confinement
 - a person reduced to an assigned number
 - information demanded but not willingly supplied

- the Village’s overt coercion
 - “By hook or by crook, we will” ([The Prisoner \(1967\), 2024](#))
 - the Village promising to obtain information despite Number Six’s refusal

Contemporary platforms obtain information through ordinary participation.

- much nicer interfaces than the Village

- the demand for information continuing without overt confinement

People Construct Social Spaces

NETWORKED PUBLICS AS TECHNOLOGY AND SHARED PRACTICE

Publics provide a space and a community for people to gather, connect, and help construct society as we understand it.

([boyd, 2014, p. 9](#))

- Physical publics include parks, malls and sporting events.
- Hybrid publics combine co-present activity with networked communication.
- Online publics range from small forums to transnational movements.
- A networked public is both a technological space and an imagined community.

Physical publics include parks, malls and sporting events.

- accessible places where people gather

- communities people imagine themselves to belong to
 - overlapping memberships
 - tangled boundaries between publics

Hybrid publics combine co-present activity with networked communication.

- boyd's Nashville football game ([boyd, 2014, pp. 3–4](#))
 - teens using phones briefly to find one another and share what they saw
 - parents paying more sustained attention to their own devices
- Facebook extending the event after the stadium emptied
 - photos, tags and comments carrying the game into a networked public
 - online participation extending local relationships rather than replacing them

Online publics range from small forums to transnational movements.

- different scales of gathering

- different forms of mediation
 - neither scale nor technology sufficient by itself
 - shared practice still required

A networked public is both a technological space and an imagined community.

- a mediated place to gather and connect

- a social construct made through people and practice
 - technology shaping possibilities
 - participants supplying norms and meaning

Epigraph: “Publics provide a space and a community for people to gather, connect, and help construct society as we understand it.” ([boyd, 2014, p. 9](#))

- publics as spaces where people gather and connect

- publics as communities that participants imagine and construct

Four Affordances Shape Participation

WHAT CAN BE REMEMBERED, SEEN, SHARED, AND FOUND

- Affordances connect a space's design with people's practices and expectations.
- **Persistence** concerns how long content can remain available.
- **Visibility** concerns who can encounter the content.
- **Spreadability** concerns how easily content can be copied and shared.
- **Searchability** concerns how easily content can be found.
- Affordances shape possibilities without determining behaviour.

Affordances connect a space’s design with people’s practices and expectations.

- the chair and table from last week
- possibilities understood through users and settings
 - technical capacity
 - social expectation
- boyd’s four labels distinguishing time, audience, circulation and discovery
 - persistence and visibility
 - “the durability of online expressions and content”
 - “the potential audience who can bear witness” ([boyd, 2014, p. 11](#))
 - spreadability and searchability
 - “the ease with which content can be shared”
 - “the ability to find content” ([boyd, 2014, p. 11](#))

Persistence concerns how long content can remain available.

- conversation becoming durable record

- later access beyond the original moment

Visibility concerns who can encounter the content.

- technically available audiences

- actually imagined audiences

Spreadability concerns how easily content can be copied and shared.

- movement into another setting

- context failing to travel with the contribution

Searchability concerns how easily content can be found.

- persistent but practically invisible material

- discovery changing the effective audience

Affordances shape possibilities without determining behaviour.

- participant strategies and community norms

- possibility without automatic acceptability

- boyd’s anti-determinist boundary
 - the window analogy
 - “The window’s affordances don’t predict how people will communicate, but they do shape the situation nonetheless.” ([boyd, 2014, p. 11](#))
 - an affordance shaping a situation without fixing one behaviour
 - practice remaining contingent
 - the same technical possibility supporting different practices
 - community norms affecting which possibility is taken up

Kidlink as a Networked Public

FROM A SHORT EXPERIMENT TO A GLOBAL DIALOGUE

- Kidlink began as a fourteen-day online project connecting 260 children in three countries ([Mizell et al., 1992, p. 5](#)).
- The next annual project connected about 2,600 children in 31 countries ([Mizell et al., 1992, p. 6](#)).
- Successive annual projects turned a temporary event into an ongoing global dialogue.
- Separate forums supported conversation, self-description and collective action.
- Internet Relay Chat provided live text conversation in named channels.
- In 1995, children used Kidlink to speak with Zlata Filipović, a Bosnian teenager known for her wartime diary.

From the [raw Kidlink IRC log](#):

[Zlata] To everyone: is Bosnia still on TV? Are people still talking about Sarajevo?

[JodyGee] Zlata: there is a lot of talk about Bosnia lately

[A-Kim] Zlata: Sarajevo was on the front page a few days ago in Toledo, Ohio, USA

([KIDLINK, 1995](#))

Speaker notes

Kidlink began as a fourteen-day online project connecting 260 children in three countries ([Mizell et al., 1992, p. 5](#)).

- short duration

- small international public

The next annual project connected about 2,600 children in 31 countries ([Mizell et al., 1992, p. 6](#)).

- tenfold increase from the first experiment

- a wider public spanning 31 countries

Successive annual projects turned a temporary event into an ongoing global dialogue.

- continuity beyond one experiment

- shared expectations and institutional history

Separate forums supported conversation, self-description and collective action.

- children’s conversation and self-description spaces

- adult planning spaces

- participation rules changing by forum

Internet Relay Chat provided live text conversation in named channels.

- live exchange under nicknames as one shared social situation

- scrolling transcript retained as a later record

In 1995, children used Kidlink to speak with Zlata Filipović, a Bosnian teenager known for her wartime diary.

- questions and messages from children across countries

- a moderated encounter in which Filipović also questioned the participants
 - questions sorted through the separate #Z-Mod channel
 - “Z-mod is another channel where we are sorting out your questions”
 - the moderation channel named directly in the live exchange ([KIDLINK, 1995](#))
 - adult facilitators managing turns in the main channel
 - “A-Barabara is next for a question”
 - the facilitator prompt and the subsequent exchange surviving together ([KIDLINK, 1995](#))

Displayed extract: From the [raw Kidlink IRC log](#) ([KIDLINK, 1995](#)).

[Zlata] *To everyone: is Bosnia still on TV? Are people still talking about Sarajevo?*

[JodyGee] *Zlata: there is a lot of talk about Bosnia lately*

[A-Kim] *Zlata: Sarajevo was on the front page a few days ago in Toledo, Ohio, USA*

- the four-hour session as recorded on 2 June 1995

- the later [edited presentation](#) as a different representation ([de Haas, 1995](#))
 - selected and rearranged dialogue
 - added narrative and moderator explanation

Kidlink Changed Its Affordances

DURATION, ACCESS, DISTRIBUTION, AND SEARCH

- **Persistence:** stored messages and annual projects outlasted individual exchanges.
- **Visibility:** child and adult forums created different audiences.
- **Spreadability:** mailing lists, newsletters and printouts moved contributions into other settings.
- **Searchability:** children's introductions became a searchable database.
- Kidlink's organisers closed a peace forum while the technology still worked.
- Changing an affordance changed what the public could become.

Persistence: stored messages and annual projects outlasted individual exchanges.

- continuity for the public
 - messages available after the moment of sending
 - one annual project continuing into the next
 - electronic letters could be saved and collected “today, tomorrow or next month” ([Mizell et al., 1992, p. 7](#))
 - “Then, KIDS-93 will start. The structure and objectives are the same.” ([Mizell et al., 1992, p. 12](#))
- a retained record for Kidlink
 - children’s responses stored beyond the original exchange
 - accumulated material available for later retrieval
 - “The responses could be stored on a hard disk somewhere”
 - “anybody can search the messages to find information of interest” ([Mizell et al., 1992, p. 7](#))

Visibility: child and adult forums created different audiences.

- separated participation and planning
 - four discussion forums intended for children aged 10–15
 - KIDS-91 and KIDPLAN used for adult information and planning
 - KIDPLAN described as “The planning room”
 - its discussion ranging from network problems to policy decisions ([Mizell et al., 1992, p. 10](#))
- access rules marking the boundary
 - adults excluded from children’s discussion
 - adult forums described as open to everyone
 - “Adults were not allowed to join the discussion.”
 - “The two forums for adults” had “free access for everybody” ([Mizell et al., 1992, pp. 9–10](#))

Spreadability: mailing lists, newsletters and printouts moved contributions into other settings.

- circulation across networks
 - project discussion and news sent to mailing-list subscribers
 - electronic distribution reaching many recipients quickly
 - “All the discussion and news about the project were distributed electronically”
 - mailing lists allowing messages to be “distributed to many people, much quicker” ([Mizell et al., 1992, p. 8](#))
- redistribution through local settings
 - newsletters received electronically in participating countries
 - paper copies distributed beyond the network connection
 - participants could “print them out for local distribution on paper”
 - a networked contribution therefore reaching people through a physical copy ([Mizell et al., 1992, p. 7](#))

Searchability: children’s introductions became a searchable database.

- contributions available by category
 - responses organised around four recurring questions
 - RESPONSE retaining those answers as a collection
 - the questions asked who participants were, what they wanted to become, and what future they wanted for the world
 - the same structure made answers comparable and retrievable ([Mizell et al., 1992, pp. 8–9](#))
- educators and researchers as later audiences
 - a participant’s self-description becoming information for someone else
 - a social contribution gaining an institutional research use
 - RESPONSE had become “a searchable data base filled with information for educators, researchers, and other interested persons”
 - the report presents this later use as a benefit, not as the children’s stated purpose ([Mizell et al., 1992, p. 9](#))

Kidlink’s organisers closed a peace forum while the technology still worked.

- closure of KIDPEACE
 - an existing discussion forum about war and peace
 - closure imposed by the project’s organisers
 - “KIDPEACE was later closed.”
 - the report gives an organisational reason rather than a technical failure ([Mizell et al., 1992, p. 9](#))
- organisers refusing identification with peace activism
 - the public meaning of the forum conflicting with Kidlink’s desired institutional identity
 - technical capacity remaining while institutional permission disappeared
 - “we did not want people to think that KIDS-91 was a peace activist movement”
 - the same infrastructure continued to support other child and adult forums ([Mizell et al., 1992, pp. 9–10](#))

Changing an affordance changed what the public could become.

- different people able to participate or observe
 - age and role affecting access to particular forums
 - later readers encountering records without joining the original exchange
- stored contributions available for new purposes
 - children’s answers sustaining later Kidlink activity
 - the same answers serving education and research
- organisers retaining control over access
 - software possibilities bounded by membership and forum rules
 - KIDPEACE showing that an available forum could still be withdrawn
- the four affordances changing together rather than independently
 - persistence making later search and redistribution possible
 - visibility rules deciding who could use those possibilities

Why This Matters

FROM TECHNICAL FEATURE TO SOCIAL CONSEQUENCE

- Persistence helped Kidlink outlast its first experiment.
- Access rules shaped who could speak, observe and organise.
- Distribution and search brought children's contributions to later audiences.
- A feature matters when it changes participation and its consequences.

Persistence helped Kidlink outlast its first experiment.

- successive annual projects creating continuity
 - reported growth across consecutive projects
 - “260 children from three countries participated in the dialog”
 - “some 2600 children from 31 countries participated in the Kids-91 project” ([Mizell et al., 1992, pp. 5–6](#))
 - change from a fourteen-day experiment to a continuing network
 - the first project bounded by one event and three countries
 - the following annual project spanning 31 countries, with a third already under way
- stored messages giving the public a history
 - discussion extending beyond any one live exchange
 - accumulated responses available to later participants
 - “a continuous global dialog between them”
 - the report’s description of a change from collecting responses to sustaining dialogue ([Mizell et al., 1992, p. 8](#))

Access rules shaped who could speak, observe and organise.

- children and adults assigned different forums
 - child discussion reserved for participants aged 10–15
 - adult planning conducted in separately accessible forums
 - “Adults were not allowed to join the discussion”
 - different access rules for different roles ([Mizell et al., 1992, p. 10](#))
- organisers able to close a forum while the technology remained
 - KIDPEACE closed by Kidlink’s organisers
 - institutional identity limiting an available technical possibility
 - “we did not want people to think that KIDS-91 was a peace activist movement”
 - closure explained as an organisational choice rather than a technical failure ([Mizell et al., 1992, p. 9](#))

Distribution and search brought children’s contributions to later audiences.

- mailing lists and printouts reaching other settings
 - project news distributed electronically to subscribers
 - “All the discussion and news about the project were distributed electronically”
 - mailing lists moving the exchange beyond the originating forum ([Mizell et al., 1992, p. 8](#))
 - newsletters printed for local distribution
 - participants could “print them out for local distribution on paper”
 - circulation extending through local settings that were not themselves connected to the live forum ([Mizell et al., 1992, p. 7](#))

- searchable records serving educators and researchers
 - RESPONSE retaining children’s answers by category
 - later users approaching the record for purposes other than participation
 - “a searchable data base filled with information for educators, researchers, and other interested persons”
 - participants recast as sources for a later audience ([Mizell et al., 1992, p. 9](#))

A feature matters when it changes participation and its consequences.

- what people can do and on what terms
 - live conversation, stored response and searchable record as different possibilities
 - access rules and organiser choices shaping which possibility exists

- who gains access or control over the resulting record
 - participants gaining connection in the original public
 - organisers and later users gaining durable control or access
 - technical capacity distributing possibilities unevenly
 - social and institutional rules determining who may use them

One Act Becomes Many Things

RELATIONSHIP, RECORD, ARCHIVE, PROFILE, CORPUS

Every single person will be closely followed all his life by his own data shadow where everything he has ever done, learnt, bought, achieved, or failed to achieve is expressed in binary numbers for eternity. His shadow will often be taken for himself and we will have to pay for all the inaccuracies as if they were his own; he will never know who is looking at [his] shadow or why.

Kerstin Anér, quoted in ([Bellovin, 2021](#))

- A social act can remain meaningful participation while also becoming usable data.
- A platform can use recorded activity to define audiences or license data access.
- An archive can preserve the record after the original event.
- Researchers, advertisers and model developers can use traces differently.
- A data shadow preserves some features and leaves others out.

Epigraph: “Every single person will be closely followed all his life by his own data shadow where everything he has ever done, learnt, bought, achieved, or failed to achieve is expressed in binary numbers for eternity. His shadow will often be taken for himself and we will have to pay for all the inaccuracies as if they were his own; he will never know who is looking at [his] shadow or why.” (Kerstin Anér, quoted in [Bellovin, 2021](#))

- the reproduced source prints “looking at is shadow”, with the bracketed correction marking the apparent typo

- a 1970s warning about institutional records preceding the current cases

A social act can remain meaningful participation while also becoming usable data.

- meaning for the participant
 - a question, answer or joke inside an ongoing relationship
 - shared context making the contribution intelligible to that public
- utility for a later actor
 - the same words treated as evidence, a signal or an example
 - usefulness determined by the later actor’s purpose

A platform can use recorded activity to define audiences or license data access.

- participation classified for advertising
 - recent subscriptions and engagement used to define an audience
 - the resulting advertisements shown beyond the community that supplied the signal
 - Reddit says people who have “joined or recently engaged with” selected communities may be shown the advertisement ([Reddit, n.d.](#))
 - a platform claim about its current audience product, not a measure of how accurately it represents participants
- accumulated conversation licensed for other products
 - current structured access offered to a model developer
 - platform control over the terms and format of access
 - Reddit promises access to “authentic and constantly-updated human-generated conversations” ([Reddit, 2024](#))
 - “authentic” is the seller’s description of its product

An archive can preserve the record after the original event.

- persistence beyond the live public
 - Archive.org retaining Kidlink pages after the site changed
 - captured pages acquiring a much longer life than the exchange

- later access without the original situation
 - words remaining while tone, relationships and surrounding activity fall away
 - a later reader reconstructing context from an incomplete capture

Researchers, advertisers and model developers can use traces differently.

- the trace selected for each actor’s purpose
 - archived text as evidence for a research question
 - the Kidlink method uses “categories constructed from our initial research question” ([Johnston et al., 2026](#), sec. “Introduction”)
 - the categories make some archived records visible to researchers
 - participation as an advertiser’s inferred interest or audience category
 - recent community engagement treated as evidence of likely interest ([Reddit, n.d.](#))
 - an advertiser buying access to a classified audience rather than to a participant’s own account of the relationship
 - selected text as one example in a model developer’s corpus
 - Reddit markets licensed access as providing “more efficient ways to train models” ([Reddit, 2024](#))
 - the sales claim does not establish which contributions enter a particular corpus

- each use retaining different features
 - provenance and source context for research
 - account, topic and engagement for advertising
 - patterns across the collection for model development

A data shadow preserves some features and leaves others out.

- a partial representation adequate for another actor’s purpose
 - selected fields standing in for a person or activity
 - Anér’s warning about a useful representation mistaken for the person
 - the warning reproduced from *New Scientist* in Bellovin’s source-history post
 - the early coinage attributed to Anér without claiming that the exact first use has been established ([Bellovin, 2021](#))

- the person and situation exceeding the record
 - intention, relationship and lived context remaining outside the fields
 - decisions still made as if the partial record were the person

Persistence Without Searchability

URL-BASED ACCESS AND MODEL-ASSISTED DISCOVERY

- Archive.org kept Kidlink pages long after the live site changed.
- The pages survived, but the archive did not make their text searchable.
- Nearly 300,000 Kidlink text records are too many to browse.
- Later research made different parts of the archive visible.
- A model created another partial way into the archive.
- This was AI-assisted classification, not chatting or model training.

Archive.org kept Kidlink pages long after the live site changed.

- enormous temporal snapshot
 - captures accumulated across changing versions of the site
 - one domain producing 384,700 archived sources
 - Johnston and colleagues report 384,700 sources collected from the Kidlink domain
 - 293,679 of those sources form the textual subset used in their account ([Johnston et al., 2026](#), sec. “Corpus”)
- broken or missing routes through the old site
 - individual pages surviving without a complete original structure
 - navigation depending on the URLs a researcher already knows

The pages survived, but the archive did not make their text searchable.

- captured pages retaining content without a complete discovery interface
 - survival of a page separated from the ability to locate it
 - no claim that every original Kidlink search tool disappeared
- the Wayback Machine searching by address rather than page text
 - known URL as the ordinary entry point
 - no full-text search across the archived pages
 - “only supports search by URL, not the text on the actual pages”
 - “looking for a needle in a haystack, without being able to grasp the size of the haystack” ([Johnston et al., 2026](#), sec. “Introduction”)

Nearly 300,000 Kidlink text records are too many to browse.

- the size of the preserved collection
 - 293,679 textual records
 - 384,700 captured sources overall
 - corpus totals reported by Johnston and colleagues
 - sources including more than the textual subset ([Johnston et al., 2026](#), sec. “Corpus”)
- persistence without practical navigability
 - scale defeating ordinary browsing
 - manual reading unable to provide systematic coverage
 - “It is not feasible to read individual pages manually”
 - some form of distant or targeted reading becoming necessary ([Johnston et al., 2026](#), sec. “Archival source discovery in the age of abundance”)

Later research made different parts of the archive visible.

- age claims, informal language and topic shifts used as search categories
 - categories derived from the historians’ research question
 - each document tested against a declared area of interest
- each category bringing some records forward and leaving others obscure
 - discovery shaped by the wording and boundaries of a category
 - unselected documents remaining persistent but unseen

A model created another partial way into the archive.

- five categories
 - five research-relevant routes combined for this investigation
 - category judgements recorded for individual documents

- 428 candidates for human investigation
 - the reported result of combining the selected categories
 - “428 sources were extracted”
 - sources surfaced that traditional navigation had not located ([Johnston et al., 2026](#), sec. “Results of last iteration”)
 - selective searchability shaped by the research question
 - candidate status depending on the researchers’ categories
 - other persistent records remaining outside this particular result

This was AI-assisted classification, not chatting or model training.

- inference over existing documents
 - a fixed model applying categories to each record
 - output consisting of classifications and supporting extracts
- categories supplied by researchers
 - research judgement expressed before the model ran
 - human checking performed during calibration
 - a sample of 210 documents reviewed manually against the model’s reasoning
 - categories removed or refined when their results did not answer the researchers’ question ([Johnston et al., 2026](#), sec. “Results of first iterations”)

The Same Record Serves Different Actors

DISTRIBUTING BENEFIT, COST, RISK, AND CONTROL

- Participants used Kidlink to build international relationships and act together.
- Kidlink stored contributions so the public could continue across later projects.
- Archivists and researchers gained records that preserve history and support new knowledge.
- The people recorded could not control every later audience or use.
- Opportunities and risks differ by stakeholder.

Participants used Kidlink to build international relationships and act together.

- live exchanges across countries
 - children addressing one another rather than merely submitting data
 - geographical distance reduced inside the exchange
- relationships and joint projects produced through participation
 - recurring projects sustaining contact over time
 - shared activity emerging through the public
 - “Geographical distance between people are quickly disappearing as a major obstacle”
 - the organisers’ contemporary account of talking and acting together ([Mizell et al., 1992, p. 8](#))

Kidlink stored contributions so the public could continue across later projects.

- stored messages carrying activity into later projects
 - earlier contributions available to new participants
 - stored contributions carrying continuity across projects
- recurring forums and project structures supporting coordination
 - adult KIDPLAN forums supporting project administration
 - annual projects providing an organised place for shared planning

Archivists and researchers gained records that preserve history and support new knowledge.

- preservation making later public access possible
 - a surviving record of children’s early online participation
 - incomplete web captures remaining useful historical sources

- academic use as another interaction between communities
 - researchers constructing questions that make some records newly visible
 - ethical and technical controls shaping what researchers may do
 - “the ethics protocols governing the corpus do not permit data to leave university infrastructure”
 - self-hosted processing as a condition of this research use ([Johnston et al., 2026](#), sec. “Building document-level binary classifications through inference”)

The people recorded could not control every later audience or use.

- public availability over three decades
 - Archive.org allowing later readers to revisit the captured exchange
 - persistence extending beyond participants’ likely temporal horizon

- original context and expectations still missing
 - archival capture unable to preserve every relationship or intention
 - later analysis asking questions unavailable to the original participants
 - copied excerpts still reconnectable to the public archive through wording, dates and links

Opportunities and risks differ by stakeholder.

- participants gaining connection while later actors gain evidence
 - social benefit located in the original encounter
 - historical benefit located in later access
- researchers controlling selection while participants bear later visibility
 - selection, interpretive control and visibility distributed to different actors
 - ethical judgement required because public availability does not settle every later use

Social Steganography

VISIBLE CONTENT AND UNEQUAL ACCESS TO MEANING

hiding messages in plain sight by leveraging shared knowledge and cues embedded in particular social contexts

([boyd, 2014, p. 65](#))

- Social steganography uses shared context to hide meaning in content visible to several audiences.
- “The Iranian yogurt is not the issue here” became Reddit shorthand for fixating on a detail while missing the larger issue.
- Reddit can record, search and monetise activity without sharing the community’s interpretation.

Social steganography uses shared context to hide meaning in content visible to several audiences.

- Outsiders and insiders can obtain the same words or image.
 - the post remaining publicly accessible across audiences
 - no technical restriction required to make interpretation unequal
- Shared history gives the intended public another reading.
 - prior conversation and community knowledge supplying the key
 - access to content separated from access to meaning
 - “limiting access to meaning can be a much more powerful tool for achieving privacy than trying to limit access to the content itself” ([boyd, 2014, p. 69](#))
 - boyd distinguishes protection of meaning from successful control over who can obtain the content

“The Iranian yogurt is not the issue here” became Reddit shorthand for fixating on a detail while missing the larger issue.

- A 2019 commenter proposed the idiom after a relationship dispute about yogurt.
 - the original post contrasting an absurd literal detail with a dispute about permission, a relationship and a shared home
 - the community proposal turning one post’s phrase into a reusable judgement ([r/AmltheAsshole, 2019](#))
 - “Anytime someone seems to hang up on an absurd detail that misses the larger point, people should just say ‘The Iranian yogurt is not the issue here.’”
 - a community-made interpretation rather than a platform rule

- Its later reuse exposed unequal understanding without establishing deliberate concealment.
 - the phrase appearing alone in an unrelated 2022 dispute ([r/AmltheAsshole, 2022a](#))
 - “The Iranian yogurt is not the issue here.”
 - no explanation accompanying the reused phrase
 - an outsider asking for the missing community history ([r/AmltheAsshole, 2022b](#))
 - “What’s the reference for the Iranian yogurt, I haven’t seen the post.”
 - archived evidence of unequal interpretation, not proof of the commenter’s intent

Reddit can record, search and monetise activity without sharing the community’s interpretation.

- Reddit processes posts, comments, profiles and engagement as public records and targeting signals.
 - public content grouped as posts, comments, usernames and profiles
 - Reddit describes these records as part of a platform that is “public and accessible to everyone” ([Reddit, 2025](#))
 - the policy states Reddit’s current category rather than a participant’s imagined audience
 - search and audience classification acting on the recorded activity
 - Reddit search covers posts, comments, communities, media and people ([Reddit, 2026a](#))
 - advertising audiences may be based on communities people have “joined or recently engaged with” ([Reddit, n.d.](#))

- Community history and insider meanings remain outside those stored fields and classifications.
 - community history not represented as an explicit field and not always recoverable from the record
 - an insider reading not reducible to the platform’s classification

Epigraph: “hiding messages in plain sight by leveraging shared knowledge and cues embedded in particular social contexts” ([boyd, 2014, p. 65](#))

- the term boyd developed with Alice Marwick

- the definition placing the method in shared social knowledge rather than technical concealment

Advertising Changed Online

FROM PAGE CONTEXT TO INFERRED AUDIENCES

- Contextual advertising matches an advertisement to page content. Audience targeting matches it to people grouped by earlier activity.
- **Persistence:** Reddit can use recent subscriptions and visits to define an audience.
- **Visibility:** public posts, comments, usernames and profiles are widely accessible.
- **Spreadability:** reposting moves eligible content between communities.
- **Searchability:** users find content, while advertisers work from classified audiences.

Contextual advertising matches an advertisement to page content. Audience targeting matches it to people grouped by earlier activity.

- a page topic attracting a related advertisement
 - the current page supplying the relevant context
 - no prior account activity required for that match
 - Google defines contextual targeting as matching an advertisement to a webpage's analysed theme
 - current first-party documentation, not evidence of how every ad system operates ([Google, n.d.](#))
- a community visit becoming an audience signal that can affect advertising elsewhere on Reddit
 - earlier behaviour used to place an account in a group
 - the advertisement following the inferred audience rather than the page

Persistence: Reddit can use recent subscriptions and visits to define an audience.

- earlier community participation retained
 - subscriptions and recent engagement treated as audience evidence
 - a recent-activity window turning past activity into a current classification
- later advertising opportunity ([Reddit, n.d.](#))
 - advertisements shown outside the community that supplied the signal
 - Reddit's current public page confirms recent engagement as the basis
 - "People on Reddit who have joined or recently engaged with those subreddit communities will see your ad as they browse Reddit."
 - Reddit does not define an exact duration for "recent" on this page

Visibility: public posts, comments, usernames and profiles are widely accessible.

- access beyond the account holder's intended public
 - viewing possible without a Reddit account
 - general availability extending beyond the community addressed

- platform rules defining public content
 - general availability asserted by Reddit
 - "Most of Reddit's platform is public and accessible to everyone"
 - the statement describing Reddit's current policy rather than every participant's imagined audience ([Reddit, 2025](#))
 - content types grouped together by that policy
 - posts, comments, usernames, profiles, karma and related metadata
 - public status defined by the platform across several kinds of record

Spreadability: reposting moves eligible content between communities.

- a platform mechanism for circulation
 - Reddit's own description of the action
 - "take a post from one community and share it with another community"
 - the originating post and account retained in a preview ([Reddit, 2026c](#))
 - destination-community rules deciding whether reposting is available
 - moderators able to disable the mechanism
 - circulation therefore depending on the receiving public's governance

- the public changing with the destination
 - a second community applying different expectations and shared context
 - the same contribution becoming legible inside a different conversation

Searchability: users find content, while advertisers work from classified audiences.

- participant-facing discovery through Reddit search ([Reddit, 2026a](#))
 - posts, comments, communities, media and people as searchable types
 - search scoped to one community when a user chooses
- advertiser targeting as classification rather than content search ([Reddit, n.d.](#))
 - communities selected as proxies for people likely to be interested
 - ads shown across Reddit to people who interacted with those communities
 - the same platform offering different possibilities to participants and advertisers
 - targeting supporting paid access to an inferred audience ([Bucher & Helmond, 2018](#))

I Didn't Set a Cookie

A PARTIAL PROFILE FROM BROWSER-EXPOSED SIGNALS

- Rejecting cookies blocks one way of recognising a browser.
- Some signals can be arranged into the form of an advertising bid request.
- Platforms, data providers and advertisers can use partial profiles to define an audience and price access to it.

Displayed demonstration: **Kuber Studio's "I didn't set a cookie" page**

(**Mehta, 2026**)

Rejecting cookies blocks one way of recognising a browser.

- cookies as stored identifiers
 - a site placing a value in browser storage
 - that value returning on a later visit
- other signals still exposed
 - information available without accepting that stored identifier
 - hardware, configuration and installed resources as examples
- browser fingerprinting
 - a combination of ordinary browser-exposed features used to distinguish one browser from others
 - probabilistic recognition rather than a name supplied by the user
 - “most of [the browser interfaces] are freely accessible from any JavaScript script”
 - stateless recognition without storing an identifier in the browser ([Laperdrix et al., 2020, p. 2](#))

Some signals can be arranged into the form of an advertising bid request.

- the device and impression represented in OpenRTB
 - the advertising opportunity described in a standard message
 - an exchange broadcasting the request to potential bidders
- inferred audience data attachable by a provider
 - a user-data object carrying provider-defined audience segments
 - the example leaving those paid-for segments as placeholders
 - “bid requests are broadcast to bidders, responses are evaluated under prevailing auction rules, and a winner is selected”
 - OpenRTB specifying the message format rather than proving which fields any particular auction contains ([OpenRTB 2.6, 2022, § 2](#))

Platforms, data providers and advertisers can use partial profiles to define an audience and price access to it.

- patterns in recorded behaviour supporting an inference
 - Target’s purchasing data used to infer pregnancy
- the reported inference occurring before some customers disclosed the pregnancy to their families
 - purchasing data used to “determine when women were pregnant even before they told their family members”
 - Gates reprises the reported Target case rather than independently validating its model ([Gates, 2025, p. 52](#))

- signals linked to one profile
 - an exchange-specific user identifier available for retargeting
 - a buyer-specific identifier mapped to the same user
 - OpenRTB describes the User object as information “known or derived about the human user of the device”
 - its identifiers and segments specify the transaction format, not the accuracy of the attached profile ([OpenRTB 2.6, 2022, § 3.2.20](#))
- an inferred audience classification
 - provider-defined segments attachable to the user representation
 - a probable interest or condition rather than a directly stated fact
 - Segment objects convey specific units of data supplied by a named provider
 - the standard permits the representation but does not establish how an inference was produced ([OpenRTB 2.6, 2022, § 3.2.22](#))
- bidding for the opportunity to reach that audience
 - a bid request offered to potential buyers
 - “bid requests are broadcast to bidders” and their responses evaluated under the auction rules
 - the standard describes competitive buyers of advertising inventory ([OpenRTB 2.6, 2022, § 2](#))
 - an addressable opportunity rather than the sale of a complete person
 - the message includes selected information “known or derived about the human user”
 - the auction prices an opportunity to display an advertisement through that representation ([OpenRTB 2.6, 2022, § 3.2.20](#))

Displayed demonstration: [Kuber Studio’s “I didn’t set a cookie” page](#)

- browser-accessible signals
 - operating system, screen, browser features and installed resources
 - a deliberately startling inventory of what the page can observe
 - “This page set none. It asked for zero permissions. You have clicked nothing.”
 - a claim about this live page rather than every possible browser configuration ([Mehta, 2026](#))
- a constructed OpenRTB-shaped example
 - placeholders where a buyer could attach audience segments
 - no proof of a real auction ([Mehta, 2026](#))
 - “We can’t show yours because we’re not a paying buyer. The bidders can.”
 - a live demonstration whose exact output varies by browser and device

Why This Matters

FROM ADDRESSABLE AUDIENCE TO LICENSABLE CORPUS

- Reddit participation and browser signals can become usable data.
- Advertisers and platforms act through partial representations of audiences.
- A partial profile omits participants' meanings but still shapes what happens to them.
- Platforms can commercialise access to audiences and to accumulated conversation.

Reddit participation and browser signals can become usable data.

- community activity used to define an advertising audience
 - earlier participation persisting as an interest signal
 - that signal acting beyond the original community
 - “People on Reddit who have joined or recently engaged with those subreddit communities will see your ad as they browse Reddit.”
 - Reddit describes the audience rule, but does not disclose how well the resulting category represents any particular participant ([Reddit, n.d.](#))
- browser features arranged as a bid request
 - a person represented through selected technical fields
 - a commercial actor valuing the opportunity through that representation
 - OpenRTB defines the User object as information “known or derived about the human user of the device” ([OpenRTB 2.6, 2022, § 3.2.20](#))
 - the Kuber example shows the message structure, but cannot display the paid audience segments held by a data provider ([Mehta, 2026](#))

Advertisers and platforms act through partial representations of audiences.

- advertisers selecting whom to reach
 - a community or segment standing in for likely interest
 - a bid calculated without knowing the whole person
 - an inferred group is sufficient when the commercial purpose is to decide whether an opportunity to display an advertisement is valuable
 - OpenRTB carries audience data as provider-defined segments rather than as a complete account of a person ([OpenRTB 2.6, 2022, §§ 3.2.21–3.2.22](#))
- platforms defining the available profile
 - fields and categories chosen for the platform’s commercial purpose
 - some traces made usable while other context disappears
 - the standard supplies fields for the transaction; it does not preserve the relationships that made a post or community visit meaningful
 - browser fingerprinting similarly distinguishes devices through a selected combination of exposed features ([Laperdrix et al., 2020, p. 2](#))

A partial profile omits participants’ meanings but still shapes what happens to them.

- community context missing from behavioural signals
 - an in-joke or relationship reduced to a visit, click or category
 - social meaning unavailable to the targeting system
 - social steganography depends on “shared knowledge and cues embedded in particular social contexts” ([boyd, 2014, p. 65](#))
 - the targeting category records activity associated with a community, not the meaning that activity held within the community
- classification influencing who receives which message
 - aggregate associations applied to one linked profile
 - a probable category producing a personal consequence
 - the reported Target case used purchasing patterns to infer pregnancy before some customers had disclosed it to their families
 - Gates reprises that case as an illustration; it is not independent verification of Target’s model ([Gates, 2025, p. 52](#))

Platforms can commercialise access to audiences and to accumulated conversation.

- advertising access as one product
 - advertisers paying to reach a classified audience
 - continued participation supplying both attention and signals
 - “The users themselves are an important source of revenue” ([Birch & Cochrane, 2022, p. 49](#))
 - the platform sells access to an audience while retaining control of the records and rules that constitute that audience commercially
- structured conversation as another
 - the accumulated record valuable to search and model developers
 - controlled access creating another revenue stream
 - Reddit markets its record as “authentic and constantly-updated human-generated conversations”
 - the partnership promises “more efficient ways to train models” from that licensed access ([Reddit, 2024](#))

Some Participation Becomes Training Material

REDDIT, DATA LICENSING, AND “AUTHENTIC” CONVERSATION

- Communities created Reddit’s public conversations through years of participation.
- Public pages and Reddit’s earlier data-access service made much of that conversation collectable.
- Reddit now controls structured access and charges for broader commercial use.
- “Authentic human conversation” is a corporate sales claim.

Communities created Reddit’s public conversations through years of participation.

- accumulated relationships and knowledge
 - answers, disagreements and recommendations built across recurring use
 - community histories making later contributions intelligible
- shared meaning making the record valuable
 - conversational context distinguishing Reddit text from disconnected sentences
 - participant labour producing what the platform can later package

Public pages and Reddit’s earlier data-access service made much of that conversation collectable.

- public pages available through web crawling
 - pages technically reachable without an account
 - this route distinct from Reddit’s structured Data API
 - Google distinguishes licensed API access from “publicly available, crawlable content”
 - the partnership statement establishes the two routes, not the contents of a particular crawl ([Patel, 2024](#))
- structured records collected into external datasets
 - Pushshift collecting Reddit data from 2015 and making it available to researchers
 - its 2020 paper reporting historical data back to Reddit’s inception
 - a documented example of an external copy existing before the current licensing business
 - the paper does not establish that every Reddit contribution was copied ([Baumgartner et al., 2020](#))

Reddit now controls structured access and charges for broader commercial use.

- infrastructure and access controlled by the platform
 - authenticated interfaces supplying data in platform-defined fields
 - rate limits and contracts separating different uses
- licensable value produced through community participation
 - digital rentiership through control of access
 - free and paid routes set by the platform
 - “Bulk exporting of Reddit data will be significantly limited by default”
 - broader access available for a fee ([Reddit, 2026b](#))
- control over the ecosystem creating the rent
 - the platform controlling data, rules and enforcement
 - outsiders paying for access to a concentrated public and its record
 - enclave rents as revenues derived from control of an ecosystem and its user-generated data
 - “The users themselves are an important source of revenue” ([Birch & Cochrane, 2022, p. 49](#))

“Authentic human conversation” is a corporate sales claim.

- Reddit using “authentic” to market licensed conversation
 - access sold for products and model training
 - Reddit’s own description
 - “one of the internet’s largest open archives of authentic and constantly-updated human-generated conversations”
 - “more efficient ways to train models” ([Reddit, 2024](#))
- the seller’s description rather than an independent finding
 - Google repeating the value claim as the buyer and partner
 - authenticity functioning as a reason to prefer this licensed corpus
 - “an incredible breadth of authentic, human conversations and experiences”
 - real-time, structured and fresher access through the Data API ([Patel, 2024](#))

Corpora Are Constructed

COMMON CRAWL AND LICENSED PLATFORM ACCESS

- A training corpus is assembled from selected sources.
- Common Crawl stores dated captures from part of the public web.
- A licensed platform API can provide structured and current access.
- Every collection route includes some material and omits other material.

A training corpus is assembled from selected sources.

- collection before learning
 - documents acquired through crawls, archives, licences or direct creation
 - the collection route determining what is available to include
- inclusion and exclusion before aggregation
 - technical access setting part of the available pool
 - collector choices determining which sources enter the corpus

Common Crawl stores dated captures from part of the public web.

- public pages reached by a particular crawl
 - technical accessibility making a page eligible rather than guaranteeing capture
 - capture depending on what the crawler selected and when it visited
- neither all Reddit nor the whole internet
 - crawl scope and capture dates
 - pages absent, blocked or changed outside a particular capture
 - “petabytes of data, regularly collected since 2008”
 - raw page data, metadata extracts and text extracts rather than a complete internet ([Common Crawl Foundation, n.d.](#))

A licensed platform API can provide structured and current access.

- platform-defined fields and rate limits
 - posts and comments delivered as structured records
 - freshness marketed as exceeding an occasional public-web crawl

- negotiated terms and permitted uses
 - bounded free access alongside enterprise arrangements
 - Reddit documenting approval and rate limits for its Data API
 - broader commercial access potentially incurring fees ([Reddit, 2026b](#))
 - commercial uses governed through platform terms
 - licensing and models trained on Reddit data listed as commercial uses
 - the terms establishing Reddit’s current route rather than every API provider’s rules ([Reddit, 2026b](#))
 - current structure and freshness marketed as benefits of licensed access
 - Google describing “real-time, structured, unique content” and “fresher information”
 - the partnership leaving Google’s separate use of crawlable public content unchanged ([Patel, 2024](#))

Every collection route includes some material and omits other material.

- timing and technical accessibility
 - a crawl retaining what was reachable at a collection date
 - an API returning what its current fields and permissions expose

- selection and commercial control
 - different corpora from the same public
 - no model simply reading the whole internet
 - Common Crawl exposing its corpus by dated crawl and URL index
 - Reddit governing bulk and commercial API access through approval, limits and fees ([Common Crawl Foundation, n.d.](#); [Reddit, 2026b](#))

Training Changes Model Parameters

STATISTICAL PATTERNS ACROSS MANY EXAMPLES

- One selected contribution enters a corpus alongside many others.
- Training repeatedly asks a model to predict the next piece of text.
- Prediction errors adjust numerical settings called parameters.
- Repeated updates encode aggregate statistical patterns.

One selected contribution enters a corpus alongside many others.

- the contribution included in a selected corpus
 - one document contributing text to the collection
 - selection already reflecting the corpus builder’s boundaries
- its statistical effect depending on its relation to many other examples
 - recurring patterns reinforced across the collection
 - this aggregate account providing no readable ledger from one contribution to one resulting capability
 - Karpathy compares roughly 10 TB of text with a 140 GB parameter file
 - his “lossy compression” metaphor illustrates aggregation without quantifying one contribution’s effect ([Karpathy, 2023, pp. 00:05:25–00:06:06](#))

Training repeatedly asks a model to predict the next piece of text.

- pieces of text supplied as input
 - text represented as a sequence the model can process
 - preceding pieces supplying the prediction context
- a target continuation used for comparison
 - the known next piece treated as the training target
 - predicted probabilities compared with what appeared in the example
 - Karpathy’s deliberately simplified account is “trying to predict the next word in a sequence”
 - the assigned transcript says “word”, not the more precise token-level account ([Karpathy, 2023, pp. 00:06:51–00:07:51](#))

Prediction errors adjust numerical settings called parameters.

- the next-word task supplying a measure of improvement
 - a prediction compared with the continuation in the training sequence
 - better performance defined by the prediction task
 - Karpathy says “we can measure that it’s getting better at the next word prediction”
 - the talk does not teach the loss function or back-propagation ([Karpathy, 2023, pp. 00:12:21–00:12:30](#))
- parameters adjusted iteratively
 - many updates producing a changed model state
 - the individual parameters remaining difficult to interpret
 - “we know how to adjust these parameters iteratively to make the network as a whole better at the next word prediction task”
 - Karpathy immediately qualifies that “we don’t actually really know what these 100 billion parameters are doing” ([Karpathy, 2023, pp. 00:12:03–00:12:21](#))

Repeated updates encode aggregate statistical patterns.

- aggregate learning as the ordinary account
 - likely continuations changing across the corpus
 - no clean path from one contribution to one capability
 - parameters as “a kind of a Gestalt of the text that we trained on”
 - lossy compression rather than an identical copy ([Karpathy, 2023, pp. 00:05:29–00:06:06](#))
- memorisation remaining a distinct possibility
 - some sequences reproduced more directly than the aggregate account suggests
 - no simple inspection revealing which output follows which route
 - “some of the stuff could be memorized and some of it is not memorized and you don’t exactly know which is which”
 - ordinary generation described as “dreaming internet text from its data distribution” ([Karpathy, 2023, pp. 00:11:12–00:11:23](#))

Training Is Not Chatting

UPDATING PARAMETERS VERSUS GENERATING RESPONSES

- Training changes parameters. Chatting generates responses with existing parameters.
- A conversation supplies temporary context.
- A provider can store chats or memories without changing the base model during the conversation.
- Telling a chatbot something does not ordinarily teach the base model for future users.

Training changes parameters. Chatting generates responses with existing parameters.

- parameter updates during training
 - repeated comparison against training examples
 - a new model state produced after substantial computation
 - Karpathy describes a “long process of optimization” producing the neural network
 - the final parameter file is the product used for later generation ([Karpathy, 2023, pp. 00:13:29–00:13:41](#))
- inference during ordinary use
 - the stored parameter file used to calculate a continuation
 - no retraining required for each response
 - Anthropic engineer Jason D. Clinton writes that “The model file never changes and is immutable once loaded”
 - one static file is loaded across many serving instances (quoted in [Willison, 2024](#))

A conversation supplies temporary context.

- messages inside the current context window
 - earlier turns included with the next prompt
 - a bounded sequence influencing the immediate response
- context ending without a parameter update
 - a new conversation beginning with cleared conversational context unless a separate memory feature supplies retained information
 - the base model remaining the same across those conversations
 - “Every time you start a new chat conversation, you clear the slate.”
 - “Once trained, the model remains static and unchanged” ([Willison, 2024](#))

A provider can store chats or memories without changing the base model during the conversation.

- service logs existing outside the model’s parameters
 - prompts reasonably assumed to be logged for compliance and abuse reasons
 - later exposure or reuse remaining a storage and policy question
 - Willison warns that logging creates a possibility of exposure
 - he does not claim to document every provider’s current retention policy ([Willison, 2024](#))

- separate memory features retaining selected information
 - stored notes inserted into future context
 - product memory distinct from changing the base model
 - the memory feature records a short note “invisibly included in the chat context passed in future conversations”
 - stored context rather than a parameter update ([Willison, 2024](#))

Telling a chatbot something does not ordinarily teach the base model for future users.

- no immediate base-model update
 - another user’s conversation not inheriting the new statement
 - memory and retention remaining separate from parameter training
 - conversational imitation creating a misleading appearance of learning
 - stored material not itself updating the base model
- later reuse requiring a separate training decision
 - provider retention allowing later selection as training data
 - another training run required to update a future model
 - no immediate update for other users
 - the operational distinction
 - “There is no point at all in ‘telling’ a model something in order to improve its knowledge for future conversations.”
 - “do not directly learn from and memorize everything that you say to them” ([Willison, 2024](#))

Why This Matters

FROM SUPPORTED CLAIM TO DEFENSIBLE RESPONSE

- Reddit now controls current structured access, while earlier external datasets remain outside that route.
- Evidence supports only the process or change that a source documents.
- A defensible response names an actor with control and the consequence their action should change.

Reddit now controls current structured access, while earlier external datasets remain outside that route.

- licensing changing one route into data
 - contracts governing current commercial API access
 - fresher structured records sold through the partnership
 - Reddit says bulk export is limited by default and broader access may incur fees ([Reddit, 2026b](#))
 - Google describes the licensed route as “real-time, structured, unique content” ([Patel, 2024](#))
- earlier collected copies remaining outside the new arrangement
 - Pushshift as one documented dataset collected before the current terms
 - the 2020 paper reporting a real-time archive with data back to Reddit’s inception
 - evidence for one external dataset, not every claimed historical copy
 - current access rules unable by themselves to erase a separately held research dataset ([Baumgartner et al., 2020](#))

Evidence supports only the process or change that a source documents.

- platform documents establishing access and terms
 - first-party evidence for what Reddit and Google say they provide
 - corporate descriptions not proving authenticity or social value

- technical sources establishing how training works
 - Karpathy and Willison supporting the training-chat distinction
 - neither source establishing that a particular Reddit post trained a particular model
 - mechanism evidence separated from provenance evidence
 - a bounded claim more defensible than an attractive inference ([Karpathy, 2023](#); [Willison, 2024](#))

A defensible response names an actor with control and the consequence their action should change.

- control located in a platform, institution or lawmaker
 - a platform changing access or contract terms
 - an institution or lawmaker changing enforceable conditions
- expected effect stated for an affected public
 - the relevant community and burden named
 - a causal path connecting the proposed action to the intended change
 - evidence establishing the problem before a response is recommended
 - a response judged against what it can actually alter

When Models Train on Generated Data

COLLAPSE, CURATION, AND THE MYTH OF PURITY

- Generated material can return to online platforms and later training corpora.
- Repeated replacement with generated material can amplify errors and remove uncommon patterns.
- Synthetic data can help when they are selected, mixed and tested.
- The evidence does not support a simple human-good, synthetic-bad rule.

Generated material can return to online platforms and later training corpora.

- model output published on collectable surfaces
 - generated text published as posts, pages or articles
 - human and generated material entering the same collection surface
- later collection determining whether it becomes training data
 - publication making collection possible rather than inevitable
 - crawler scope, licensing and filtering deciding later inclusion
 - Shumailov and colleagues study the case in which model-generated text found online becomes material for successor models
 - the experiment establishes a feedback regime, not the composition of any current commercial corpus ([Shumailov et al., 2024, p. 755](#))

Repeated replacement with generated material can amplify errors and remove uncommon patterns.

- the original distribution replaced across generations
 - each new generation sampled from its predecessor
 - original examples absent from later training rounds
- narrowing and error compounding across recursive generations
 - low-probability events disappearing first
 - “tails of the original content distribution disappear”
 - the reported loss beginning in the distribution’s tails
 - model errors becoming later inputs
 - recursive replacement of original data as the tested regime
 - “indiscriminate use of model-generated content” rather than the mere presence of synthetic data ([Shumailov et al., 2024, p. 755](#))

Synthetic data can help when they are selected, mixed and tested.

- generated examples retained alongside original data
 - selected cases serving a defined task
 - accumulation replacing recursive substitution
 - Gerstgrasser and colleagues “demonstrate that accumulating the successive generations of synthetic data alongside the original real data avoids model collapse”
 - the result holds for their tested models and regimes rather than every possible use of synthetic data ([Gerstgrasser et al., 2024](#))
- performance checked against an independent target
 - the accumulation study testing later models against original data
 - an analogous research-method principle in Ross and Ballsun-Stanton
 - verification “anchored to an external source” rather than a model checking its own persuasive output
 - a parallel about independent validation, not technical evidence about synthetic training ([Ross & Ballsun-Stanton, 2026, p. 2](#))

The evidence does not support a simple human-good, synthetic-bad rule.

- the two studies testing different data regimes
 - recursive replacement producing collapse in Shumailov and colleagues
 - accumulation with original data avoiding collapse in Gerstgrasser and colleagues
- selection, mixture and testing carrying the relevant distinction
 - authorship naming origin rather than demonstrated quality
 - usefulness judged against the model’s intended task

Is the Internet Dead?

BOT TRAFFIC, GENERATED CONTENT, AND HUMAN PRESENCE

- Dead Internet Theory claims that apparent online social life is increasingly synthetic.
- A platform can present people, bots and generated propaganda through the same interface.
- Traffic estimates do not count people or publics.
- Evidence of automation does not establish that a public is synthetic.

Dead Internet Theory claims that apparent online social life is increasingly synthetic.

- bots or generated personas appearing as people
 - an account or reply presenting the signs of a participant
 - no corresponding human author behind some apparent participation
- apparent interaction without a corresponding human public
 - the strong version extending from some synthetic actors to synthetic social life
 - weaker complaints about repetition, platform decline or bad content not establishing that claim
 - the influential 2021 formulation opens, “The Internet feels empty and devoid of people”
 - the same post moves from felt absence to sweeping claims about bots, corporations and government manipulation ([IlluminatiPirate, 2021](#))

A platform can present people, bots and generated propaganda through the same interface.

- shared account, post and reply forms concealing how content was produced
 - visible platform features concealing how a contribution was produced
 - surface plausibility unable to establish authorship from one post
- generated propaganda changing what a public encounters without proving that the public itself is synthetic ([Klincewicz et al., 2025](#))
 - mass-produced material altering the public's informational environment
 - influence operating through scale, targeting and repetition
 - “altering the informational environment en masse to reach a desired aggregate change in decision-making”
 - a mechanism of influence rather than proof that human participants have disappeared ([Klincewicz et al., 2025, p. 139](#))

Traffic estimates do not count people or publics.

- Prada’s headline claim of one-third of internet traffic
 - a current-affairs claim about automated traffic
 - a headline that turns a traffic statistic into a claim about the internet as a social space

- Cloudflare traffic cited as the underlying measure
 - the underlying measurement unit not preserved in Prada’s article
 - the article inferring bots communicating with bots
 - “nearly one-third of all internet traffic now comes from bots”
 - Prada’s article reports the Cloudflare figure but does not preserve the underlying measurement method ([Prada, 2025](#))
 - no direct measure of account or public authenticity
 - traffic share unable to identify the proportion of participants
 - a measured connection or transfer not equivalent to a person

Evidence of automation does not establish that a public is synthetic.

- bounded studies answering narrower questions
 - Ng and Carley’s classification of accounts across seven events
 - “Chatter on social media about global events comes from 20% bots and 80% humans.”
 - one platform, operational definition and classifier rather than an estimate of the whole internet ([Ng & Carley, 2025, p. 1](#))
 - traffic, accounts, posts and relationships remaining different units
 - evidence about one unit not establishing the composition of another
 - the broad Dead Internet claim requiring evidence across those units
- an investigation comparing the measure with the conclusion
 - the exact claim and unit of analysis stated first
 - source chain followed to the underlying dataset or method
 - uncertainty retained when that method cannot be recovered
 - evidence kept within its measured scope
 - traffic evidence answering a traffic question
 - account classification answering a bounded platform question

What Changes the Relationship?

SYMBOLIC, TECHNICAL, INSTITUTIONAL, AND MARKET RESPONSES

- A “NO AI TRAINING” notice expresses refusal and signals affiliation.
- A community norm does not itself change access, rules or authority.
- Communities can alter collected data, restrict access, seek legal force, leave or fund alternatives.
- A response can be judged by what it changes, for whom and at whose cost.

A contemporary “NO AI TRAINING” notice

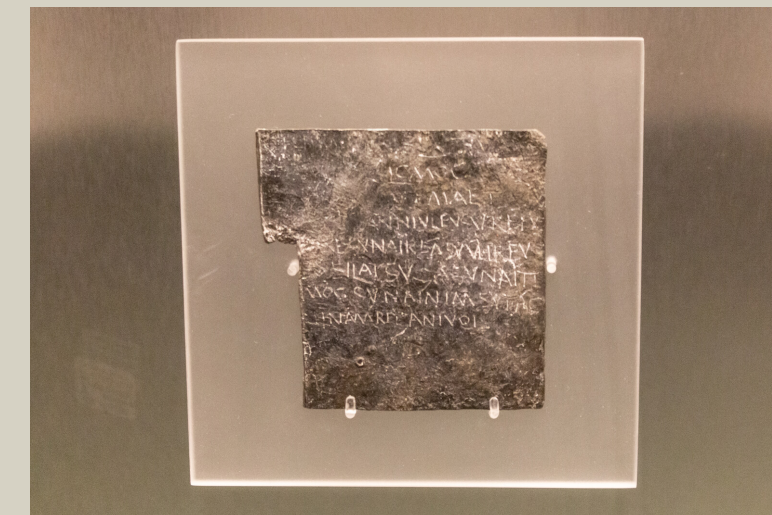
NO AI TRAINING: Without in any way limiting the author's and publisher's exclusive rights under copyright, any use of this publication to “train” generative artificial intelligence (AI) technologies to generate text is expressly prohibited. The author reserves all rights to license uses of this work for generative AI training and development of machine learning language models.

It is a very odd feeling seeing history repeat itself
cause back in *my* day, guide writers on gamefaqs would plaster copyright notices at the top of their faqs, in a vain attempt to get scrapers to stop stealing their work
and as the months/years went by their 'copyright notices' would get increasingly lengthy and unhinged, yelling into the void
because it turns out that whether it's the 90's or today
content scrapers give zero fucks about legality

A contemporary notice prohibiting use of a publication for AI training

Private Discord correspondence, reproduced with permission (personal communication, 13 August 2026)

A Roman curse tablet



A metal Roman curse tablet complaining about the theft of Vilbia

May he who has stolen VILBIA from me become as liquid as water ...

([Roman Inscriptions of Britain, n.d.](#))

Tablet photograph by -JvL-, CC BY 2.0 ([-JvL-, 2022](#))

A “NO AI TRAINING” notice expresses refusal and signals affiliation.

- a community norm and political position
 - a desired boundary stated in terms other members recognise
 - authors aligning themselves with a shared position
 - social force without technical enforcement
 - the notice making refusal legible to its intended public
- a speech act whose force depends on recognition
 - the curse-tablet comparison
 - an invocation making a norm and threatened consequence visible
 - efficacy depending on a reader recognising or respecting the claim
 - the notice as one performative example
 - “any use of this publication to ‘train’ generative artificial intelligence” declared prohibited
 - no evidence in the image that a crawler detected or obeyed it

A community norm does not itself change access, rules or authority.

- a public notice remaining technically ignorable without another control
 - public technical access remaining unchanged
 - no machine-readable control demonstrated by wording alone
- a public notice able to become part of what is collected
 - refusal posted on a collectable surface becoming another record
 - the private Discord example demonstrating the speech act, not public accessibility or scraping

Communities can alter collected data, restrict access, seek legal force, leave or fund alternatives.

- technical and legal responses changing access or authority
 - data poisoning aimed at collection or model performance
 - Nightshade targets prompt-image associations with image-text pairs that appear benign
 - its controlled experiments do not establish effectiveness against a current frontier or web-scale training pipeline ([Shan et al., 2023](#))
 - access controls and law altering available routes or enforceable duties
 - authentication, rate limits and interface rules leaving distributed copies untouched
 - jurisdiction and enforcement limiting legal reach

- exit and user funding changing the economic relationship
 - exit withholding participation at the cost of connection
 - Number Six’s resignation and the continuing demand “by hook or by crook”
 - withholding future contribution as the user’s available leverage
 - prior records and the institution’s demand remaining
 - Kagi as a subscription-funded contrast
 - Kagi says it has “no ads, no tracking” and is “fully funded by its users”
 - the direct customer is the person buying search access ([Kagi, n.d.](#))
 - subscription changing the funding relationship without proving virtue
 - a user-funded service contrasted with advertiser-funded access to an audience
 - first-party business-model evidence rather than an independent privacy audit or proof of virtue
 - a paying user’s exit as a current response
 - Rands considering cancellation over Anthropic’s announced watermarking
 - objection tied to a product decision and a recurring payment
 - “Getting in the mindset to cancel my Claude Max plan.”
 - the author’s EU-regulation explanation explicitly presented as a guess ([Rands, 2026](#))

A response can be judged by what it changes, for whom and at whose cost.

- the response’s mechanism and target
 - symbolic force establishing affiliation and norms
 - technical, legal or economic force altering a named system

- the distribution of benefit, cost and control
 - evidence that the response changes something for the affected public
 - costs borne by institutions, communities and departing participants

Displayed visual pair: contemporary “NO AI TRAINING” notice and Roman curse tablet

- two invocations making a desired boundary and consequence legible in a recognised form
 - the contemporary notice declaring AI training prohibited
 - the Roman tablet asking for a thief to suffer until the loss is remedied

- the comparison concerning performative form rather than equivalent beliefs, institutions or enforcement
 - the tablet invokes, “May he who has stolen VILBIA from me become as liquid as water ...” ([Roman Inscriptions of Britain, n.d.](#))
 - tablet photograph by -JvL-, CC BY 2.0 ([-JvL-, 2022](#))
 - “Vilbia” possibly a person or a pointed object, with that uncertainty not affecting the comparison between the two written invocations

- “Private Discord correspondence, reproduced with permission (personal communication, 13 August 2026)”

Books Become Data

TEXT PRESERVED, PHYSICAL COPIES DESTROYED

- In *Rainbows End*, a library preserves the text by destroying the books.
- 404 Media used an AirTag to follow a book to an Amazon facility that scans books for AI training.
- The report does not show what happened to the tagged book.

[404 Media, “We Tracked a Shipment of Rare Books,” 17 August 2026](#)

In *Rainbows End*, a library preserves the text by destroying the books.

- fictional destructive digitisation
 - pages separated and imaged for digital reconstruction
 - words preserved while bound copies disappear
- a difference between preserving a text and preserving its physical form
 - one object becoming raw material for another
 - fiction supplying the analogy rather than evidence about Amazon

404 Media used an AirTag to follow a book to an Amazon facility that scans books for AI training.

- a tracker placed in one book from an order of about 1,000
 - the book arriving at Amazon's VGT3 operation in Las Vegas
 - the journey established by the tracking data

- employees describing the facility's scanning process
 - bindings removed so books can be scanned quickly
 - physical copies destroyed in the process
 - [404 Media, 17 August 2026](#)

The report does not show what happened to the tagged book.

- tracking data establishing a destination
 - direct evidence about the book's journey
 - no direct observation after its arrival
- employee accounts establishing the facility's usual practice
 - a strong inference about the likely outcome
 - an inference rather than an observed fate

What Does “Rare” Mean?

A HEADLINE WITHOUT THE TITLES

- 404 Media calls the tracked shipment “rare books” but does not name them.
- Earlier bulk orders included *The Insider’s Guide to Metro Denver* (1995) and *How to Use Corel WordPerfect* (1991).
- Becker called books like these dead inventory and “not necessarily rare.”
- Rarity depends on importance and demand, not scarcity alone.
- The headline can circulate while the books needed to assess it remain unidentified.

[404 Media, 17 August 2026](#) · [Charlie Becker, 10 May 2026](#) · [RBMS, “Your Old Books”](#)

404 Media calls the tracked shipment “rare books” but does not name them.

- no titles or editions available for comparison
 - no way to check library holdings or the book market
 - rarity remaining the report’s description
- the scanning operation more firmly established than the rare-book claim
 - tracking data and employee accounts supporting the operation
 - unidentified books limiting the appraisal

Earlier bulk orders included *The Insider’s Guide to Metro Denver* (1995) and *How to Use Corel WordPerfect* (1991).

- Charlie Becker’s account of unusually large orders in May 2026
 - obsolete guides that ordinarily did not sell
 - examples from earlier orders, not the AirTagged shipment
- titles making the character of those earlier orders visible
 - old and obscure rather than evidently collectable
 - no basis for assigning them to 404 Media’s later shipment
 - [Becker, 10 May 2026](#)

Becker called books like these dead inventory and “not necessarily rare.”

- limited information or surplus stock
 - books that buyers were not collecting or seeking
 - obscurity without corresponding demand

- financial arbitrage offered as an alternative explanation
 - price and availability gaps between platforms
 - a seller’s hypothesis rather than evidence about Amazon’s later facility

Rarity depends on importance and demand, not scarcity alone.

- Rare Books and Manuscripts Section guidance
 - demand exceeding supply
 - demand shaped by perceived importance
- few surviving copies insufficient by themselves
 - condition, edition, provenance and distinguishing features also mattering
 - little economic value when importance and demand are absent
 - [RBMS, “Your Old Books”](#)

The headline can circulate while the books needed to assess it remain unidentified.

- a clear and repeatable claim
 - “rare books” travelling with the Amazon and AI story
 - the underlying titles unavailable to readers
- the earlier titles supplying contrast rather than proof
 - evidence for obsolete stock in other orders
 - no evidence that those titles entered the tracked shipment

What If Someone Brings a Very Big Cheque?

OUR ASSESSMENTS ARE DATA

- University systems accumulate assignments, drafts, feedback and forum posts.
- These records show how people learn, not only what they submit.
- Weller imagines companies selling access to this corpus for AI training.
- What happens when someone arrives with a very big cheque for all of it?

[Martin Weller, “Slop Will Eat Itself \(then us\),” 17 August 2026](#)

Speaker notes

University systems accumulate assignments, drafts, feedback and forum posts.

- higher education producing new student work every teaching period
 - new cohorts, topics and responses
 - a continuing source rather than a fixed publication
- Weller's list extending beyond submitted answers
 - artefacts, variations, revisions and edits
 - detailed conversations in online forums

These records show how people learn, not only what they submit.

- revisions connecting earlier and later versions
 - changes in reasoning and expression
 - feedback becoming part of the record
- forum exchanges recording learning with other people
 - questions, explanations and disagreement
 - process becoming legible alongside outcomes

Weller imagines companies selling access to this corpus for AI training.

- student work already held by assessment platforms
 - institutions handing data to Turnitin or similar services
 - accumulation making the corpus commercially attractive
- a possible business model rather than a reported sale
 - Weller saying he does not think the companies currently sell it
 - his imagined discount to institutions in return for access

What happens when someone arrives with a very big cheque for all of it?

- assessment evidence becoming a valuable dataset
 - the same work serving educational and commercial purposes
 - new value arising without a new act by the student
- authority over the change in purpose
 - who may sell access, who benefits and who can refuse
 - **Are We Data?** returning as a question about who can act on our work
- Martin Weller, “Slop Will Eat Itself (then us),” 17 August 2026

References

umgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). *The Pushshift Reddit Dataset*. <https://doi.org/10.48550/ARXIV.2001.08435>

Illovin, S. M. (2021, June 26). *Where Did “Data Shadow” Come From?* SMBlog. <https://www.cs.columbia.edu/~smb/blog/2021-06/2021-06-26.html>

rch, K., & Cochrane, D. T. (2022). Big Tech: Four Emerging Forms of Digital Rentiership. *Science as Culture*, 31(1), 44–58. <https://doi.org/10.1080/09505431.2021.1932794>

yd, danah. (2014). *It's Complicated: The Social Lives of Networked Teens*. Yale University Press.

icher, T., & Helmond, A. (2018). The Affordances of Social Media Platforms. In J. Burgess, A. Marwick, & T. Poell (Eds.), *The SAGE Handbook of Social Media* (pp. 233–253). SAGE Publications Ltd. <https://doi.org/10.4135/9781473984066.n14>

ommon Crawl Foundation. (n.d.). *Overview*. Common Crawl. Retrieved August 15, 2026, from <https://commoncrawl.org/overview>

Haas, M. (1995, June 2). *Zlata Chat (Edited Version)*. KIDLINK. <https://web.archive.org/web/20010512082947/http://www.kidlink.org:80/KIDPROJ/WritersCorner/EVENTS/zlata.html>

ites, K. (2025). “A High-Tech Company Masquerading as a Retailer”: Target's Video Infrastructure. In *Targeted: Corporations and the Police Surveillance Economy* (pp. 52–84). New York University Press. <https://www.jstor.org/stable/jj.27300020.6>

rstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., Hughes, J., Korbak, T., Agrawal, R., Pai, D., Gromov, A., Roberts, D. A., Yang, D., Donoho, D. L., & Koyejo, S. (2024). *Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data*. <https://doi.org/10.48550/ARXIV.2404.01413>

oogle. (n.d.). *Contextual Targeting*. Google Ads Help. Retrieved August 16, 2026, from <https://support.google.com/google-ads/answer/1726458>

uminatiPirate. (2021, January 5). *Dead Internet Theory: Most of the Internet Is Fake* [Forum post]. Agora Road's Macintosh Cafe. <https://forum.agoraroad.com/index.php?threads/dead-internet-theory-most-of-the-internet-is-fake.3011/>

hnston, V. H., Ballsun-Stanton, B., Jensen, H. S., Kjelsen, C. K., & Thøgersen, J. (2026). *Exploring the Archived Web through AI Assisted Document Discovery* (Preprint bc77ead95211793cbaaebdfce200e50e6bbb365d). <https://github.com/WEB-CHILD/exploring-the-archived-web-through-ai-assisted-document-discovery> (Pre-published)

vL-. (2022). *A Metal Curse Tablet (Defixio) with a Complaint about the Theft of a Vilbia* [Graphic]. [https://commons.wikimedia.org/wiki/File:A_metal_curse_tablet_\(defixio\)_with_a_complaint_about_the_theft_of_a_Vilbia.jpg](https://commons.wikimedia.org/wiki/File:A_metal_curse_tablet_(defixio)_with_a_complaint_about_the_theft_of_a_Vilbia.jpg)

Kagi. (n.d.). *Pricing*. Kagi Search. Retrieved August 16, 2026, from <https://kagi.com/pricing>

Karpathy, A. (Director). (2023, November 23). *[1hr Talk] Intro to Large Language Models* [Video recording]. https://www.youtube.com/watch?v=zjKBMFhNj_g

KIDLINK. (1995, June 2). *Zlata Filipovic on KIDLINK IRC*. <https://web.archive.org/web/20001117135000/http://www.kidlink.org:80/resources/irczlata.html>

Klincewicz, M., Alfano, M., & Fard, A. E. (2025). Slopaganda: The interaction between propaganda and generative AI. *Filosofiska Notiser*, 12(1), 135–162. <https://www.filosofiskanotiser.com/KlincewiczAlfanoFard.pdf>

Laperdrix, P., Bielova, N., Baudry, B., & Avoine, G. (2020). Browser Fingerprinting: A Survey. *ACM Transactions on the Web*, 14(2), 8:1–8:33. <https://doi.org/10.1145/3386040>

Mehta, K. (2026, August 5). *I Didn't Set a Cookie*. Kuber Studio. <https://kuber.studio/cookie/>

Mizell, A. P., Carroll, P., & Haddad, V. (1992). *Global Communications: Kids-92* (Reports – Descriptive; Speeches/Meeting Papers No. ED368342; p. 19). <https://eric.ed.gov/?id=ED368342>

Ng, L. H. X., & Carley, K. M. (2025). A global comparison of social media bot and human characteristics. *Scientific Reports*, 15(1), 10973. <https://doi.org/10.1038/s41598-025-96372-1>

OpenRTB 2.6, Technical specification Nos. OpenRTB 2.6 (2022). <https://github.com/InteractiveAdvertisingBureau/openrtb2.x/blob/main/2.6.md>

Patel, R. (2024, February 22). *An Expanded Partnership with Reddit*. Google. <https://blog.google/company-news/inside-google/company-announcements/expanded-reddit-partnership/>

Prada, L. (2025, September 15). *One-Third of the Internet Is Just Bots Now. Seriously*. VICE. <https://www.vice.com/en/article/yep-one-third-of-the-internet-is-just-bots-now/>

r/AmitheAsshole. (2019, May 1). *AITA for Throwing Away my Boyfriend's Potentially Illegal Yogurt Collection?* [Reddit post]. https://www.reddit.com/r/AmitheAsshole/comments/bjd41e/aita_for_throwing_away_my_boyfriends_potentially/

r/AmitheAsshole. (2022a, October 20). *The Iranian yogurt is not the issue here*. [Reddit comment]. https://www.reddit.com/r/AmitheAsshole/comments/y8gzd5/aita_for_sitting_on_my_wifes_bag_after_she/it0t39y/

r/AmitheAsshole. (2022b, October 20). *What's the reference for the Iranian yogurt, I haven't seen the post. Can someone please link it* [Reddit comment]. https://www.reddit.com/r/AmitheAsshole/comments/y8gzd5/aita_for_sitting_on_my_wifes_bag_after_she/it1ghz3/



MACQUARIE
University

SYDNEY • AUSTRALIA

Rands. (2026, August 13). *RIP Claude*. Rands in Repose. <https://randsinrepose.com/archives/rip-claude>

Reddit. (n.d.). *Community and Interest Targeting*. Reddit for Business. Retrieved August 13, 2026, from <https://www.business.reddit.com/advertise/targeting/community-and-interest>

Reddit. (2024, February 22). *Expanding Our Partnership with Google*. Reddit. <https://redditinc.com/news/reddit-and-google-expand-partnership>

Reddit. (2025, May 29). *Public Content Policy*. Reddit Help. <https://support.reddithelp.com/hc/en-us/articles/26410290525844-Public-Content-Policy>

Reddit. (2026a, May 26). *How Does Reddit Search Work?* Reddit Help. <https://support.reddithelp.com/hc/en-us/articles/19695647891988-How-does-Reddit-search-work>

Reddit. (2026b, May 28). *Developer Platform and Accessing Reddit Data*. Reddit Help. <https://support.reddithelp.com/hc/en-us/articles/14945211791892-Developer-Platform-Accessing-Reddit-Data>

Reddit. (2026c, July 13). *What Is Reposting (FKA Crossposting)?* Reddit Help. <https://support.reddithelp.com/hc/en-us/articles/4835584113684-What-is-reposting-fka-crossposting>

Roman Inscriptions of Britain. (n.d.). *Tab. Sulis 4: Curse Tablet* [Inscription]. Roman Inscriptions of Britain. Retrieved August 16, 2026, from <https://romaninscriptionsofbritain.org/inscriptions/TabSulis4>

Ross, S., & Ballsun-Stanton, B. (2026, July 24). *Paper B: Reliability in research with large language models is a property of the human–AI system*. <https://doi.org/10.17605/OSF.IO/M376W>

Shan, S., Ding, W., Passananti, J., Wu, S., Zheng, H., & Zhao, B. Y. (2023). *Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models*. <https://doi.org/10.48550/ARXIV.2310.13828>

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8022), 755–759. <https://doi.org/10.1038/s41586-024-07566-y>

The Prisoner (1967): *HD Opening Titles*. (2024, March 15). [Video recording]. https://www.youtube.com/watch?v=0a1_v-BJ-Wc

Willison, S. (2024, May 29). *Training is not the same as chatting: ChatGPT and other LLMs don't remember everything you say*. Simon Willison's Weblog. <https://simonwillison.net/2024/May/29/training-not-chatting/>

AI-use disclosure

- Claude (Anthropic) and Codex (OpenAI) produced about 4.3 million output tokens between 11 and 18 August.
- About 2.7 million tokens came from main threads and 1.6 million from subagents.
- The tools supported source research, planning, drafting, production and verification.
- Brian selected the argument, sources and teaching form.

Speaker notes

Claude (Anthropic) and Codex (OpenAI) produced about 4.3 million output tokens between 11 and 18 August.

- exact total of 4,294,709 output tokens
 - 338,071 from Claude
 - 3,956,638 from Codex
- one project window from the beginning of Week 4 development to Brian’s instruction to rerun the measurement
 - 11 August 2026 at 15:37:40 AEST
 - 18 August 2026 at 10:21:29 AEST

About 2.7 million tokens came from main threads and 1.6 million from subagents.

- exact main-thread total of 2,727,708 output tokens
 - human-steered root conversations
 - Claude and Codex combined
- exact subagent total of 1,567,001 output tokens
 - replay-adjusted child-thread output
 - 36 per cent of the measured output

The tools supported source research, planning, drafting, production and verification.

- source and claim checking
 - documentary research and comparison
 - independent audits of evidence and measurements
- teaching production
 - lecture planning and prose development
 - slide composition, rendering and package checks

Brian selected the argument, sources and teaching form.

- Brian making the substantive decisions
 - claims selected and corrected
 - optional codas and their place in the lecture approved
- the lecturer retaining final responsibility
 - source boundaries adjudicated
 - rendered teaching artefacts reviewed

