

Indirect Crowdsourcing and the Hidden Flows of Data for AI

Yan Hill¹, Carlos Eduardo Barbosa^{1,2}, Matheus Argôlo¹, Lucas Nóbrega¹, Luiz Felipe Martinez¹ and Jano de Souza¹

¹Systems Engineering and Computer Science Program of the Alberto Luiz Coimbra Institute for Graduate Studies and Research in Engineering, Federal University of Rio de Janeiro, Brazil

²Naval Systems Analysis Centre, Brazilian Navy, Rio de Janeiro, Brazil

yanhill@cos.ufrj.br

eduardo@cos.ufrj.br

matheusargolo@cos.ufrj.br

lucasnobrega@cos.ufrj.br

felipem@cos.ufrj.br

jano@cos.ufrj.br

Abstract: AI systems depend on labelled datasets, yet producing high-quality labels remains costly and labour-intensive. An increasing share of “data for AI” is therefore generated by reusing the digital traces people leave as byproducts of everyday interaction. This work argues that this practice should be analysed systematically as a knowledge management (KM) phenomenon: it captures dispersed, situated know-how and codifies it into reusable knowledge assets. We formalise indirect crowdsourcing as the ex-post appropriation of human-generated digital traces into structured datasets or labels, characterised by unintentional participation, the absence of ex-ante task framing, and mediation or curation that enables reuse beyond the traces’ original purpose. We apply a unified analytic rubric—the Four Pillars of crowd, task, platform, and crowdsourcer—to three cases: Medical X, Pandemic Prediction, and reCAPTCHA. Using directed content analysis over publicly accessible artefacts and documentation, we reconstruct how traces move through a trace-to-asset lifecycle central to KM: production, selection and validation, codification and labelling, storage and access modalities, and downstream reuse. We further distinguish indirect crowdsourcing from adjacent phenomena relevant to KM and data work, including explicit crowdsourcing, human computation, and passive crowdsensing. Across cases, the analysis reveals recurring KM mechanisms (embedded versus ex-post mediation), differences in primary beneficiaries (platform owners, researchers, public agencies, etc.), and governance risks that undermine responsible knowledge reuse (opaque consent, licensing and provenance gaps, fairness concerns, and skewed value distribution). These findings highlight platforms as knowledge intermediaries that shape what can be captured, who can access it, and how value is allocated. Rather than offering a large-scale empirical evaluation of AI systems, the paper provides a conceptual and comparative analysis of how digital traces are converted into knowledge assets through indirect crowdsourcing. We contribute (i) an operational definition with observable implications for auditing indirect knowledge capture, (ii) a cross-case comparative rubric for evaluating trace-to-asset pipelines as KM processes, (iii) knowledge governance guidelines for responsible and equitable reuse, and (iv) a research agenda on boundary conditions, stewardship, and measurement for managing indirectly sourced knowledge at scale.

Keywords: Knowledge management, Knowledge governance, Digital trace data, Data provenance, Indirect crowdsourcing

1. Introduction

Artificial Intelligence (AI) has become a foundational technology across modern society, driving applications ranging from recommendation systems to complex decision-support mechanisms in healthcare, where models assist physicians with drug prescribing and risk prediction (Bortolini et al., 2024; LeCun et al., 2015). These systems depend fundamentally on data, particularly labelled datasets, which serve as the core infrastructure for training machine learning models (Sambasivan et al., 2021). However, obtaining labelled data remains a critical bottleneck, as it is costly, time-consuming, and heavily reliant on human labour (Zhou, 2017). This challenge has motivated both academia and industry to explore alternative mechanisms for large-scale data generation, with crowdsourcing emerging as a prominent approach.

The concept of crowdsourcing, introduced in 2006, has since been interpreted in multiple ways across disciplines (Hosseini et al., 2014; Howe, 2006). Despite variations in emphasis—whether on the crowd, the task, or the platform—definitions generally involve people intentionally contributing—whether for payment, recognition, or voluntary engagement—to complete specific tasks. Such explicit participation has supported large-scale labelling and problem-solving efforts, which are crucial for AI development.

Yet human contributions to AI are not always intentional. In digital environments, people generate massive volumes of data as a byproduct of everyday activities, such as communication, interaction, and authentication. When these traces are later repurposed as labelled data for training AI systems, they reveal an alternative, underexamined phenomenon—what this work terms indirect crowdsourcing.

The goal of this work is to formalise this concept and examine how large-scale, unintentional human participation can generate valuable datasets for machine learning. To do so, three illustrative cases are analysed: the Medical X (formerly Twitter) Community, Pandemic Prediction, and reCAPTCHA. Each case demonstrates a distinct configuration of how everyday online behaviour can be transformed into structured data for AI development. From a knowledge management perspective, the central issue is how dispersed experiential traces are transformed into codified, reusable knowledge assets and how they are governed across organisational boundaries. This framing is consistent with established KM literature that treats knowledge as an organisational resource created, stored, transferred, and applied through socio-technical systems (Alavi and Leidner, 2001). It also connects to the distinction between tacit and explicit knowledge, since indirect crowdsourcing often captures situated, experiential traces and later converts them into more explicit, codified artefacts that can be reused beyond their original context (Nonaka, 1994). Thus, the KM relevance of the phenomenon lies not only in knowledge sharing, but also in knowledge creation, acquisition, codification, storage, reuse, and governance.

Because the literature does not provide a consolidated definition for this specific phenomenon, we propose the following operational definition by synthesising insights from crowdsourcing, human computation, digital trace reuse, weak supervision, and passive crowdsensing. We define indirect crowdsourcing as the ex-post appropriation of human-generated digital traces into structured datasets or labels, characterized by the following jointly sufficient conditions: (C1) participation is unintentional with respect to the target task; (C2) the task is not presented to participants beforehand (no ex-ante task framing or call); (C3) mediation by a platform or third-party curation transforms traces into labeled or otherwise structured resources; (C4) the primary beneficiary is an actor different from the original trace producers; and (C5) reuse extends beyond the original purpose for which the traces were produced. Cases failing any of C1–C3 are excluded; C4–C5 distinguish indirect crowdsourcing from adjacent phenomena, such as explicit crowdsourcing or mere user-generated content reuse.

The observable implications are as follows: no recruitment message or call-to-action for the target task is found at the time of trace production; a data pipeline or curation step exists that converts traces into labeled or structured artifacts; benefits accrue primarily to platform owners, firms, or researchers rather than to trace producers; consent is typically mediated by platform Terms of Service rather than explicit task-level consent; and tasks may be embedded or mandatory within unrelated workflows, such as access control.

This work makes four main contributions. The first is an operational definition of indirect crowdsourcing with falsifiable criteria and observable implications. The second is a unified comparative rubric applied to three cases (Medical X, Pandemic Prediction, and reCAPTCHA), consolidating findings into a single cross-case matrix. The third is a set of governance guidelines for the responsible and equitable reuse of indirectly sourced data. The fourth is a research agenda that identifies boundary conditions, measurement strategies, and open problems for future work.

2. Crowdsourcing

Crowdsourcing is commonly understood as the transfer of tasks traditionally performed inside organisations to a large, undefined group through an open call (Howe, 2006). Subsequent work has shown that definitions vary across the crowd, task, platform, and crowdsourcer dimensions (Hosseini, 2014; Hosseini et al., 2014), but most accounts still assume intentional, task-framed participation. Indirect crowdsourcing differs because participants do not knowingly respond to a target task; instead, traces generated through ordinary digital activity are later curated into labelled or structured artefacts. This places the phenomenon near human computation, weak supervision, and passive crowdsensing, but with an important KM distinction: the traces are not merely sensed or shared, but codified into reusable knowledge assets.

To structure the comparison, this paper adopts Hosseini et al.'s Four Pillars model: the crowd, the crowdsourcer, the task, and the platform. The crowd dimension captures who produces the traces and whether they are specialised or general. The crowdsourcer identifies the actor that later recognises, controls, or benefits from the traces. The task describes the activity that generated the traces and the latent downstream task imposed on them. The platform mediates capture, access, and reuse. These dimensions provide the basis for comparing Medical X, Pandemic Prediction, and reCAPTCHA as cases with different degrees of intentionality, openness, and mediation.

3. Methodology

We conduct a qualitative, multiple-case comparative study. Following a directed content analysis, we utilise the Four Pillars (crowd, task, platform, and crowdsourcer) as an a priori analytic framework introduced earlier in the paper. The cases are treated as theoretical exemplars rather than as a large-N empirical sample. The purpose is not to measure the prevalence or performance of indirect crowdsourcing, but to clarify its boundary conditions and illustrate how trace-to-asset transformations operate across different settings. Our goal is to explicate boundary conditions for indirect crowdsourcing and derive governance implications.

3.1 Case Selection

We apply purposive sampling guided by our operational definition. A case qualifies as indirect crowdsourcing if and only if it satisfies all of the following necessary conditions: (C1) participation is unintentional relative to the target task; (C2) no ex ante task framing or call is presented to participants; and (C3) traces are transformed into labelled or structured artefacts via platform mediation or third-party curation. We additionally require two distinctiveness conditions: (C4) the primary beneficiary differs from trace producers; and (C5) reuse extends beyond the original purpose of trace production. We select three cases to maximise variation along domain and mediation: (i) Medical X (professional social media; research curation), (ii) Pandemic Prediction (public social media; public-health curation), and (iii) reCAPTCHA (embedded human computation; platform-owned curation).

3.2 Data Sources

We use publicly accessible documentation and artefacts for each case. For Medical X, we use platform posts and metadata accessible via official interfaces, as well as research artefacts and dataset documentation, such as curated pathology image-text pairs used to train open models. For Pandemic Prediction, we use public, aggregated digital traces, such as topic frequencies, and reports describing their use for city-level surveillance and forecasting, which are triangulated with public-health records. For reCAPTCHA, we use the service documentation and technical descriptions of embedded human-verification workflows, along with secondary analyses of downstream reuse, such as in vision tasks. Where applicable, we rely on multiple independent sources for each case and maintain an audit trail of the consulted documents.

3.3 Analytic Protocol

Our analysis proceeds in four steps: (S1) construct a structured narrative for each case (actors, traces, mediation, beneficiaries); (S2) apply a unified rubric operationalizing the Four Pillars plus cross-cutting attributes (intentionality, access modality, mediation, beneficiary); (S3) perform cross-case pattern matching to contrast mechanisms (embedded vs. ex post; public vs. proprietary curation); and (S4) synthesize governance implications linked to observed mechanisms. To make the knowledge management dimension explicit, the rubric also tracks a trace-to-asset lifecycle across cases: trace production, selection and validation, codification and labelling, storage or access modality, and downstream reuse.

3.4 Coding Procedure

Two authors independently coded each case using an a priori codebook. Disagreements were resolved through adjudication. Given the small number of cases and heterogeneous sources, we emphasise qualitative agreement and memoing rather than statistical reliability; nonetheless, we report code-level clarifications and decision rules in the analytic narrative to support replicability.

For Table 4, assignments were made using the following decision rules. “Participation mode” was coded from evidence of whether participants received an explicit task request for the downstream use. “Mediation/curation” was coded as ex-post when existing traces were later collected and structured by researchers or agencies, and as embedded when the task was incorporated into another workflow. “Access modality” was coded according to whether the resulting traces or labels were publicly accessible, available through curated datasets, aggregated, or proprietary. “Primary beneficiary” refers to the actor that controls, analyses, or directly reuses the resulting knowledge asset. “Governance flags” were assigned when the case raised visible issues of consent, provenance, privacy, representativeness, or value distribution.

3.5 Validity and Limitations

We address construct validity by grounding the rubric in an operational definition and explicit inclusion and exclusion criteria (C1–C5). Internal validity is strengthened by pattern matching across cases with different mediation modes. External validity is limited: cases are theoretical exemplars rather than a sample for statistical

generalisation. An audit trail, a shared codebook, and independent coding with adjudication support reliability. Limitations include reliance on secondary sources and the evolving nature of platform policies and APIs.

3.6 Ethical Considerations

Indirect crowdsourcing surfaces predictable failure modes—opaque consent, licensing gaps, privacy risks, and skewed value distribution. Evidence from medical social media underscores the need for stringent de-identification and context-aware sharing (Crane and Gardner, 2016); data cascades demonstrate how upstream documentation and curation choices can propagate harms in high-stakes AI (Sambasivan et al., 2021); and public-sector data governance emphasises the importance of accountability, auditability, and public-interest oversight (Belli and Doneda, 2020). In health domains, emerging risk catalogues further motivate continuous monitoring and traceability (Bortolini et al., 2024). We therefore adopt four minimal safeguards: (G1) task-level transparency whenever feasible; (G2) license- and ToS-aligned access with documented provenance; (G3) privacy-preserving aggregation and de-identification by default; and (G4) benefit-sharing mechanisms proportionate to downstream reuse, such as attribution and data access for the contributing community. All analyses use publicly accessible, aggregate, or documented artefacts. No attempt is made to deanonymize individuals. We discuss governance, consent, and responsible reuse in a dedicated section, including alignment with platform terms and applicable data-protection principles.

4. Case Studies

This section presents three illustrative cases—Medical X Community, Pandemic Prediction, and reCAPTCHA—showing how indirect crowdsourcing operates across professional exchange, public social-media activity, and embedded web infrastructure.

4.1 Medical X Community

This case study examines how professional medical communities can unintentionally generate large-scale labelled data through online interaction. In the mid-2010s, a growing number of pathologists and medical professionals began using X as a space for informal knowledge exchange and case discussion. Studies from 2014 onward have already described the emergence of these practices (Allen, 2014; Misialek and Allen, 2016) and guidelines for safely sharing pathology images on social media (Crane and Gardner, 2016). By 2017, the active presence of pathologists on X was well established, forming an online ecosystem dedicated to professional collaboration and education (Cohen et al., 2017).

Within this community, it became common for specialists to post anonymised medical images—such as microscope slides—using hashtags like #pathology, often accompanied by brief descriptions or diagnostic questions. Other professionals would respond with their interpretations, hypotheses, or final diagnoses, transforming each post into a collective reasoning exercise. Over time, this routine exchange evolved into a collaborative dynamic that unintentionally generated an extensive repository of labelled medical images and expert commentary. Figure 1 illustrates a representative pathology post, and Figure 2 shows how diagnostic interpretation unfolds in the replies.

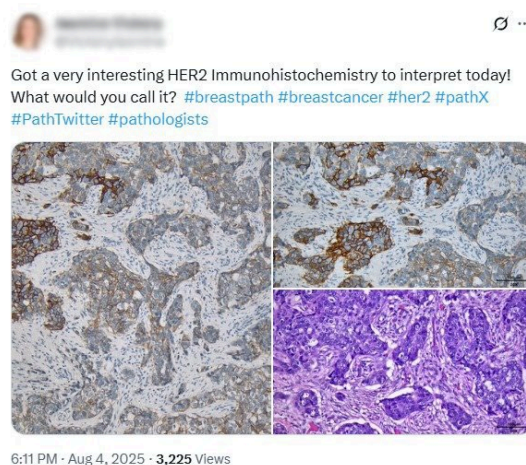


Figure 1: Example of a pathology case posted on X, where a specialist shares anonymised histopathology images and a diagnostic prompt



Figure 2: Replies to the original post showing pathologists’ diagnostic discussion

Years later, in 2023, researchers from Cambridge and Harvard identified this ecosystem as a valuable, though unintended, source of labelled data for AI development (Huang et al., 2023b). Using X’s API, they automatically collected more than 116,000 image-text pairs from pathology-related tweets and approximately 60,000 from their corresponding replies (Huang et al., 2023a). These data were then combined with other publicly available sources to form OpenPath, the largest pathology image dataset ever compiled, comprising over 208,000 labelled images. From this material, the researchers trained a visual-language model for pathology (PLIP), which achieved strong performance in a range of diagnostic tasks. Thus, Medical X illustrates specialised indirect crowdsourcing: professional knowledge exchange is retrospectively codified into image-text datasets for pathology AI.

Table 1: Medical X community case analysis

Pillar	Description
Crowd	Thousands of specialised professionals in pathology
Crowdsourcer	Independent researchers and dataset curators
Task	Sharing medical images and discussing diagnoses
Platform	X

4.2 Pandemic Prediction

This case study explores how general social media activity can unintentionally generate valuable data for public health analysis. During the COVID-19 pandemic, researchers in Brazil examined whether information shared by users on social platforms could be used to anticipate epidemic trends (Gori Maia et al., 2023). The study collected posts from several Brazilian cities mentioning COVID-related terms such as “COVID,” “mask,” and “vaccine,” aiming to assess whether online discourse could reflect real-world infection dynamics.

By correlating the frequency of these keywords with official COVID-19 case and death statistics, the researchers found a strong relationship between digital discussions and subsequent epidemiological patterns. Cities with higher volumes of pandemic-related posts tended to experience increases in reported cases and deaths in the following weeks, suggesting that social media activity could serve as an early indicator of outbreaks. Notably, the study also revealed that predictive accuracy was highest when the data were analysed at the local level, reinforcing that digital behaviour often mirrors community-specific conditions.

Although users were merely expressing personal opinions or sharing news, their collective activity became an indirect and consequential source of data for public health monitoring. This demonstrates how uncoordinated, large-scale communication can generate information of significant societal value, without the crowd’s explicit awareness or participation in a structured task. Thus, Pandemic Prediction illustrates large-scale indirect crowdsourcing: ordinary public communication is retrospectively aggregated into signals for public-health forecasting.

Table 2: Pandemic Prediction analysis

Pillar	Description
Crowd	Millions of non-specialised users are communicating on social media.
Crowdsourcer	Public health researchers and agencies
Task	Sharing information and reactions about the pandemic
Platform	X

4.3 ReCAPTCHA

This case study examines how a security tool originally designed to distinguish humans from bots has evolved into a large-scale mechanism for data labelling in AI. CAPTCHA, which stands for Completely Automated Public Turing Test to Tell Computers and Humans Apart, was developed to protect online services by presenting users with challenges that are easy for humans but difficult for automated systems (Hossain et al., 2015). Over time, this mechanism evolved into reCAPTCHA, integrating image-based tasks that require users to identify objects such as traffic lights, crosswalks, or buses. Figure 3 presents a typical object-identification challenge embedded in this workflow.

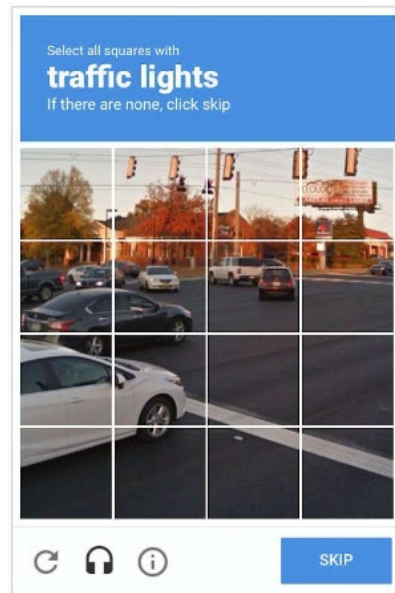


Figure 3: Example of a reCAPTCHA challenge requiring users to identify traffic lights

The reCAPTCHA system operates with a dual purpose: preventing automated access while simultaneously generating labelled data for AI training. Each time a user completes a verification challenge, their selection helps improve the accuracy of computer vision models (IBM, 2021). For instance, when millions of people select all images containing a specific object, these responses collectively form a massive dataset that can be used to train recognition algorithms. Thus, reCAPTCHA illustrates embedded indirect crowdsourcing: access-control tasks generate labels that can be reused in platform-controlled AI pipelines.

Table 3: ReCAPTCHA case analysis

Pillar	Description
Crowd	Millions of non-specialised users performing verification tasks
Crowdsourcer	Google (Alphabet Inc.)
Task	Identifying objects in images to confirm human access
Platform	Google

5. Discussion

The comparative analysis of the three case studies reveals a new pattern in the relationship between human activity and data production. While traditional crowdsourcing depends on explicit calls for participation, indirect crowdsourcing operates invisibly, transforming ordinary online behaviours into structured information that can be reused for AI development. This discussion examines how this process unfolds, the conditions that make it possible, and its broader implications for knowledge management. Read through a knowledge management lens, Table 4 also summarises a trace-to-asset pipeline in which dispersed traces are selected, validated, codified, stored or made accessible, and then reused.

Table 4: Cross-case matrix for indirect crowdsourcing using the Four Pillars and cross-cutting attributes

Attribute	Medical X	Pandemic Prediction	reCAPTCHA
Crowd size & expertise	Large, highly specialised professional community (pathologists).	Broad public, non-specialist participants.	Massive general user base; non-specialist.
Participation mode	Unintentional; no ex ante task framing or call.	Unintentional public chatter; no task presented beforehand.	Not voluntary; fully embedded in access-control workflow.
Trace source	Professional posts (image-text pairs; pathology discourse).	Public social-media signals (e.g., topic frequencies; city-level aggregation).	Human verification prompts (image/text) within access checks.
Latent/ex-post task	Pairing/labelling of pathology images and findings; dataset formation.	Syndromic signal extraction and trend forecasting.	Micro-labelling for vision/OCR tasks.
Mediation/curation	Third-party academic curation via platform interfaces; research datasets.	ex-post aggregation by public-health researchers/agencies.	Platform-owned curation by Google; embedded workflow.
Access modality	Public posts; curated research datasets (openly described).	Public posts; aggregated signals (city-level).	Proprietary labels and internal datasets.
Crowdsourcer / primary beneficiary	Academic researchers and dataset curators (e.g., OpenPath/PLIP).	Public-health researchers and agencies.	Google (Alphabet Inc.).
Reuse beyond original purpose.	Yes, training/benchmarking of pathology VLMs.	Yes, early warning and forecasting for public health.	Yes, improvements to OCR/vision and platform services.
Example downstream artifacts	OpenPath; PLIP (pathology vision-language model).	Forecast dashboards and public-health reports.	Vision/OCR improvements; non-public training corpora.
Governance flags (brief)	ToS-mediated consent; attribution/anonymisation risks for medical content.	Aggregation, representativeness, and location/privacy concerns.	Opaque consent, limited user agency, value distribution issues.

A central observation is that collective value can be generated without participant awareness or coordination. From a KM perspective, Medical X shows knowledge creation and codification, as diagnostic exchanges become reusable image-text datasets. Pandemic Prediction illustrates acquisition and combination, as social signals are aggregated with public-health records. reCAPTCHA illustrates codification and storage, as human micro-judgements are embedded into platform-controlled labelling pipelines. Across the cases, different motivations converge on the same dynamic: human activity leaves traces that can be organised into reusable data and knowledge assets.

This convergence highlights an essential property of indirect crowdsourcing: the value of participation is retrospective. Users often act without intention to contribute, and only later do they recognise and structure their output. In this sense, contribution is not a voluntary act but an interpretive process imposed on existing behaviour. The same dataset that emerges spontaneously from collective activity can thus be reframed as a resource for training AI models, policy analysis, or other secondary purposes. What defines indirect crowdsourcing is not what the crowd does, but how and when others decide that what was done has value.

The three cases also highlight the growing importance of platforms as intermediaries for human contributions. They define the boundaries of what can be captured, accessed, and reused. X, in the first two cases, acts as an open social environment where users produce data through visible interaction; Google, in reCAPTCHA, embeds participation directly into its infrastructure, making contribution invisible and automatic. Despite these contrasts, both systems operate under the same logic: they mediate between human expression and data appropriation. The degree of openness—whether public, restricted, or fully closed—determines who can benefit from the collective output and whether it remains a public or private good.

Indirect crowdsourcing, therefore, reveals a transformation in online collective intelligence. Participation no longer depends on the willingness to contribute, but on the capacity of digital systems to record and reinterpret everyday behaviour. This has two conceptual implications. First, the boundary between social activity and data production becomes porous: communication, verification, and interaction are simultaneously personal actions and potential inputs for computational models. Second, the notion of the crowd itself changes—from an

organised collective acting toward a shared goal to a distributed population whose aggregated traces acquire meaning only through external analysis.

These findings raise important ethical and governance considerations. The first concerns transparency and consent, as users often remain unaware that their actions are being reused for AI development or scientific research. The second relates to ownership and control, as the value created by collective human activity often remains concentrated in the hands of platform owners. The third is data equity: while indirect crowdsourcing can generate insights of high public relevance, such as epidemic prediction or diagnostic datasets, access to these data is unevenly distributed between private companies and public research institutions. Addressing these asymmetries will require new forms of collaboration, such as negotiated data-sharing agreements that ensure the social benefits of collectively generated data are accessible while respecting privacy and ethical standards.

Ultimately, the three cases demonstrate that indirect crowdsourcing is not an exception but a pervasive feature of the contemporary digital ecosystem. It reflects a shift from explicit collaboration to implicit participation, where collective value emerges from activities never intended as contributions. Recognising and governing this process is essential to ensure that the data produced through everyday human behaviour serves not only technological progress but also public interest and ethical accountability.

6. Conclusion

AI systems rely heavily on large amounts of labelled data, yet producing such data remains costly and labour-intensive. In this context, the rise of indirect crowdsourcing—where data is generated as byproducts of everyday digital interactions rather than intentional contributions—creates new opportunities and challenges. Understanding how collective digital behaviour is transformed into usable information is increasingly central to both technological and ethical debates in data-driven AI.

This study examined indirect crowdsourcing through a qualitative analysis of three cases: the Medical X Community, pandemic prediction models using social media data, and reCAPTCHA. Guided by the four-pillar crowdsourcing framework, the analysis identified shared structures and conditions that enable routine online activity to produce valuable datasets for AI. Although the cases differ in specialisation, coordination, and intent, all demonstrate how everyday digital behaviour can be repurposed for computational use.

The results indicate that indirect crowdsourcing is marked by unintentional participation, retrospective recognition of value, and heavy reliance on platform mediation. In all three cases, data were generated without explicit coordination, and their reuse depended on external actors or platform owners who identified and systematised the information. These findings show a shift from intentional collaboration to latent participation as a key mechanism of data production in digital environments, while underscoring the central role of platforms in determining who can access, organise, and benefit from collective human activity.

The primary contribution of this work is to formalise indirect crowdsourcing as a distinct mode of collective data generation and to demonstrate its analytical relevance across various domains. By integrating cases from health, social communication, and web infrastructure, the study provides a structured view of how uncoordinated participation can be transformed into organised knowledge assets and information flows.

This study has limitations related to scope and data selection. The analysis focused on a small number of cases, which, although representative, may not fully encompass the diversity of indirect crowdsourcing phenomena across different cultural and technological contexts. Additionally, the assessment relied on publicly available documentation and secondary sources, which limits the depth of technical or legal evaluation in each case. Therefore, the paper should be read as a conceptual and comparative contribution rather than as a comprehensive empirical mapping of indirect crowdsourcing practices.

Future research could expand this analysis by employing quantitative approaches, using larger datasets, and engaging directly with platform stakeholders to refine the understanding of how unintentional participation is transformed into structured data.

Ethics declaration: Ethical clearance was not required for the development of this research.

AI declaration: AI tools such as Grammarly and ChatGPT were used solely for language revision. The authors' analysis and interpretations are their own.

References

- Alavi, M., Leidner, D.E., 2001. Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *MIS Quarterly* 25, 107–136. <https://doi.org/10.2307/3250961>
- Allen, T.C., 2014. Social media: pathologists' force multiplier. *Arch. Pathol. Lab. Med.* 138, 1000–1001.
- Belli, L., Doneda, D.C.M., 2020. Municipal Data Governance: an analysis of Brazilian and European practices.
- Bortolini, V.S., Engelmann, W., Garcia, A. de S., 2024. Artificial Intelligence in Medicine: A systematic literature review of emerging risks and challenges. *Brazilian Journal of Law, Technology and Innovation* 2, 1–18. <https://doi.org/10.59224/bjlti.v2i2.1-18>
- Cohen, D., Allen, T.C., Balci, S., et al., 2017. #InSituPathologists: how the #USCAP2015 meeting went viral on Twitter and founded the social media movement for the United States and Canadian Academy of Pathology. *Mod. Pathol.* 30, 160–168.
- Crane, G.M., Gardner, J.M., 2016. Pathology image-sharing on social media: Recommendations for protecting privacy while motivating education. *AMA J. Ethics* 18, 817–825.
- Gori Maia, A., Martinez, J.D.M., Marteleto, L.J., Rodrigues, C.G., Sereno, L.G., 2023. Can the content of social networks explain epidemic outbreaks? *Popul. Res. Policy Rev.* 42, 9.
- Hossain, M., Nguyen, K., Rahman, M.A., 2015. Labeled-Image CAPTCHA: concept of a secured and universally useful CAPTCHA, in: 2015 UBT International Conference. University for Business and Technology, Pristina, Kosovo.
- Hosseini, M., 2014. Crowdsourcing Definitions and Its Features: An Academic Technical Report (Monograph). RCIS 2014 conference.
- Hosseini, M., Phalp, K., Taylor, J., Ali, R., 2014. The four pillars of crowdsourcing: A reference model, in: 2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS). pp. 1–12. <https://doi.org/10.1109/RCIS.2014.6861072>
- Howe, J., 2006. The Rise of Crowdsourcing. *Wired Magazine*.
- Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T., Zou, J., 2023a. Leveraging medical Twitter to build a visual-language foundation model for pathology AI. *bioRxiv*.
- Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J., 2023b. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat. Med.* 29, 2307–2316.
- IBM, 2021. What Is Data Labeling?
- LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature* 521, 436–444.
- Misialek, M.J., Allen, T.C., 2016. You're on social media! So now what? *Arch. Pathol. Lab. Med.* 140, 393.
- Nonaka, I., 1994. A Dynamic Theory of Organizational Knowledge Creation. *Organization Science* 5, 14–37. <https://doi.org/10.1287/orsc.5.1.14>
- Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., Aroyo, L.M., 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3411764.3445518>
- Zhou, Z.-H., 2017. A brief introduction to weakly supervised learning. *National Science Review* 5, 44–53. <https://doi.org/10.1093/nsr/nwx106>