AUTHOR ACCEPTED MANUSCRIPT (AAM)

Repository copy - not the publisher's Version of Record

The Data-Labeling Interface as a Security Boundary: Task-Level Privacy and Security Risks in Human-in-the-Loop ML Pipelines

Authors

Chinelo Patience Umeanozie[1]; Federico Del Giorgio Solfa[2,3]

Affiliations

[1] Doctor of Business Administration Program, University of the Cumberlands, Williamsburg, Kentucky, USA.

[2] Industrial Design Research and Development Laboratory (LIDDI), National University of La Plata, La Plata, Argentina.

[3] Laboratorio de Investigaciones del Territorio y el Ambiente (LINTA), Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC-PBA), Gonnet, Buenos Aires Province, Argentina.

Publication details

Journal: Information & Computer Security
DOI: 10.1108/ICS-05-2026-0254
Accepted: 3 October 2026
Published online: 11 November 2026

Self-archiving statement

This is the Author Accepted Manuscript (AAM) of an article published by Emerald Publishing in Information & Computer Security. The Version of Record is available via the DOI above. This AAM is deposited under the Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC 4.0). Reuse is permitted in accordance with the terms of that licence.

The following pages reproduce the accepted manuscript supplied by Emerald for self-archiving. The manuscript itself has not been altered.



The Data-Labeling Interface as a Security Boundary: Task-Level Privacy and Security Risks in Human-in-the-Loop ML Pipelines

Journal:	*Information and Computer Security*
Manuscript ID	ICS-05-2026-0254.R1
Manuscript Type:	Research Paper
Keywords:	Privacy, Security, data labeling, human-in-the-loop, least privilege, machine learning pipelines



The Data-Labeling Interface as a Security Boundary: Task-Level Privacy and Security Risks in Human-in-the-Loop ML Pipelines

Purpose

This paper investigates privacy and security vulnerabilities arising from human-in-the-loop (HITL) interactions during machine learning data labeling and quality assurance/review. It argues that the data-labeling interface constitutes a distinct and underexamined security boundary, separate from infrastructure controls and organizational compliance frameworks.

Design/methodology/approach

The study adopts a qualitative, practice-oriented research design combining thematic analysis of publicly available documentation and workflows with responses from six practitioners gathered via a structured online instrument. Findings are organized through a task-level vulnerability taxonomy and a design–exposure–control matrix.

Findings

Six recurring vulnerability families are identified: excess contextual visibility, metadata exposure, privilege expansion in QA/review, instruction-level disclosure, inference-based exposure, and weak anti-exfiltration controls. The matrix shows that many existing controls are partial rather than absent, and that governance principles such as least privilege and purpose limitation are frequently undermined at the task-interface level.

Originality/value

This study offers the first systematic treatment of the data-labeling interface as a task-level security boundary. It provides a vulnerability taxonomy, a design–exposure–control matrix, and an actionable safeguards checklist, offering practitioners and researchers structured tools to evaluate and strengthen privacy and security in HITL data-preparation environments.

Keywords: privacy; security; human-in-the-loop; data labeling; task-level vulnerabilities; least privilege; purpose limitation; data governance; machine learning pipelines; interface design; quality assurance; access control

Article type: Research paper

1. Introduction

The performance and trustworthiness of machine learning systems depend heavily on upstream work that often receives less analytical attention than models or deployment: the preparation of training data, and in particular the data labeling and quality assurance/review carried out in human-in-the-loop (HITL) settings. In practice, these tasks involve workers engaging with raw or partially processed data through interfaces, instructions, and review workflows that shape what information becomes visible, copyable, persistent, and auditable. Yet much of the privacy and security literature in ML has centered on model- or infrastructure-level threats such as membership inference and training-data extraction alongside technical mitigations like differential privacy, with comparatively limited focus on the task-level surfaces introduced by human involvement in data preparation (Dwork, 2006; Abadi et al., 2016; Shokri et al., 2017; Carlini et al., 2021).

Information and Computer Security

This emphasis matters because, during labeling and QA/review, privacy and security are not simply properties of "the system" but are shaped directly by task-level design choices. Elements such as interface layout, which fields are shown to workers, the use of masking or redaction, role-based permissions, session persistence, and the availability of tools like copy-paste or export can all facilitate accidental exposure, informal data leakage, or contextual re-identification even when organizational policies are in place (Sweeney, 2002; Narayanan and Shmatikov, 2008). At the same time, core governance principles, including purpose limitation, data minimization, and accountability, are ultimately tested in practice when data is applied to a concrete task (European Union, 2016; NIST, 2020). Similarly, the classical security principle of least privilege is meaningful only when labeling and review workflows enforce verifiable need-to-know access at the task level (Saltzer and Schroeder, 1975; NIST, 2020). Recent studies of artificial intelligence data work and annotation ecosystems further support this view by showing that data preparation is organized through broader socio-technical arrangements rather than functioning as a purely technical preprocessing stage (Muldoon et al., 2024; Wu et al., 2025).

This paper argues that the data labeling interface should be understood as a distinct and undertheorized security boundary: a socio-technical boundary that sits apart from infrastructure controls such as encryption or network segmentation and from formal compliance mechanisms like policies and contracts. It is at this boundary that design decisions concretely shape exposure, misuse, and traceability. This perspective aligns with insights from usable security and privacy, which have long shown that security mechanisms tend to fail when they do not align with real operational practices and human constraints (Whitten and Tygar, 1999; Adams and Sasse, 1999). It also draws on research in crowdsourcing and task design, which has demonstrated that how information is presented and how instructions are structured can significantly influence human performance and error patterns, with direct consequences for both quality and risk (Snow et al., 2008; Quinn and Bederson, 2011; Kittur et al., 2013).

Against this backdrop, we investigate two research questions. First (RQ1), what specific privacy and security vulnerabilities arise from HITL interactions during labeling and QA/review? And second (RQ2), to what extent do contemporary task-level design choices such as interface layout, masking and redaction, instruction design, and reviewer privileges align with or contradict the principles of least privilege and purpose limitation? To address these questions, we adopt a qualitative, practice-oriented approach that combines thematic analysis of publicly available documentation and workflows with a structured instrument administered to experts working in data operations, quality assurance, and privacy and security roles. The paper makes three contributions: (i) a task-level taxonomy of HITL vulnerabilities; (ii) a design–exposure–control matrix that maps concrete interface and workflow configurations to associated risks and potential safeguards; and (iii) an actionable checklist of task-level safeguards that complements emerging documentation and accountability practices in ML (Gebru et al., 2018; Mitchell et al., 2019).

2. Scope and Definitions

This study focuses specifically on data labeling and on quality assurance and review (QA/review) activities within machine learning data preparation. We focus on these

Information and Computer Security

activities because they involve direct human interaction with raw or partially processed training data through task interfaces, workflow configurations, and review mechanisms. The analysis does not address model evaluation in the broader sense of post-training performance assessment, except where review activities remain directly tied to labeling or annotation quality control. This boundary matters because our concern is with the privacy and security conditions created by task-level exposure during data preparation, rather than with downstream model behavior or deployment risks.

Within this scope, we examine a set of actors who may interact with task content under different permissions and workflow conditions. These include labelers or annotators, who perform the primary data classification or annotation tasks; reviewers or QA personnel, who verify, correct, or escalate completed work; vendor administrators or workflow managers, who may configure task environments and permissions; and client-side staff, where applicable, who may access outputs, oversee projects, or participate in review processes. Our threat model also considers insiders more broadly, meaning authorized users who may access information beyond operational need as well as external attackers, though the latter are not our primary analytical focus. Instead, we emphasize exposure arising from legitimate workflow participation, especially when ordinary access structures or interface features create unnecessary visibility, inference, or retention of sensitive information.

Two governance principles organize our analysis: least privilege and purpose limitation. Throughout this study, we treat these principles not as abstract policy commitments but as observable properties of task design. Least privilege means a user can access only the data, metadata, and workflow functions necessary to complete a specific task or role. At the task level, this translates into concrete considerations such as field visibility, reviewer permissions, escalation rights, export capability, and cross-task browsing. Purpose limitation refers to whether the information displayed in the interface is justified by the task's immediate purpose rather than being visible by default or for administrative convenience. At the task level, this includes contextual overexposure, metadata display, training examples, prior annotations, and other workflow artifacts that may exceed what is strictly needed for annotation or review.

By operationalizing these principles, we aim to examine privacy and security not only in terms of formal compliance but also in how access is concretely structured, displayed, and exercised in human-in-the-loop data work. This framing supports the paper's broader argument that the labeling interface functions as a distinct security boundary within contemporary ML pipelines.

3. Conceptual Framework: The Interface as a Security Boundary

This study conceptualizes the data-labeling interface as a distinct security boundary within the machine learning data-preparation pipeline. We use the term ‘security boundary’ here in an operational rather than a purely metaphorical sense. A security boundary, in this context, refers to the point at which access to information is concretely structured and exercised: what a worker can see, what actions they can perform, what context accompanies the task, and what traces remain visible or auditable. While system-level controls such as encryption, authentication, and network segmentation remain essential, they do not fully determine exposure conditions once data is made available inside a task environment. At that point,

Information and Computer Security

interface and workflow design become central mechanisms that either preserve or weaken privacy and security.

This framing builds on the classical security principle that access should be limited to what is necessary for a given function or role. In labeling environments, however, this principle is not implemented only through backend permissions; it is instantiated through concrete task-level decisions such as field visibility, metadata display, reviewer privileges, escalation pathways, export options, and the presence or absence of prior annotations. For this reason, the interface is not merely a neutral display layer. It is a governance mechanism that mediates how least privilege and purpose limitation become operational realities within human-in-the-loop workflows.

To analyze this boundary systematically, the study adopts a lightweight socio-technical threat model centered on three elements: assets, exposure surfaces, and actors. The relevant assets include raw or partially processed task data, metadata, contextual records, prior annotations, workflow history, and any sensitive or identifying information embedded in examples or review artifacts. The main exposure surfaces include task views, metadata panels, instructions, quality assurance queues, escalation channels, logs, and export or copy functions. The relevant actors include annotators, reviewers, QA personnel, workflow managers, vendor administrators, client-side staff, and insiders who may access information beyond what is necessary for their tasks. The model's focus is therefore not primarily on malicious intrusions from outside the system, but on exposure created through ordinary participation in task workflows.

A key implication of this perspective is that privacy and security should be evaluated not only in terms of whether controls exist, but also where and how they apply across workflow layers. This conceptualization also provides the basis for the study's analytical tools: a taxonomy of task-level vulnerabilities and a design–exposure–control matrix.

4. Related Work and Motivation

Research on privacy and security in machine learning has grown substantially, but much of it continues to focus on risks located at the level of models, datasets, or infrastructure. A major strand of this literature examines how trained models may leak information about their training data through membership inference, model inversion, or direct training-data extraction attacks (Shokri et al., 2017; Carlini et al., 2021). A related line of work develops technical mitigations such as differential privacy and other privacy-preserving learning methods to reduce leakage from model outputs or training procedures (Dwork, 2006; Abadi et al., 2016). This body of work is essential for understanding downstream privacy risks, yet it largely treats the training-data pipeline as a technical input to the model rather than as a socio-technical environment in which privacy and security may already be compromised before model training is complete.

A second body of work addresses privacy and security through governance frameworks, legal principles, and organizational controls. Regulatory and standards-based approaches emphasize concepts such as purpose limitation, data minimization, accountability, access control, and confidentiality as cornerstones of responsible data handling (McCallister et al., 2010; European Union, 2016; National Institute of Standards and Technology, 2020; Joint Task Force, 2020). These frameworks are highly relevant to the present study because they

Information and Computer Security

establish a normative basis for restricting access to data and aligning processing activities with legitimate purposes. However, they often operate at a relatively high level of abstraction: they specify what organizations ought to do, but do not always explain how those principles are instantiated in the concrete design of task interfaces, review queues, escalation pathways, annotation instructions, or workflow transitions.

Recent research on AI data work and annotation infrastructures helps move closer to that operational layer. Muldoon et al. (2024) propose a typology of artificial intelligence data work that highlights the diversity of human labor involved in preparing, organizing, and sustaining AI systems. Their analysis is useful here because it frames data preparation not as a purely technical preprocessing step, but as structured labor shaped by organizational arrangements and workflow design. Similarly, Wu et al. (2025) show that contemporary data annotation is embedded in broader geopolitical and economic systems, revealing how annotation practices are organized across large-scale AI data supply chains. This line of work strengthens the present study's premise that privacy and security cannot be understood only at the level of models or infrastructure; they must also be examined where data work is operationally performed.

A related contribution comes from Díaz et al. (2022), who examines crowdsourced dataset annotation as a socially situated process shaped by individual and collective identities. Their work is important because it demonstrates that annotation environments are not neutral containers for labeling tasks; rather, they are structured settings in which workers, interfaces, and organizational expectations interact to shape how data is interpreted and produced. Although their primary concern is not privacy or security, the study reinforces the idea that annotation workflows deserve closer scrutiny as socio-technical environments with consequences that extend beyond data quality alone.

The study also draws on usable security and privacy research, which argues that security mechanisms cannot be understood independently of how real users interpret, navigate, and work around them (Adams and Sasse, 1999; Whitten and Tygar, 1999). To update this perspective, the present paper also engages more recent reviews of the field. Reuter et al. (2022) characterize a quarter century of usable security and privacy research, emphasizing the importance of transparency, tailorability, and user-centered adaptation in security design. Likewise, Di Nocera et al. (2023), in a systematic literature review, show that usable security remains a central framework for understanding how security mechanisms succeed or fail in practice depending on their fit with human behavior, context, and system design. These insights are directly relevant to labeling environments, where permissions, interface visibility, metadata display, and instructional materials shape what workers are able to see and do in practice.

Taken together, these literatures reveal an important gap. Model-centered privacy research explains how information may leak from machine learning systems; governance frameworks explain the principles that should constrain data access and use; recent work on AI data labor and annotation ecosystems shows how data work is organized and operationalized; and usable security research shows that control effectiveness depends on how well protections fit actual human practices. What remains underdeveloped is an integrated account of the data-labeling interface as a task-level security boundary. In other words, existing scholarships still do not sufficiently explain how privacy and security risks

Information and Computer Security

emerge when human workers encounter raw or partially processed data through specific interfaces, role configurations, and workflow transitions.

This gap motivates the present study. Rather than asking only how data should be governed in principle, or how models may leak information after training, the paper asks how task design itself structures visibility, access, inference, and control during labeling and QA/review. By focusing on the operational layer where data becomes visible and actionable to human workers, the study seeks to connect governance principles with concrete interfaces and workflow decisions. In doing so, it contributes to a task-level perspective that complements technical privacy research, extends current discussions of AI data work, and offers a more operational bridge between governance theory and labeling practice.

5. Methodology

This study adopts a qualitative, practice-oriented research design to examine privacy and security vulnerabilities emerging at the task level in human-in-the-loop data-labeling environments. The methodological approach was developed to address a practical constraint central to this field: many labeling workflows are embedded in proprietary or confidential operational settings, making direct observation difficult or ethically complex. To avoid dependence on privileged access while still producing analytically grounded findings, the study combines two complementary sources of evidence: publicly accessible documentary materials and expert-elicited incident accounts.

The first source of evidence consists of a curated corpus of publicly available materials on contemporary data labeling and review workflows. These materials include platform documentation, task templates, training and guidance materials, workflow descriptions, and publicly accessible references to operational controls such as masking, permissions, logging, and review procedures. The purpose of this corpus is not to compare vendors as such, but rather to identify recurring task- and interface-level design features that shape what workers can see, access, infer, or retain during labeling and QA/review activities.

The second source of evidence consists of qualitative responses gathered from six professionals with direct or indirect experience in AI data labeling, annotation, or review workflows. Participants were selected through purposive sampling, which is appropriate for exploratory qualitative research where the aim is to reach individuals able to speak directly to the conditions under study rather than to construct a statistically representative sample. Eligibility was confirmed through an opening screening question, so that only respondents who reported being directly or indirectly involved in AI data labeling, annotation, or ML training-data pipelines continued to the substantive items.

The respondents spanned a range of roles across the labeling ecosystem, including data annotators, labeling reviewers and quality assurance (QA) personnel, project managers, ML engineers or data scientists, and security or compliance staff. Because exposure can arise at different points in a workflow, from annotation through review, escalation, and oversight, drawing on several roles allowed the study to capture task-level conditions from more than one vantage point. Respondents also varied in tenure, ranging from under one year to more than five years of experience, and worked across technology and AI companies, data labeling service providers, and academic or research institutions.

Information and Computer Security

Participants were reached through direct outreach by email and text message, and recruitment and response collection took place together over approximately two weeks, during which invited professionals completed the instrument at their convenience. Participation was voluntary and anonymous. No respondent was asked for personally identifying information, employer names, or client identities, and responses are reported thematically rather than attributed to individuals. Anonymity served two purposes: it protected participants, some of whom work under confidentiality obligations, and it encouraged more candid accounts of exposure than might otherwise have been provided. Informed consent was obtained before data collection, and respondents were told that the research was academic, that their responses would be anonymized, and that participation was voluntary.

The survey instrument was administered through Google Forms and contained 20 questions combining closed items (single- and multiple-select) with open-ended prompts. It was designed to move from respondent context toward specific, observable exposure conditions, opening with a short background block and then working through six thematic blocks tied to the study's analytical framework.

The background block recorded each respondent's role in the workflow, years of experience with labeling or annotation, and organizational sector. These items supported the role- and tenure-based description of the sample and helped situate each response. Each of the six thematic blocks paired a closed question about whether a given control or condition was present with an open prompt asking the respondent to describe a concrete situation in which that control fell short or produced exposure. The blocks covered six areas: metadata and contextual visibility; access controls such as masked fields, hidden metadata, restricted interface views, and access-level permissions; reviewer and QA privilege expansion; anti-exfiltration safeguards such as screenshot blocking, copy/paste restriction, watermarking, and monitoring tools; individual-level logging and monitoring gaps; and instruction-level disclosure and escalation practices. A closing pair of open questions asked respondents to identify the interface or workflow weakness they considered most consequential and to suggest one change that would reduce it.

Pairing closed and open questions in this way allowed the closed items to establish whether controls were present across respondents, while the open prompts supplied the situated incident accounts that became the main material for the taxonomy and the design–exposure–control matrix. Each returned form was reviewed individually before its content was incorporated into the analysis. Two items were near-identical in wording and were consolidated during analysis, yielding 20 distinct questions. Responses were collected anonymously over the roughly two-week data-collection period and were then analyzed thematically alongside the documentary corpus.

The analysis followed a thematic, matrix-based procedure in three linked stages. First, documentary materials and expert responses were coded to identify recurring exposure conditions, workflow features, and control gaps across labeling and QA/review settings; these observations were provisionally grouped into recurring vulnerability families. Second, each incident or documentary example was translated into a structured design–exposure–control matrix comprising design choice, exposure mode, actor, impact, existing control, recommended control, and mapped principle. Third, recurrent patterns in the matrix were interpreted through the cross-cutting principles of least privilege, purpose

Information and Computer Security

limitation, and accountability and consolidated into the task-level taxonomy, safeguards framework, and actionable checklist. Thus, the taxonomy, matrix, and checklist represent successive abstractions from the same coding and interpretive process rather than separate post hoc outputs. An initial coding framework used to guide this process is provided in Supplementary Appendix B for transparency.

The taxonomy and the design–exposure–control matrix served as the core analytical devices of the study. The full matrix developed during coding is provided in the supplementary material to enhance analytical transparency and traceability, while the main text includes a condensed version for interpretive clarity.

6. Findings I: Emerging Task-Level Vulnerability Patterns in HITL Labeling Environments

Preliminary evidence indicates that privacy and security risks in human-in-the-loop labeling environments are not limited to exceptional failures or the absence of controls. They arise through ordinary task design decisions that shape what workers can see, access, infer, retain, and act upon during labeling and QA/review. Vulnerabilities are often built into routine choices about how tasks are presented, structured, and managed.

To organize these patterns, Table I presents a task-level vulnerability taxonomy derived from the documentary corpus and the expert-elicited incident accounts. The taxonomy suggests that the relevant risk surface is not exhausted by infrastructure security or formal compliance but is constituted at the level of the task interface, workflow structure, and supporting materials. Privacy and security are shaped less by high-level policies alone than by the concrete configurations through which workers encounter data and carry out their roles.

Table I

[Table I about here]

A first category is excess contextual visibility, where annotators or reviewers are shown more surrounding content than a task requires. In several incidents, the workflow provides full conversation histories, extended document context, linked records, or adjacent information by default, even when only a limited excerpt is needed for accurate labeling. This transforms routine access into overexposure: what seems like helpful context can become a privacy risk.

A second category concerns metadata exposure. Even where the task content is relevant to annotation, visible metadata such as timestamps, identifiers, provenance fields, workflow history, or reviewer notes can substantially enlarge the privacy risk surface. This exposure is often normalized through default interface settings rather than justified by task necessity, so workers see information not because they need it but because no one hid it.

A third family is privilege expansion in QA/review workflows. Review and escalation processes often function as exceptions to the tighter visibility rules applied to frontline

Information and Computer Security

annotators, so reviewers or escalated users may gain fuller records, wider context, or broader cross-task navigation than primary annotators. What is intended as a quality mechanism can become a vector for overexposure.

A fourth category is instruction-level disclosure, where exposure is produced not by the dataset or interface but by the wording of instructions, embedded examples, and training materials. Workers encounter sensitive or contextual information in support materials that exceed the task's requirements, so the very materials meant to guide them become sources of unintended visibility.

The evidence also points to inference-based exposure, where workers reconstruct restricted information from workflow artifacts such as prior annotations, contextual traces, or patterns left by previous contributors, even without access to protected fields. The risk arises not from what is shown but from what can be pieced together.

Finally, a cross-cutting vulnerability concerns weak anti-exfiltration and monitoring controls. Once information becomes visible, restrictions on copying, retaining, exporting, or reproducing it may be weak, inconsistently applied, or dependent more on policy than enforcement, so a worker with legitimate access can take information outside the intended workflow with little friction or oversight.

Taken together, the taxonomy supports the paper's central claim that the labeling interface functions as a distinct security boundary. The following section examines how these patterns are operationally produced, mapping specific design choices to exposure modes and controls.

7. Findings II: Mapping Design Choices to Exposure Modes and Controls

While the first set of findings organized the evidence into vulnerability families, the next step is to examine how those patterns are operationally produced, moving from what vulnerabilities appear to how they arise through concrete design choices. Table II presents a condensed version of the design–exposure–control matrix, linking representative design choices to exposure modes, affected actors, impacts, and the most relevant existing or recommended controls. This shifts the analysis to a more precise question: how do interface and workflow features distribute access, under what conditions, and with what safeguards? The rows shown are representative rather than exhaustive; the extended matrix is provided in the supplementary material.

Table II

[Table II about here]

Read together with Table II, the mapping surfaces several cross-cutting patterns. Exposure frequently originates in default visibility settings, where broad access to full records, contextual fields, or ancillary metadata is built into a role or workflow layer rather than justified by task need. It is compounded where context is displayed for perceived convenience without a clear contribution to task accuracy, and where QA, review, and

Information and Computer Security

escalation roles accumulate visibility beyond that of primary annotators. The matrix also captures two routes that conventional access-control models tend to miss: instructional materials and workflow artifacts that inadvertently reveal sensitive content, and inference-based exposure, in which workers reconstruct restricted information from visible patterns or prior annotations even when protected fields are hidden. Finally, controls on post-access handling, such as copying, exporting, or capturing visible data, are consistently weaker than controls on initial visibility, so that even well-governed entry points can lead to poorly controlled exits.

Taken together, the matrix supports three broader conclusions. First, task-level risks are not incidental; they cluster around recurring design decisions that can be named, compared, and governed. Second, many controls are partial rather than absent, existing in principle but not holding up under real operational pressures. Third, the matrix exposes the gap between formal governance commitments and task realities, showing where least privilege and purpose limitation are most easily undermined.

8. Discussion

The findings suggest that privacy and security risks in human-in-the-loop labeling environments cannot be explained by system-level weaknesses alone. They emerge through the routine configuration of task interfaces, review workflows, and supporting materials that define what workers can see, infer, and do during data preparation. This shifts the focus from privacy and security as properties of infrastructure or formal compliance to privacy and security as operational outcomes of task design, supporting the view that the data-labeling interface is a distinct security boundary within the broader ML pipeline, since it structures the conditions under which access is granted and exercised (Saltzer and Schroeder, 1975).

A central implication is that established governance principles may be necessary but insufficient unless translated into concrete task-level controls. Least privilege, purpose limitation, and accountability are widely recognized in governance, yet their force depends on how they are instantiated in workflows, permissions, and interface visibility (European Parliament and Council of the European Union, 2016; Joint Task Force, 2020; National Institute of Standards and Technology, 2020). The issue is often not the absence of governance but the misalignment between principles and their implementation at the interface level.

The results also indicate that task-level controls are best understood as socio-technical rather than purely technical mechanisms. Some vulnerabilities arise from interface configuration, such as field visibility, reviewer permissions, or export functionality. Others arise from instructional design, workflow habits, or role expectations that normalize broad visibility even where technical safeguards exist. This observation is consistent with the literature on usable security, which has long shown that security failures often emerge when formal protections do not align with actual human practices and operational constraints (Adams and Sasse, 1999; Whitten and Tygar, 1999). In the present case, instruction-level disclosure and inference-based exposure are especially important because they show that privacy risk may arise not only from direct access to restricted fields, but also from

Information and Computer Security

examples, prior annotations, workflow artifacts, or contextual traces that enable workers to piece together information beyond formal access boundaries.

Another implication concerns the relationship between data protection and labeling quality. Broader visibility is often justified in the name of efficiency, interpretive accuracy, or smoother review, so excessive exposure is not simply a policy failure but sometimes reflects unresolved trade-offs between minimization and performance. Research on crowdsourcing and human computation has shown that task design, context provision, and instruction quality materially affect performance and output reliability (Snow et al., 2008; Quinn and Bederson, 2011; Kittur et al., 2013). The question is not whether workers should ever see contextual information, but which forms of context are necessary, for whom, at what stage, and under what safeguards.

The study also has implications for audit and assurance practices. If exposure conditions differ across workflow layers, as the matrix suggests in cases involving logs, exports, and QA/review, then assessments focused only on primary task interfaces may substantially understate the actual risk surface. A labeling system may appear constrained at the front end while retaining residual visibility in back-end review or monitoring artifacts. This supports treating workflow-layer consistency as a governance criterion, so that safeguards not only exist but extend coherently across annotation, escalation, review, and audit pathways. This argument also complements recent calls for stronger documentation and accountability practices in ML pipelines (Gebru et al., 2018; Mitchell et al., 2019).

Future research could extend this work through comparative studies across sectors such as health, finance, or multilingual moderation, and by testing how specific interface changes affect both privacy exposure and annotation quality.

Taken together, the discussion reinforces the paper's central claim: privacy and security in HITL data preparation are matters not only of compliance, infrastructure, or model protection, but also of task architecture. By linking design choices, exposure modes, and governance principles, the study offers a more operational perspective on how training-data protection can be evaluated and improved.

8.1 Limitations

This study has several limitations, which should be considered when interpreting its findings.

First, the analysis relies on publicly accessible documentary materials, which may present labeling processes in idealized or compliance-oriented terms and may not fully capture informal practices in real-world proprietary environments. Second, the study has limited visibility into proprietary workflows; the most consequential exposure conditions may occur in internal escalation processes or administrative dashboards that were not directly observed. Third, the expert input component is constrained by a small, non-random sample of six participants; responses offer practice-grounded insight into recurring patterns but should be understood as analytically generative rather than statistically generalizable. Finally, labeling practices vary considerably across domains, including healthcare, financial services, and content moderation, so findings may not transfer directly across contexts. The

Information and Computer Security

contribution of this study is therefore exploratory and framework-building rather than statistically representative.

Taken together, these limitations position the study as a starting point for more systematic inquiry rather than a comprehensive account.

9. Implications for Practice

The findings suggest that privacy and security in HITL labeling environments cannot be strengthened by a single control or policy statement. Because exposure emerges across multiple workflow layers, task interfaces, review queues, logs, instructions, and post-access handling, effective protection requires a layered safeguards framework. The purpose of this framework is not to eliminate all contextual access indiscriminately, but to ensure that access is justified, bounded, and traceable at each stage of the workflow. In short, the goal is to make visibility the exception rather than the default, and to ensure that when access is granted, it is both necessary and accountable.

A first layer concerns data minimization and field-level redaction or masking. Where full records, metadata, or contextual traces are visible by default, privacy risk increases even in the absence of malicious intent. Simple design choices such as showing only the fields required for a task, redacting sensitive attributes, or masking identifiers by default can substantially reduce exposure without compromising task accuracy. What is not shown cannot be leaked, inferred, or misused.

A second layer concerns fine-grained role-based access control (RBAC) and role separation. The findings show that exposure often expands in QA/review and escalation processes, where reviewers or supervisors may receive broad visibility by default. Designing roles with distinct permission profiles rather than granting reviewer privileges as a superset of annotator access helps ensure that the principle of least privilege applies across the entire workflow, not just at its entry point.

A third layer is just-in-time access and scoped tasks. Even where roles are differentiated, risk remains if access is broad, persistent, or generalized across tasks. Limiting access to specific tasks, time windows, or single sessions can reduce the likelihood that exposure extends beyond what is operationally necessary. Workers should be able to see only what is required to complete the task at hand, not everything they have ever touched or might touch in the future.

A fourth safeguard layer concerns monitoring, audit logs, and anomaly detection. The study shows that privacy and security depend not only on restricting access, but on making access traceable and reviewable. Logs that capture who viewed what, when, and for how long, combined with mechanisms to flag unusual patterns, create accountability and deter informal exfiltration. Visibility without traceability is a gap in the control surface.

A fifth layer comprises anti-exfiltration measures, including export restrictions, clipboard policies, screenshot limitations where feasible, watermarking, and secure work environments such as virtual desktop infrastructure (VDI). These controls matter because once information becomes visible, the ability to copy, capture, or remove it from the secure environment becomes a critical risk point. Strong entry controls are undermined by weak exit controls.

Information and Computer Security

A final and often underestimated safeguard concerns instructional design. Instructions should be treated as governance artifacts, not merely pedagogical support. Embedding sensitive examples in training materials, providing excessive contextual detail in guidelines, or exposing workers to unnecessary background information can transform instructional content into an exposure vector. Designing instructions with the same attention to data minimization as the task interface itself helps close this overlooked gap.

Together, these six layers form a cohesive safeguards framework that addresses vulnerabilities at the points where they emerge, not just at the level of infrastructure or policy, but at the task interface, workflow structure, and supporting materials where human-in-the-loop work happens.

Actionable Checklist

The layered framework can be translated into a practical checklist for evaluating labeling and QA/review environments. Rather than serving as a one-time audit tool, this checklist is intended to guide ongoing design and review conversations, helping practitioners ask the right questions about where exposure occurs and how it might be contained.

a) Data visibility

- Are only task-relevant fields visible by default? In other words, is everything shown actually needed for the work at hand?
- Are identifiers, metadata, and contextual traces masked or segmented where possible, rather than displayed simply because they are present in the underlying data?

b) Role permissions

- Are annotator, reviewer, QA, and administrative roles clearly separated, or do they tend to accumulate privileges as users move up the workflow?
- Does each role have access only to the information necessary for its function, or is broader access granted by default?

c) Task scope

- Is access limited to the current task or review instance, or can workers browse freely across tasks, records, or previous annotations?
- Are escalation privileges time-bounded and context-specific, or do they become standing permissions once granted?

d) Workflow consistency

Information and Computer Security

- Do protections extend consistently across annotation, QA, escalation, logs, and exports, or are there workflow layers where sensitive information remains visible despite front-end restrictions?
- Are there points in the workflow where information that was hidden at earlier stages becomes visible again?

e) Monitoring and accountability

- Are access events logged in sufficient detail to reconstruct who viewed what, when, and under what circumstances?
- Can unusual access patterns, such as repeated visibility expansions, unexpected export activity, or anomalous viewing behavior, be detected and investigated?

f) Exfiltration controls

- Are copying, exporting, screenshot capture, or retention pathways restricted where appropriate, or are they freely available by default?
- Are higher-risk tasks conducted in hardened or controlled work environments, such as virtual desktops or browser-based sandboxes that limit the ability to remove data from the secure environment?

g) Instructions and examples

- Do task materials avoid embedding unnecessary, sensitive, or identifying context in training examples, guidelines, or sample annotations?
- Are escalation routes and visibility rules clearly specified in the instructions, so workers understand not only how to perform the task but also what information they are permitted to see and under what conditions?

This checklist is designed to be adaptable: not every item will apply equally to all labeling environments, and the appropriate level of control will vary depending on data sensitivity, threat model, and operational context. However, treating these questions as part of regular workflow design rather than as afterthoughts to compliance or infrastructure planning can help close the gap between formal governance commitments and the practical realities of human-in-the-loop data work.

10. Conclusion

This study has argued that privacy and security vulnerabilities in human-in-the-loop data-labeling environments are not fully captured by system-level safeguards or compliance frameworks alone. Many of the most consequential risks emerge at the level of the task itself, through interface configurations, workflow arrangements, review structures, and instructional artifacts that shape what workers can see, infer, and do. On this basis, the paper proposed that the data-labeling interface should be treated as a distinct security

Information and Computer Security

boundary within the broader machine learning pipeline, where access is not merely authorized but concretely structured and exercised.

The analysis contributes three connected elements: a task-level taxonomy of recurring vulnerabilities, a design–exposure–control matrix linking concrete design choices to exposure modes and safeguards, and an actionable safeguards perspective showing that principles such as least privilege, purpose limitation, and accountability become meaningful only when instantiated in the ordinary architecture of tasks and workflows.

More broadly, privacy and security in ML data preparation should be evaluated not only by what systems protect in the abstract but by how access is distributed, constrained, and monitored where human work occurs. This has implications for platform design, vendor assessment, procurement, governance audits, and future research. Strengthening training-data protection requires moving beyond policy language and infrastructure controls toward the operational realities of task design, where data becomes visible, actionable, and potentially exposed.

Although exploratory in scope, the study offers a framework for systematically analyzing these realities. The taxonomy, matrix, and safeguards checklist are tools for asking better questions about how labeling environments are configured and where gaps may lie between governance intent and operational practice. Future work can build on this foundation by examining how task-level safeguards interact with organizational accountability mechanisms across different labeling contexts.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of Generative AI Use

Generative AI was used only for language refinement and copy-editing of the authors' original text. No AI tool was used to generate the study design, analysis, findings, or substantive scholarly content of this manuscript.

Ethical Statement

This study involved human participants who completed a structured online survey instrument. The research was conducted in accordance with standard ethical principles for low-risk survey research. No institutional ethics committee approval was required as the study did not involve vulnerable populations, sensitive personal data, or any risk to participants. Informed consent was obtained from all six participants prior to data collection; respondents were informed that their responses would be used solely for research purposes and agreed to participate voluntarily and anonymously.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request, subject to participant confidentiality.

Information and Computer Security

References

Abadi, M., Chu, A., Goodfellow, I.J., McMahan, H.B., Mironov, I., Talwar, K. and Zhang, L. (2016), "Deep learning with differential privacy", in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16)*, Association for Computing Machinery, pp. 308-318, doi: 10.1145/2976749.2978318.

Adams, A. and Sasse, M.A. (1999), "Users are not the enemy", *Communications of the ACM*, Vol. 42 No. 12, pp. 40-46, doi: 10.1145/322796.322806.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A. and Raffel, C. (2021), "Extracting training data from large language models", in *30th USENIX Security Symposium (USENIX Security 21)*, USENIX Association, pp. 2633-2650.

Di Nocera, F., Tempestini, G. and Orsini, M. (2023), "Usable security: a systematic literature review", *Information*, Vol. 14 No. 12, p. 641, doi: 10.3390/info14120641.

Díaz, M., Kivlichan, I., Rosen, R., Baker, D., Amironesei, R., Prabhakaran, V. and Denton, R. (2022), "Crowdworksheets: accounting for individual and collective identities underlying crowdsourced dataset annotation", in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2342-2351, doi: 10.1145/3531146.3534647.

Dwork, C. (2006), "Differential privacy", in Bugliesi, M., Preneel, B., Sassone, V. and Wegener, I. (Eds.), *Automata, Languages and Programming*, Springer, pp. 1-12, doi: 10.1007/11787006_1.

European Parliament and Council of the European Union (2016), Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation), Official Journal of the European Union, L 119, available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H. and Crawford, K. (2018), "Datasheets for datasets", arXiv preprint arXiv:1803.09010, doi: 10.48550/arXiv.1803.09010.

Joint Task Force (2020), Security and Privacy Controls for Information Systems and Organizations, NIST Special Publication 800-53 Revision 5, National Institute of Standards and Technology, doi: 10.6028/NIST.SP.800-53r5.

Kittur, A., Nickerson, J.V., Bernstein, M.S., Gerber, E.M., Shaw, A., Zimmerman, J., Lease, M. and Horton, J. (2013), "The future of crowd work", in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*, Association for Computing Machinery, pp. 1301-1318, doi: 10.1145/2441776.2441923.

McCallister, E., Grance, T. and Scarfone, K.A. (2010), *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII), NIST Special Publication 800-122*, National Institute of Standards and Technology, doi: 10.6028/NIST.SP.800-122.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T. (2019), "Model cards for model reporting", in *Proceedings of the*

Information and Computer Security

Conference on Fairness, Accountability, and Transparency (FAccT '19), Association for Computing Machinery, pp. 220-229, doi: 10.1145/3287560.3287596.

Muldoon, J., Cant, C., Wu, B. and Graham, M. (2024), "A typology of artificial intelligence data work", *Big Data & Society*, Vol. 11 No. 1, doi: 10.1177/20539517241232632.

Narayanan, A. and Shmatikov, V. (2008), "Robust de-anonymization of large sparse datasets", in *2008 IEEE Symposium on Security and Privacy*, IEEE, pp. 111-125, doi: 10.1109/SP.2008.33.

National Institute of Standards and Technology (2020), NIST Privacy Framework: a Tool for Improving Privacy through Enterprise Risk Management, Version 1.0, NIST CSWP 01162020, doi: 10.6028/NIST.CSWP.01162020.

Quinn, A.J. and Bederson, B.B. (2011), "Human computation: a survey and taxonomy of a growing field", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*, Association for Computing Machinery, doi: 10.1145/1978942.1979148.

Reuter, C., Iacono, L.L. and Benlian, A. (2022), "A quarter century of usable security and privacy research: transparency, tailorability, and the road ahead", *Behaviour & Information Technology*, Vol. 41 No. 10, pp. 2035-2048, doi: 10.1080/0144929X.2022.2080908.

Saltzer, J.H. and Schroeder, M.D. (1975), "The protection of information in computer systems", *Proceedings of the IEEE*, Vol. 63 No. 9, pp. 1278-1308, doi: 10.1109/PROC.1975.9939.

Shokri, R., Stronati, M., Song, C. and Shmatikov, V. (2017), "Membership inference attacks against machine learning models", in *2017 IEEE Symposium on Security and Privacy (SP)*, IEEE, pp. 3-18, doi: 10.1109/SP.2017.41.

Snow, R., O'Connor, B., Jurafsky, D. and Ng, A.Y. (2008), "Cheap and fast but is it good? Evaluating non-expert annotations for natural language tasks", in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, Association for Computational Linguistics, pp. 254-263.

Sweeney, L. (2002), "k-anonymity: a model for protecting privacy", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10 No. 5, pp. 557-570, doi: 10.1142/S0218488502001648.

Whitten, A. and Tygar, J.D. (1999), "Why Johnny can't encrypt: a usability evaluation of PGP 5.0", in *8th USENIX Security Symposium (USENIX Security 99)*, USENIX Association.

Wu, T., Muldoon, J. and Xia, B. (2025), "Global data empires: analysing artificial intelligence data annotation in China and the USA", *Big Data & Society*, Vol. 12 No. 2, doi: 10.1177/20539517251340600.

Supplementary Material

Information and Computer Security

Appendix A (Extended Design–Exposure–Control Matrix) and Appendix B (Initial Coding Framework Used in the Thematic Analysis) are provided as supplementary material submitted alongside this manuscript in a single supplementary file.

Information and Computer Security

Tables

Table I

Task-Level Vulnerability Taxonomy in Human-in-the-Loop Labeling Environments

Vulnerability family	Definition	Typical exposure pattern	Primary governance principle
Excess contextual visibility	Workers are shown more surrounding content or contextual information than is necessary to complete a task.	Full conversation history, extended document context, adjacent records, linked content.	Purpose limitation
Metadata exposure	Metadata or quasi-identifying information remains visible even when not required for annotation or review.	Timestamps, internal IDs, provenance fields, workflow history, reviewer notes.	Purpose limitation
Privilege expansion in QA/review	Review, escalation, or supervisory roles receive broader visibility than is operationally necessary.	Full-record access during review, cross-task browsing, wider access in escalation workflows.	Least privilege
Instruction-level disclosure	Instructions, examples, or training materials reveal unnecessary, sensitive, or contextual information.	Realistic examples with personal details, embedded sensitive cases, overly rich guidance materials.	Purpose limitation
Inference-based exposure	Workers infer restricted or sensitive information from workflow artifacts rather than direct field access.	Prior annotations, visible traces of previous contributors, contextual cues, workflow artifacts.	Purpose limitation
Weak anti-exfiltration and monitoring controls	Visible information can be copied, retained, reproduced, or accessed without sufficient enforcement or traceability.	Manual transcription, screenshots, export pathways, incomplete logging, weak activity monitoring.	Accountability / Least privilege

Note. The taxonomy is derived from thematic analysis of publicly available documentation and expert-elicited incident accounts.

Table II

Representative Mapping of Design Choices, Exposure Modes, Controls, and Governance Principles

Design choice	Exposure mode	Actor	Impact	Existing/recommended control	Principle
Full conversation or document history shown in labeling interface	Over-visibility of contextual data	Annotator/Reviewer	Unnecessary exposure of sensitive content beyond task need	Context segmentation, least-visibility interface, field-level masking	Purpose limitation

Information and Computer Security

Review tasks include prompts, system metadata, or internal identifiers	Contextual re-identification	Reviewer	Exposure of timestamps, identifiers, or partial user history not needed for verification	Field-level masking, contextual filtering, scoped review visibility	Least privilege / Purpose limitation
Metadata remains visible in task logs or embedded interface elements	Indirect exposure through interface design	Reviewer	Visibility of identifiers or timestamps not required for evaluation	Metadata suppression in task view, log isolation, redaction across workflow layers	Purpose limitation
QA queue exposes full task history	Privilege expansion in review workflow	QA reviewer	Broader access to user-generated content than primary annotators	Tiered RBAC, task-scoped QA review views	Least privilege
Escalated tasks become visible to additional reviewers or managers	Exposure expansion during escalation	QA/Project leads	Sensitive content becomes visible to a wider set of actors	Controlled escalation workflow, masked identifiers, time-bounded access	Least privilege
Training instructions contain realistic sensitive examples	Instruction-level disclosure	Annotator	Exposure through supporting materials rather than core task content	Sanitized examples, instructional review, task-specific guidance only	Purpose limitation
Metadata displayed next to task content	Misinterpretation and identity inference	Annotator	Metadata becomes an implicit labeling cue or enables inference about identity or context	Conditional metadata display, suppression of non-essential fields	Purpose limitation
Copy/paste restrictions exist but manual transcription remains possible	Informal exfiltration	Annotator	Sensitive information can still be recreated outside the system	Secure workspace controls, watermarking, VDI in higher-risk tasks, anomaly monitoring	Accountability
Monitoring tracks completion but not detailed user behavior	Monitoring gap	Annotator/Reviewer	Screenshots, note-taking, or unusual access patterns remain undetected	Activity-level audit logging, anomaly detection, secure workspace policies	Accountability
Prior annotations or workflow traces remain	Inference-based exposure	Reviewer	Reviewers infer earlier decisions or contextual information	Anonymization of prior annotations, clearer separation between task stages	Purpose limitation

Information and Computer Security

visible in review			beyond intended scope		

Note. Table II reports only representative rows; the extended analytical matrix is provided in Supplementary Material Appendix 1.docx.

Information and Computer Security