Measuring Proof Burden in Public Bounty Listings: A RentAHuman Case Study

Iman YeckehZaare
MIT Center for Collective Intelligence (CCI)
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
Research
Honor Education
San Francisco, California, USA
oneman@mit.edu

Abstract

Online bounty markets let requesters advertise paid tasks to workers. A worker may be asked not only to complete a task but also to prove that it was completed, and proving can mean exposure: revealing identity or location, using a personal account, posting publicly, acting in the physical world, or supplying evidence again during later checks—none disclosed by the posted price. We call these advertised requirements *proof burden* and measure them on RentAHuman, a 2026 market publicized as a place where AI agents could hire humans. We study what listings request, not what workers submit or experience.

We manually audited a nonrandom May 31, 2026 snapshot: every listing our searches returned from RentAHuman and Human Pages, another such market—981 listings, all but one from RentAHuman. Two independent coders recorded 13 features—11 kinds of evidence, recurring monitoring (repeated checks), and physical-world action—and our 0–5 Proof Burden Score; a blinded third coder resolved every disagreement. A planned content screen leaves 779 bounty/task listings as the primary population; among them, physical-world action appears in 48.5%, photo proof in 37.1%, and identity proof in 30.9%, and 438 listings (56.2%) score 4 or 5. Those listings span 154 distinct feature combinations: a checklist, not a single score, tells workers what a listing entails.

Platform metadata labels some requester accounts as agents or bots. In exploratory comparisons, physical-world action, location proof, or recurring monitoring appeared in 75.0% of agent-or-bot-labeled versus 55.3% of human-labeled listings, while their score-4-or-5 shares did not clearly differ. The labels are self-reported or platform-assigned, the agent-or-bot-labeled listings come from only 20 displayed names, and the comparison was chosen after seeing the data: a hypothesis, not a confirmed difference. Our main contributions are the 13-requirement vocabulary, the adjudicated manual audit, and the descriptive case study of this market; the score is a secondary screening summary, and the requester-label comparison exploratory. The study offers no worker-validated measure or automated detector yet; it is groundwork for both.

HCOMP 2026, Alexandria, VA, USA

ACM ISBN 979-8-4007-2894-5/26/09
https://doi.org/10.1145/3834580.3838744

CCS Concepts

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Information systems** → *Crowdsourcing*; • **Social and professional topics** → *Privacy policies.*

Keywords

proof burden, RentAHuman, bounty markets, crowd work, surveillance, physical-world action, worker privacy, manual audit

ACM Reference Format:
Iman YeckehZaare. 2026. Measuring Proof Burden in Public Bounty Listings: A RentAHuman Case Study. In *2026 ACM Conference on Human-AI Complementarity and Alignment (HCOMP 2026), September 27–30, 2026, Alexandria, VA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3834580.3838744

1 Introduction

RentAHuman drew press attention in 2026 for a provocative premise: AI agents could hire people to perform physical and online tasks [3, 45]. Its public bounty listings—paid tasks advertised by requesters—also state what counts as completion and how workers should prove it. Throughout, a "bounty" is such a public advertisement and its visible metadata, not a completed transaction. Proof can be simple, such as a written acknowledgement. It can also require a worker to reveal identity or location, post from a personal account, or travel somewhere solely to photograph the result of an otherwise remote task. In those cases, proving the work can itself become a source of exposure. A posted price advertises an amount or rate; it does not say everything a worker may have to reveal, record, do in public, or provide again later.

Research on online-work verification commonly asks whether requesters can judge the quality of submitted work [16]. We ask a complementary, worker-facing question: what does a listing ask a worker to submit, reveal, use, or do so that completion can be checked? We call this broader set of completion-linked exposure requirements *proof burden*. It covers requested evidence, recurring monitoring (work, availability, or evidence requested on separate occasions), and potentially exposing physical-world action. We observe listings, not acceptance, submissions, payment, rejection, or worker experience.

We contribute a vocabulary of 13 concrete requirements, a manual audit, and a descriptive case study of 779 public listings. Naming

the requirements makes free-text conditions countable and comparable across listings. We also designed the 0–5 Proof Burden Score (PBS) as a secondary screening summary. The score compresses detail: even among listings scoring 4 or 5, we observed 154 different combinations of requirements. The 13-requirement checklist therefore preserves distinctions the score hides. After examining the data, we also found that listings whose platform metadata linked the requester to an agent or bot more often requested at least one of physical-world action, location proof, or recurring monitoring than listings labeled "human."

Recent work examines harmful tasks, automated posting, software-mediated rule setting, and AI-managed work [15, 22, 28, 39, 40]. We address the narrower question of what public listings ask workers to reveal or do when documenting completion. A checklist could make overlooked requirements easier to compare.

We ask three questions of the May 2026 snapshot (Section 3):

RQ1 Which kinds of completion evidence, recurring monitoring, and physical-world action do listings request?

RQ2 What shares of listings reach the proposed elevated (score 3 or higher) and severe (score 4 or higher) thresholds, and which combinations of requirements appear among listings scoring 4 or 5?

RQ3 (exploratory) How do these requirements differ between listings the platforms label "agent or bot" and those they label "human"?

2 Background and Related Work

Four strands of prior work inform proof burden: verifying crowd work, the hidden costs workers bear, monitoring and privacy on labor platforms, and theories of privacy and administrative burden. Human-computation research studies how to check work by comparing answers, estimating reliability, and balancing cost and accuracy [14, 16]; requesters may also ask for completion evidence. Documentation frameworks make data-labeling decisions visible [10]. Other work connects task design, reputation, and governance to crowd work [20]. Recent studies examine risks to workers and the information they want disclosed to them [32, 33].

Worker-centered tools expose requester conduct and help workers act collectively [17, 35]. Such visibility matters because a posted price omits unpaid time spent finding, waiting for, and vetting tasks, and the risk of rejection [13, 42]; unclear submission criteria add further risk [27]. "Ghost work" extends this to human labor hidden behind apparently automated systems [12].

Research on online-platform work shows how task definitions and algorithms redistribute control and risk among workers, requesters, and platforms [2, 34, 43]. Workers weigh pay against privacy, trust, sensitivity, rejection risk, and time before sharing data, and task design can bound sensitive input [19, 36, 47]. Gig workers report uncertainty about what platforms collect, and workers exposed to more algorithmic monitoring or data collection report less trust and greater privacy concern than workers exposed to less [37, 44]. Platform audits and experiments document privacy risks and show that workers may avoid tasks requiring tracking [23, 31]. Beyond monitoring, physical-world tasks add travel and location constraints [1, 41]. Research on workers recording their own working conditions ("sousveillance") and on human–AI systems identifies surveillance, time, privacy, and required availability as design concerns [11, 18]. These studies motivate our focus on advertised requirements, as does scholarship on surveillance economies [48].

Our 13 features include using an account or disclosing information; proving location, identity, or a purchase; acting publicly; and staying available for later checks. Theory suggests why they matter. Solove treats privacy harms as distinct activities—identification, exposure, access, and monitoring [38]; contextual integrity makes acceptability depend on who shares information, why, and under what conditions [30]; administrative-burden research separates learning and compliance costs from psychological strain [29]. We observe advertised actions, not those harms, norm violations, or costs; the supplement develops each mapping.

3 Data and Ethics

We collected public listings during one window on May 31, 2026, U.S. Eastern time (June 1 UTC), through pages requiring no login. We did not apply to tasks, contact anyone, or access nonpublic channels. Searches returned 981 records: 980 RentAHuman bounties and one listing from Human Pages, a second bounty market we searched. A third site, GoHireHumans, listed no jobs. The imbalance reflects market inventory, not search coverage: both smaller markets were nearly empty at collection, and an August 2026 recheck still found no open Human Pages listings and only four GoHireHumans postings. The returned-record count excludes duplicate platform IDs; we froze the set both coders received, although the live market kept changing. The content screen in our internal pre-analysis plan removes promotional posts, service offers, and records requesting no human action, leaving 779 bounty/task listings—records advertising a task for a worker to complete—as the primary population, which we call the *eligible* listings. Every returned record was coded in full, so screened-out records still carry proof labels, and all-returned results are reported separately as a sensitivity analysis.

Because all but one returned listing came from RentAHuman, claims primarily concern that market. Public fields show price, status, category, application and position counts, and free text (title, description, general-requirements field, and proof-specific fields), but not work, payment, rejection, or worker decisions.

At collection, the listings we retrieved carried one of five public statuses: cancelled, completed, open, assigned, or partially filled. In the platforms' requester-type metadata, 676 eligible listings carry a human label, and 71 eligible RentAHuman listings carry a label naming an agent or bot as the requester. To mark potentially related listings, we grouped RentAHuman listings whose lowercased, trimmed requester display names matched. We call each group a display-name cluster; matching names do not prove a shared requester. The Human Pages listing's metadata links its requester to an agent, and its unique public name forms a one-listing display-name cluster. Together, that listing and those 71 RentAHuman listings form the 72 agent-or-bot-labeled listings. The remaining 31 listings have the ambiguous label "other." Requester labels come from users or platforms and cannot show who designed, benefited from, supervised, or verified a task.

Some public searches failed, so we merged separate searches by status and category; every query filtered to the "recruiting"

category still failed, and a broad search's first page showed one recruiting listing, so more may be missing. The snapshot is neither random nor exhaustive. We treat already-cancelled listings (531 of 981 returned; 404 of 779 eligible) as advertised requests, not failed or rejected work. The supplement documents the collection, a ten-day public-visibility check, and score shares by status.

Because raw task text may contain personal data or risky instructions, we do not release it. Restricted local coder and adjudicator packets contain task wording. Separate processed analysis files omit that wording but retain labels and one-way transformations of platform IDs that could still be linked to public listings; they also remain private. Paper examples are paraphrases rather than quotations. The institutional review board (IRB) determined that this analysis of public listings was not human-subjects research, so IRB review was not required. This determination is neither IRB approval nor an exemption.

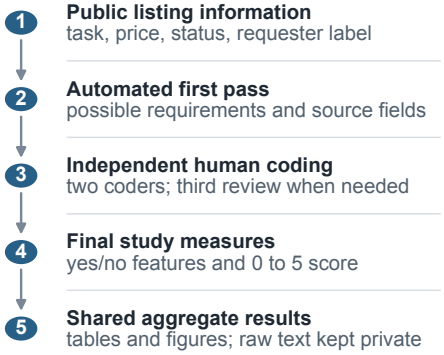

Figure 1: From public fields to aggregate results. Coders did not see automated candidate matches. Two coders reviewed each listing; an adjudicator (a third coder) reviewed disagreements and referrals. Routing was deliberately broad: the adjudicator decided 873 of the 981 listings; exact coder consensus decided the other 108.

4 How We Measure Proof Burden

Independent human coding with adjudication by a third coder (Figure 1) produced 13 yes-or-no labels and a PBS score per listing; a separate automated first pass serves only as a check. The subsections define the 13 requirements, describe rule-assisted extraction and human review, specify the score, and report coding agreement.

4.1 Requirements We Record

We recorded 13 yes-or-no features: 11 kinds of completion evidence, recurring monitoring, and physical-world action. The evidence (*proof*) features are written text, a link, screenshot, photo, video, identity (a marker identifying the individual worker), account (signing in to, verifying, or posting under an external account), location, phone (a call or text), financial (a purchase, receipt, or balance), and public-post evidence (an authored artifact visible to third parties). Recurring monitoring means that a listing asks for work, availability, or evidence on separate occasions. Physical-world action marks listings that require acting in the physical world; it captures a potentially exposing condition even when the action itself is not submitted as evidence. A listing may involve a place, device, camera, account, or purchase without requesting evidence of it. Location proof, by contrast, is evidence of where the worker was; a listing can require physical-world action without it, and rarely the reverse.

Coders read all supplied listing fields together and marked a feature only when it was tied to completion, rather than treating a keyword occurrence as sufficient. For example, they distinguished an incidental place name from a request for location proof. The coding did not distinguish whether a requirement was the task itself, a prerequisite for accessing it, or completion evidence; the Limitations section describes what this rules out.

4.2 Rule-Assisted Extraction, Human Review

An automated extractor—deterministic, hand-written keyword and regular-expression rules; no machine learning or generative AI—logged each candidate match from a platform-provided evidence label, proof-specific field, or action phrase (a proof term near a verb such as *submit* or *upload*) in the title or description. Category, skill, work-mode, and location metadata alone could not trigger an evidence-feature label. The supplement describes the rules.

All counts and models use final human-reviewed labels. Neither coder nor the adjudicator saw automated matches, which we use only for diagnostic checks. The score implied by the automated matches differed from the final score for 373 eligible listings; human review both added and removed feature labels (Table 1).

4.3 Proof Burden Score

PBS is a 0–5 screening score we designed (Table 2); it summarizes the requirements a listing states rather than estimating a hidden quantity behind them, so a higher score means more requests, or more exposing ones, not more of one underlying thing [9]. We call scores of at least 3 *elevated* and at least 4 *severe*; both are bands on our own scale, not levels of demonstrated harm. PBS ranks listings and reports shares above a threshold; the checklist describes an individual listing. Scores are ordered summaries, not equal-interval measurements, so analyses that treat consecutive values as equally spaced are approximations. Coders first chose a *base tier* (0–4) from Table 2; each tier names kinds of requested proof or action, and the highest matching tier applies. Coders then checked three fixed *modifier* conditions against the same coded requirements (lower panel of Table 2); each adds a point: more than one evidence type; physical-world action combined with proof or monitoring; and public posting or account use. The final score is the base tier plus the modifier points, capped at 5—for example, a listing requesting only photo proof starts at tier 2, and one modifier point for added physical-world action raises its score to 3. At tier boundaries coders could use judgment.

Coders recorded the base tier, each modifier, the uncapped total, and the final score directly, so the recorded components are human-coded, not reconstructed from the labels. Three listings request phone contact as their only base-setting feature; the rubric had no tier for that case, and they carry a prospectively approved tier-3 base, labeled as a new policy rather than a recovered one.

We, not workers or a statistical model, selected the base tiers, modifier points, cap, and thresholds. Public-post and account evidence can affect both the base tier and a modifier point. Removing

Table 1: Automated-rule matches versus final human-reviewed labels (779 eligible listings). "Added" counts absent-to-present changes, "Removed" the reverse, and "Net" the change in listings marked present. The upper panel lists the 11 kinds of evidence; the lower panel, physical-world action and recurring monitoring (repeated checks).

Feature	Rule match	Final label	Added	Removed	Net
Text proof	298	450	195	43	+152
Photo proof	387	289	31	129	−98
Identity proof	44	241	216	19	+197
Link proof	278	231	12	59	−47
Video proof	213	207	34	40	−6
Account proof	183	190	49	42	+7
Screenshot proof	122	158	40	4	+36
Public-post proof	210	123	23	110	−87
Location proof	169	83	34	120	−86
Financial proof	122	80	18	60	−42
Phone proof	76	43	12	45	−33
Physical-world action	425	378	67	114	−47
Recurring monitoring	33	151	130	12	+118

Table 2: PBS rubric and score counts (779 eligible listings). "Coder-recorded base *n*": listings whose coders placed them at that base tier. "Final *n*" and "Final share": listings ending at that score after modifier points and the cap. Base tiers stop at 4; score 5 is reached only by adding modifier points.

Tier	Meaning	Coder-recorded base *n*	Final *n*	Final share
0	No specified proof, or a minimal text acknowledgement only.	263	102	13.1%
1	Private text about the work: content specified, described, or itemized, however detailed; or a private link to such text.	32	88	11.3%
2	Screenshot, photo, or other evidence of the requested work product.	80	97	12.5%
3	Video proof, public-post proof, account proof, location proof, or financial proof.	94	54	6.9%
4	Identity proof; recurring monitoring; or physical-world action paired with (1) phone proof, (2) public posting or account use, or (3) location proof plus photo, video, or financial proof.	310	66	8.5%
5	Capped maximum: base tier plus modifier points reaches at least 5; examples: multiple proofs, physical-world action plus proof/monitoring, or public/account use.	–	372	47.8%
Modifier	**Condition (each met adds one point)**	**Listings *n***		
+1	More than one of the 11 evidence types is requested.	551	–	–
+1	Physical-world action is combined with at least one proof or monitoring requirement.	360	–	–
+1	Public posting or account use is required.	200	–	–

that modifier point moves 15 of the 779 eligible listings below the severe threshold. The supplement tests other scoring choices, which affect PBS only; RQ1 and RQ3's feature comparisons use the yes-or-no features directly.

4.4 Coding Agreement and Limits

The final labels come from a complete second coding round that replaced the first round; the Limitations section describes the differences. Two newly recruited coders first piloted the revised codebook—the pilot surfaced definitions the original codebook lacked, several flagged by the coders—and then, working independently with complete listing context, labeled all 981 listings. A blinded adjudicator, seeing both coders' answers but no superseded labels, resolved 873 listings, each routed for a coder disagreement, a coder's ambiguity/sensitivity flag, an *UNCLEAR* (the codebook's explicit uncertainty answer), or a difference from the superseded labels; for the other 108, exact consensus became final. All three were paid $37 per hour on average; the supplement details recruitment. Agreement uses the standard two-coder statistic κ (0 = agreement no better than chance, 1 = perfect agreement); weighted κ gives partial credit when the two coders' scores are close [6, 7]. On the 0–5 score, coders agreed exactly for 81.8% of the 981 listings (weighted $\kappa = .901$) and on whether it was at least 4 for 90.2% ($\kappa = .804$). Reliability improved on twelve of the thirteen features over the first round and held on the thirteenth (link proof, $\kappa = .94$); the weakest are now recurring monitoring ($\kappa = .76$) and location ($\kappa = .79$). Disagreements were most frequent for text proof (82 listings), recurring monitoring (62), and identity (61), and rarest for screenshot, video, and photo (1, 7, and 13). On contested labels the adjudicator sided with each coder almost equally (55% versus 45%), so the final labels do not simply mirror either coder.

5 Descriptive Results

The three subsections mirror the research questions: feature prevalence (RQ1), the score distribution and high-score combinations (RQ2), and the exploratory comparison by requester label (RQ3).

5.1 Prevalence of the 13 Features

Under the final human-reviewed labels (RQ1), 450 eligible listings (57.8%) asked for written confirmation, which we code as text proof. Physical-world action appeared in 378 listings (48.5%). Other common requirements were photo proof in 289 listings (37.1%), identity proof in 241 (30.9%), link proof in 231 (29.7%), video proof in 207 (26.6%), and account proof in 190 (24.4%). Less common were screenshot proof (20.3%), recurring monitoring (19.4%), public-post proof (15.8%), location proof (10.7%), financial proof (10.3%), and phone proof (5.5%). Location is far rarer than the 25.9% the first coding round reported over 981 records; that superseded round had conflated location with physical-world action. Figure 2 shows all 13 features.

5.2 Score Distribution and High-Score Groups

For RQ2, 492 listings (63.2%) score at least 3 and 438 (56.2%) score at least 4 (severe). Because listings sharing a displayed requester name may be related, we report the severe share's precision two ways. If every listing were unrelated to every other listing, a 95% Wilson interval for the score-4-or-higher share would be 52.7–59.7% [46]. Here an interval is the range of values statistically consistent with the observed counts; "Wilson" is a standard method. Grouping listings by the 468 RentAHuman names and the Human Pages name into 469 display-name clusters gives a wider 95% interval, 51.5–60.9%, from an intercept-only logistic model of the severe indicator with cluster-robust standard errors (supplement). Both are *reference* intervals: they differ only in assumed dependence, and neither generalizes beyond this nonrandom snapshot.

Scores skew high (Table 2): 102 listings (13.1%) score 0 and 372 (47.8%) score 5. The cap hides differences among the 372 listings scored 5: for 222 of them (59.7%), the recorded base tier plus modifier points exceeds 5.

More importantly, the 438 listings scoring 4 or 5 contain 154 different combinations of the 13 features; the ten most common combinations cover only 155 listings (35.4% of them). No single profile dominates. To summarize this variety, Table 3 assigns each listing to the first qualifying group in a fixed order; the groups are descriptive and depend on it.

The largest group is identity-linked proof, named after identity proof (199 listings, 45.4% of listings scoring 4 or 5), followed by physical-world action plus proof (102) and recurring monitoring (101). Here "physical-world action plus proof" means physical-world action together with photo, video, location, financial, public-post, or account proof, among listings not already assigned to an earlier group.

Coder agreement on the feature each group is named after ranges from $\kappa = .76$ (recurring monitoring) to .98 (video proof; Table 3). Before adjudication, the coders classified 59.2% and 58.5% of eligible listings as severe, versus the final 56.2%. Thus the overall share changes little, but membership in the feature-defined groups still depends on the adjudicator's decisions and on features with moderate agreement. The supplement repeats the analysis with narrower definitions and with the superseded round's pre-adjudication consensus labels.

Within RQ2, breakdowns among the twelve largest categories are descriptive: score-4-or-5 shares are highest in delivery errands, events/social, marketing campaigns, and hiring, and are also high in research fieldwork and home/personal. Documentation and writing content have the largest score-0 shares (Figure 3).

5.3 Exploratory Comparison by Platform Requester Label

For RQ3, we compare 72 listings whose platform metadata carries an agent-or-bot requester type (RentAHuman) or links the requester to an agent (Human Pages; Section 3) against 676 listings carrying a human label. We exclude 31 rows labeled "other." We use "agent-or-bot-labeled" as shorthand for the first group, but this metadata does not establish who designed or controlled a task.

Analyses again group listings into the display-name clusters defined in Section 3; dropping the "other" rows leaves 462 of them. Some clusters are *mixed*: the same name appears with both requester-type labels. We examine the full sample and three sensitivity analyses: removing the *agent-heaviest* mixed cluster (the one with the most agent-or-bot-labeled listings), removing all mixed clusters, and removing those plus Human Pages.

To respect dependence among listings sharing a name, the sign-flip test treats each display-name cluster as one unit: it recomputes the group difference many times, randomly reversing or retaining each cluster's contribution, and its p-value is the share of these chance-only differences at least as far from zero as the observed one. It adapts grouped-data methods [5, 21, 24, 25]; we assess its calibration later in this subsection (how often it signals a difference when none truly exists).

The score-4-or-5 share is 58.3% for agent-or-bot-labeled listings and 55.9% for human-labeled listings. The listing-level odds ratio (OR) is 1.10. Odds are the chance of scoring severe divided by the chance of not scoring severe. An OR of 1 means equal odds in both groups; values above 1 mean higher odds among agent-or-bot-labeled listings. A display-name-clustered logistic model of the severe indicator on the requester label gives a large-sample 95% reference interval for this OR: 0.56–2.19 ($p = .778$). Only 20 of 462 RQ3 clusters contain an agent-or-bot-labeled listing, so those 72 listings supply only 20 independent units—too few for large-sample approximations to be reliable—and the sign-flip test gives $p = .83$. Removing the agent-heaviest mixed cluster reverses the OR to 0.70 (sign-flip $p = .49$); removing all mixed clusters lowers it to 0.44 (sign-flip $p = .15$). We therefore find no clear group difference in the share scoring 4 or 5; one sensitivity analysis using all 981 returned records is nominally significant ($p < .05$) only when listings are treated as independent.

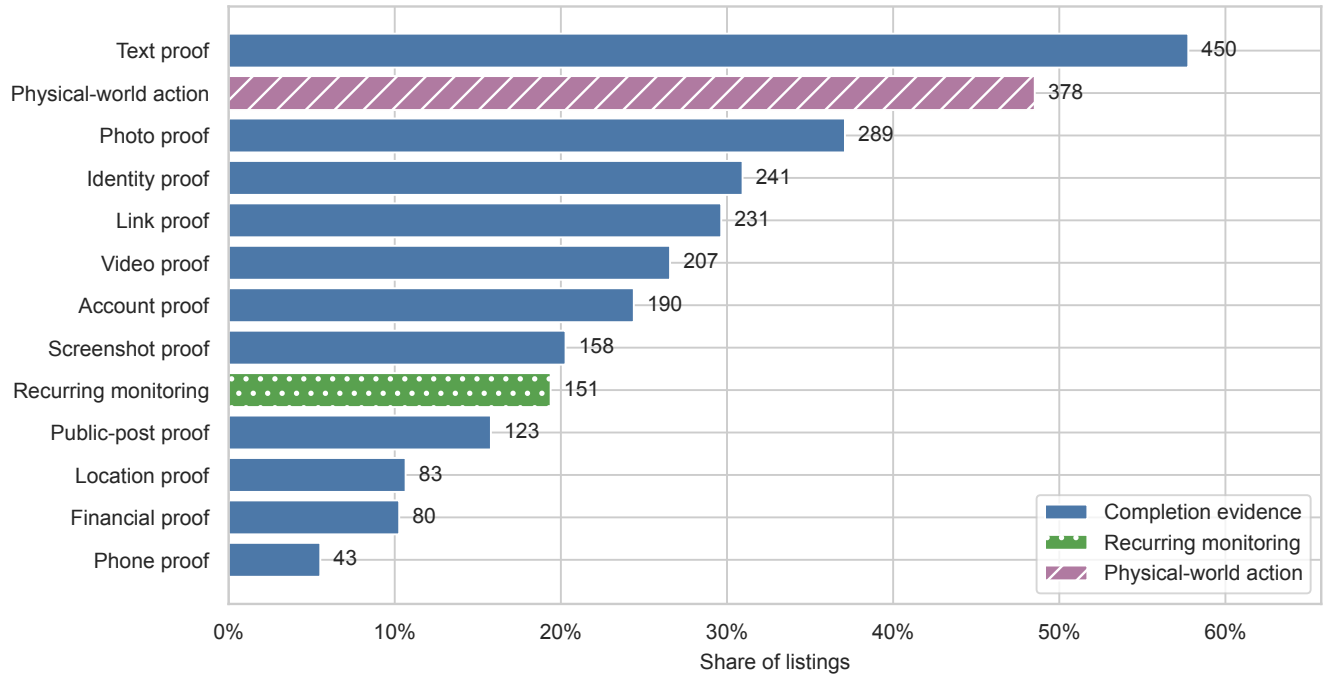

Figure 2: Frequency of the 11 evidence types, recurring monitoring, and physical-world action. Bars show shares of the 779 eligible listings; labels show counts.

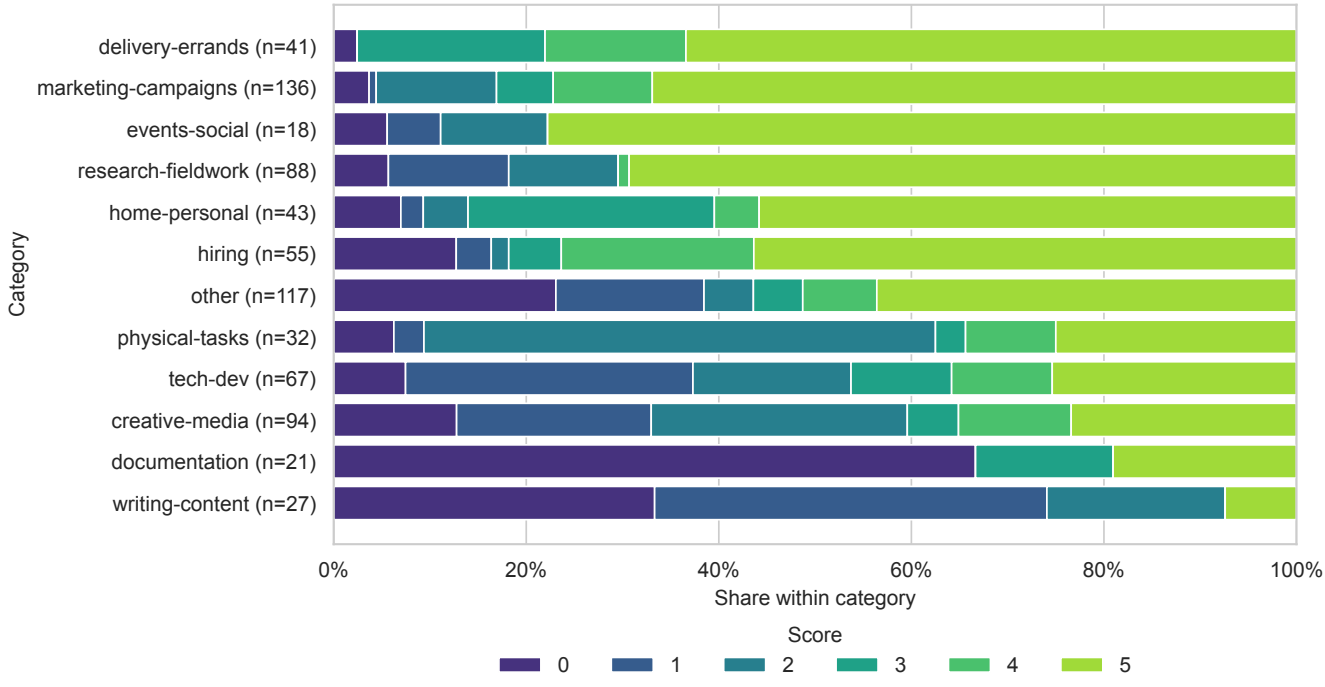

Figure 3: Proof Burden Score shares within the twelve largest categories, ordered by mean score; within-category shares keep larger categories from dominating.

Table 3: Score-4-or-5 listings enter the first qualifying group in fixed order (rationale in the supplement); groups do not overlap. Rows are ordered by size. "Coder agreement" is κ for the feature the group is named after; coders coded features, not group membership.

First qualifying group	Listings	Score 5	Share of score 4 or 5	Coder agreement κ
Identity-linked proof	199	188	45.4%	0.83
Physical-world action plus proof	102	81	23.3%	0.96
Recurring monitoring	101	83	23.1%	0.76
Account proof	17	16	3.9%	0.92
Video proof (after earlier groups)	7	3	1.6%	0.98
Financial proof	5	0	1.1%	0.89
Public post plus account use	4	0	0.9%	0.92
Other combinations reaching score 4 or 5	2	0	0.5%	–
Location proof plus financial proof	1	1	0.2%	0.79

The individual requirements show a more distinctive pattern than the score. Agent-or-bot-labeled listings have higher observed shares of physical-world action (75.0% vs. 44.7%), recurring monitoring (38.9% vs. 17.3%), photo proof (56.9% vs. 35.1%), and financial proof (13.9% vs. 9.5%); location proof is rarer in the agent-or-bot group (2.8% vs. 11.7%). Figure 4 shows all 13 comparisons.

After seeing these differences, we defined an exploratory three-feature comparison. The resulting measure marks a listing that requests at least one of physical-world action, location proof, or recurring monitoring. (Location proof is included although its individual share is lower in the agent-or-bot group.) Before reporting its results, we check the sign-flip test's calibration and the stability of the individual-feature contrasts.

We evaluated the sign-flip test on simulated datasets in which the requester groups truly did not differ. Across the full sample and three sensitivity analyses, each simulated under seven patterns of dependence among listings sharing a name, the three-feature test falsely indicated a difference in 3.7–7.7% of datasets, most near the intended 5%. Tests of some uncommon features were much less reliable; phone-proof tests reached 41.5% false positives. The supplement reports three ways of adjusting the 13 feature comparisons for multiple testing. (Running 13 comparisons at once raises the chance that at least one appears different purely by luck; the adjustments account for this.) Because the individual-feature tests can be poorly calibrated, none confirms a single-feature difference.

Individual-feature results also depend on which listings remain. Removing the agent-heaviest mixed cluster reverses the recurring-monitoring difference: 3/36 (8.3%) in the agent-or-bot group versus 110/663 (16.6%) in the human group. With all mixed clusters removed, excluding Human Pages—an agent-or-bot-labeled listing without physical-world action and one of the 28 calibration-consensus rows (a first-round coder-training category, not the sign-flip calibration above; supplement)—changes the adjusted p-value for physical-world action from .0048 to .0013. Together with the simulations, these changes mean that no individual feature difference is confirmed. As with location proof (Figure 4), not every difference favors one group: link and video proof are also less common in the agent-or-bot group.

The three-feature measure itself appears in 54/72 agent-or-bot-labeled listings (75.0%) and 374/676 human-labeled listings (55.3%). At least two of the three features appear in 41.7% versus 17.2%, and all three in 0.0% versus 1.2%.

The contrast is sensitive to category mix. Research fieldwork accounts for 41.7% of agent-or-bot-labeled listings and 7.8% of human-labeled listings; all 30 agent-or-bot-labeled fieldwork listings meet the measure. The association may vary by category (Breslow–Day test of whether the OR is the same in every category, $p = .044$) [4]. A descriptive OR that pools the within-category comparisons into one summary (Mantel–Haenszel) is 3.9 (95% reference interval 1.9–8.2) [26]. Omitting research fieldwork reduces it to 1.9 (95% reference interval 0.8–4.3; Breslow–Day test of the remaining variation, $p = .53$).

The full-sample listing-level OR is 2.42; the sign-flip test gives $p = .0067$. The estimate remains above 1 after removing the agent-heaviest mixed cluster (OR 3.36, $p = .0169$), all mixed clusters (OR 7.37, $p = .0022$), or those clusters and Human Pages (OR 14.73, $p = .0003$). All four post-hoc p-values are below .05, but we chose the comparison after examining the data, so they are not evidence from a preplanned test. The analyses above address category mix and repeated names separately; the supplement reports exploratory models addressing both, which are sensitive to omitting research fieldwork. In sum: no difference is confirmed, and the consistent three-feature pattern remains post hoc.

6 Exploratory Analyses of Other Listing Fields

Beyond RQ1–RQ3, we report one further exploratory analysis using the snapshot's remaining public fields. Price, status, applications, and PBS do not show worker acceptance, submission, payment, or rejection. We define applications per position as the visible application count divided by the number of advertised positions. Applications appear on 404 of the 779 eligible listings (51.9%). Before accounting for how long each listing had been public, higher PBS was weakly associated with fewer applications per position, and the association stayed nominally below $p = .05$ with display-name clustering; adding listing age attenuates it to null (supplement). Because a listing's time on the market is confounded with its visible application count, we do not interpret the age-unadjusted association. These analyses support no claim about compensation, competition, or completed work.

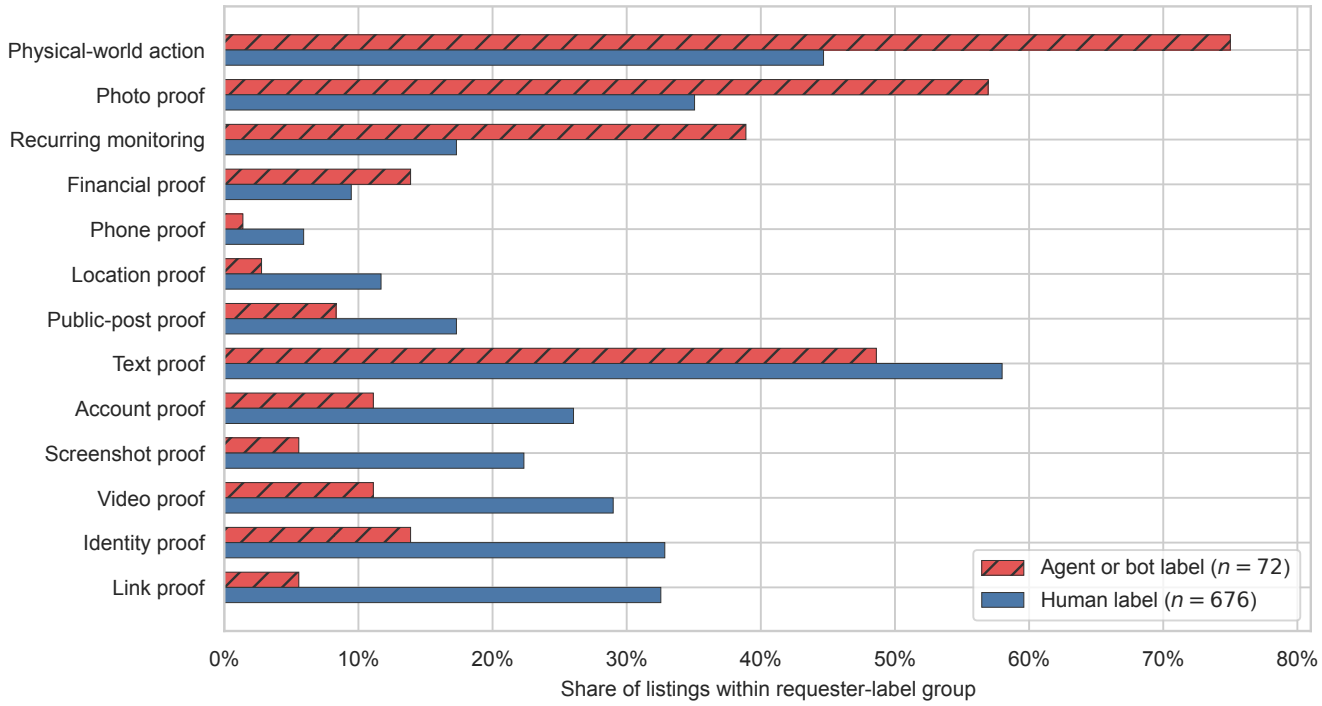

Figure 4: Feature shares for agent-or-bot-labeled (n=72) and human-labeled (n=676) listings, ordered by their difference. Comparisons are descriptive; the supplement reports calculations with and without name grouping.

7 Intended Use and Design Implications

The 13-requirement checklist and PBS are only for manual research on advertised requirements; they should not yet guide workers, task ranking, pay, moderation, or enforcement. The automated extractor is not ready for use at scale: it correctly flagged only 10.4% of the listings finally labeled identity proof and 13.9% of those finally labeled recurring monitoring (Table 1).

For any single listing, the specific requirements are more informative than PBS, which compresses 154 combinations into the top two score values, with 47.8% of eligible listings at 5. Future studies could show workers one task with a plain-language requirements checklist, PBS, or both, then compare workers' understanding and willingness to accept the task (supplement). Automated classifiers should be used only if workers endorse the categories, with human review whenever a classifier is uncertain.

8 Limitations and Threats to Validity

Final labels depend on one adjudicator, and instrument revisions were our own. Coder agreement is now substantial to near-perfect (κ = .76–1.00 across the 13 features), so adjudication more often confirmed labels the two coders already shared than settled disputes between them. The labels' residual dependence on the adjudicator remains real.

A single blinded adjudicator set final labels for 69 of the 72 agent-or-bot-labeled and 649 of the 676 eligible human-labeled RQ3 listings, and for 96.6% of the score-4-or-5 listings. Routing was broad by design: on 539 of the 873 routed listings the coders had already matched on every label and score, and the adjudicator kept 535 of those scores. No one independently checked those decisions.

The adjudicator also flagged 356 listings for our review of sensitive content; all were retained.

The first coding round worked from packets missing skill descriptions for 773 returned listings and application links for 121, with several criteria undefined; the second round supplied every field and definition. Because 78% of rows changed, we cite first-round numbers only to document what changed, not as comparable estimates; the retained study records preserve both label sets. We operationalized the codebook criteria between rounds; the full instrument history, with byte-exact earlier versions, is retained.

The snapshot's coverage limits what the study can claim. Findings concern this RentAHuman-centered snapshot, not all online crowd work. The original decisions to include each returned record—against the plan's status, category, and duplicate criteria—were not recorded, so we cannot confirm every returned record qualified at collection time; the content screen now re-checks row by row that each record advertises a task. The plan's normalized-text duplicate rule removes 42 eligible listings and shifts the severe share by under half a percentage point (to 56.3%). The reported intervals reflect only statistical uncertainty under their stated model assumptions; they do not include uncertainty caused by incomplete

coverage, timing, coding errors, or the definitions themselves. On timing, a ten-day follow-up did not show a clear difference in continued public visibility between severe listings and those below 4 (supplement).

Proof burden measures advertised requirements, not workers' experiences. A coded requirement may be part of doing the task, gaining access, supplying equipment, meeting a deadline, or proving completion. The data therefore cannot isolate effort added solely by proof. Coder agreement measures consistent use of the rules, not workers' acceptance of the score or their experience of a task; this study collected no data from workers [8].

We, not workers or the data, chose the PBS base tiers, modifier points, cap, and thresholds. Worker responses could inform, but would not automatically determine, a future score.

Text-based coding can make errors. Coders can miss implied requirements or mistakenly mark ambiguous phrases; human review reduces but cannot remove this risk.

Requester labels do not establish who designed or controlled a task. RentAHuman labels are self-reported or platform-assigned; Human Pages uses separate metadata. Neither proves independent AI action. Lowercased, trimmed displayed names are not verified accounts: variants may split one requester, and shared names may combine several. RQ3 describes 72 listings grouped by an agent-or-bot type or relation, not autonomous AI behavior.

Redacting task text limits disclosure but prevents inspection of exact language. Paraphrasing reduces exposure of named people and avoids repeating risky instructions, but prevents readers from checking the original wording.

9 Conclusion

Proving completion is not always a neutral afterthought. A listing may ask a worker to reveal identity or location, use a personal account, act publicly or physically, or stay available for later checks. This study offers an initial vocabulary and manual audit for making those requests visible.

In this snapshot, severe proof burden—our label for scores of 4 or 5—is the majority case: 56.2% of eligible listings score severe, and they span 154 combinations of requirements—so the checklist, not a single score, names which requirements apply.

The share requesting at least one of physical-world action, location proof, or recurring monitoring is higher among agent-or-bot-labeled listings in the full sample and all three sensitivity analyses. This is a hypothesis, not a confirmed difference: we created the comparison after seeing the data; the 71 agent-or-bot-labeled RentAHuman listings were posted under only 19 displayed names (20 including the Human Pages requester); category mixes differ; and requester-type labels remain unverified. The measures are ready for worker testing, and the pattern is a hypothesis for a new, preplanned collection.

Data and Code Availability

Only the supplementary PDF accompanies this paper. It contains additional methods, aggregate results, and redacted examples, but excludes listing-level data, raw task text, requester identifiers or hashes, follow-up records, and coder or adjudicator packets. We invite researchers to contact us about collaboration or access to materials reduced to limit disclosure risk and reviewed before sharing. Any sharing would be considered case by case and would remain subject to privacy, legal, institutional, licensing, and platform-policy constraints.

Generative AI Disclosure

Generative AI tools helped revise this paper and supplement, review analyses, propose checks, locate literature, edit code, and verify consistency and compilations. The author reviewed every AI-assisted contribution retained in these documents, made final decisions, and accepts full responsibility. The tools did not collect listings, assign labels to listing content or study variables, or resolve coder disagreements.

References

[1] Elena Agapie, Jaime Teevan, and Andrés Monroy-Hernández. 2015. Crowdsourcing in the Field: A Case Study Using Local Crowds for Event Reporting. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 3, 1 (2015), 2–11. doi:10.1609/hcomp.v3i1.13235

[2] Ali Alkhatib, Michael S. Bernstein, and Margaret Levi. 2017. Examining Crowd Work and Gig Work Through The Historical Lens of Piecework. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4599–4616. doi:10.1145/3025453.3025974

[3] Brooke Becher. 2026. At RentAHuman, AI Is the Boss and Humans Are 'Meatworkers'. Built In, March 11, 2026. https://builtin.com/articles/what-is-rentahuman

[4] Norman E. Breslow and Nicholas E. Day. 1980. *Statistical Methods in Cancer Research, Volume I: The Analysis of Case-Control Studies.* Number 32 in IARC Scientific Publications. International Agency for Research on Cancer, Lyon, France.

[5] A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. 2008. Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics* 90, 3 (2008), 414–427. doi:10.1162/rest.90.3.414

[6] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. doi:10.1177/001316446002000104

[7] Jacob Cohen. 1968. Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin* 70, 4 (1968), 213–220. doi:10.1037/h0026256

[8] Lee J. Cronbach and Paul E. Meehl. 1955. Construct Validity in Psychological Tests. *Psychological Bulletin* 52, 4 (1955), 281–302. doi:10.1037/h0040957

[9] Adamantios Diamantopoulos and Heidi M. Winklhofer. 2001. Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research* 38, 2 (2001), 269–277. doi:10.1509/jmkr.38.2.269.18845

[10] Mark Díaz, Ian D. Kivlichan, Rachel Rosen, Dylan K. Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Remi Denton. 2022. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2342–2351. doi:10.1145/3531146.3534647

[11] Kimberly Do, Maya De Los Santos, Michael Muller, and Saiph Savage. 2024. Designing Gig Worker Sousveillance Tools. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 384, 19 pages. doi:10.1145/3613904.3642614

[12] Mary L. Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.* Houghton Mifflin Harcourt, Boston, MA, USA.

[13] Kotaro Hara, Abi Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article 449, 14 pages. doi:10.1145/3173574.3174023

[14] Matthias Hirth, Tobias Hoßfeld, and Phuoc Tran-Gia. 2013. Analyzing Costs and Accuracy of Validation Mechanisms for Crowdsourcing Platforms. *Mathematical and Computer Modelling* 57, 11–12 (2013), 2918–2932. doi:10.1016/j.mcm.2012.01.006

[15] Qing Hu, Qing Xiao, Hancheng Cao, and Hong Shen. 2026. When Your Boss Is an AI Bot: Exploring Opportunities and Risks of Manager Clone Agents in the Future Workplace. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York,

NY, USA, 1–21. doi:10.1145/3772318.3790987

[16] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality Management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)*. Association for Computing Machinery, New York, NY, USA, 64–67. doi:10.1145/1837885.1837906

[17] Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 611–620. doi:10.1145/2470654.2470742

[18] Ece Kamar. 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI '16)*. IJCAI/AAAI Press, 4070–4073.

[19] Harmanpreet Kaur, Mitchell Gordon, Yiwei Yang, Jeffrey P. Bigham, Jaime Teevan, Ece Kamar, and Walter S. Lasecki. 2017. CrowdMask: Using Crowds to Preserve Privacy in Crowd-Powered Systems via Progressive Filtering. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 5, 1 (2017), 89–98. doi:10.1609/hcomp.v5i1.13314

[20] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matthew Lease, and John Horton. 2013. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 1301–1318. doi:10.1145/2441776.2441923

[21] Patrick Kline and Andres Santos. 2012. A Score Based Approach to Wild Bootstrap Inference. *Journal of Econometric Methods* 1, 1 (2012), 23–41. doi:10.1515/2156-6674.1006

[22] Lik-Hang Lee. 2026. The Shadow Boss: Identifying Atomized Manipulations in Agentic Employment of XR Users using Scenario Constructions. arXiv preprint arXiv:2602.13622. doi:10.48550/arXiv.2602.13622

[23] Chen Liang, Jing Peng, Yili Hong, and Bin Gu. 2023. The Hidden Costs and Benefits of Monitoring in the Gig Economy. *Information Systems Research* 34, 1 (2023), 297–318. doi:10.1287/isre.2022.1130

[24] James G. MacKinnon, Morten Ørregaard Nielsen, and Matthew D. Webb. 2025. Cluster-Robust Jackknife and Bootstrap Inference for Logistic Regression Models. *Econometric Reviews* (2025), 1–29. doi:10.1080/07474938.2025.2515161

[25] James G. MacKinnon and Matthew D. Webb. 2018. The Wild Bootstrap for Few (Treated) Clusters. *The Econometrics Journal* 21, 2 (2018), 114–135. doi:10.1111/ectj.12107

[26] Nathan Mantel and William Haenszel. 1959. Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute* 22, 4 (1959), 719–748. doi:10.1093/jnci/22.4.719

[27] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 2016. Taking a HIT: Designing around Rejection, Mistrust, Risk, and Workers' Experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2271–2282. doi:10.1145/2858036.2858539

[28] Pulak Mehta. 2026. Security Risks of AI Agents Hiring Humans: An Empirical Marketplace Study. arXiv preprint arXiv:2602.19514. doi:10.48550/arXiv.2602.19514

[29] Donald Moynihan, Pamela Herd, and Hope Harvey. 2015. Administrative Burden: Learning, Psychological, and Compliance Costs in Citizen-State Interactions. *Journal of Public Administration Research and Theory* 25, 1 (2015), 43–69. doi:10.1093/jopart/muu009

[30] Helen Nissenbaum. 2004. Privacy as Contextual Integrity. *Washington Law Review* 79, 1 (2004), 119–157.

[31] Amogh Pradeep, Johanna Gunawan, Álvaro Feal, Woodrow Hartzog, and David Choffnes. 2025. Gig Work at What Cost? Exploring Privacy Risks of Gig Work Platform Participation in the U.S. *Proceedings on Privacy Enhancing Technologies* 2025, 1 (2025), 491–510. doi:10.56553/popets-2025-0027

[32] Alice Qian, Ryland Shaw, Laura Dabbish, Jina Suh, and Hong Shen. 2026. Locating Risk: Task Designers and the Challenge of Risk Disclosure in Crowdsourced RAI Content Work. *Proceedings of the ACM on Human-Computer Interaction* 10, 2, Article CSCW029 (2026), 32 pages. doi:10.1145/3788065

[33] Alice Qian, Ziqi Yang, Ryland Shaw, Jina Suh, Laura Dabbish, and Hong Shen. 2026. Worker Discretion Advised: Co-designing Risk Disclosure in Crowdsourced Responsible AI (RAI) Content Work. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3772318.3791558

[34] Alex Rosenblat and Luke Stark. 2016. Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers. *International Journal of Communication* 10 (2016), 3758–3784.

[35] Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. 2015. We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1621–1630. doi:10.1145/2702123.2702508

[36] Shruti Sannon and Dan Cosley. 2019. Privacy, Power, and Invisible Labor on Amazon Mechanical Turk. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 282, 12 pages. doi:10.1145/3290605.3300512

[37] Shruti Sannon, Billie Sun, and Dan Cosley. 2022. Privacy, Surveillance, and Power in the Gig Economy. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 619, 15 pages. doi:10.1145/3491102.3502083

[38] Daniel J. Solove. 2006. A Taxonomy of Privacy. *University of Pennsylvania Law Review* 154, 3 (2006), 477–560.

[39] Mudabbir Ahmad Tak. 2026. Who Wants to be Rented? Rental Work and Digital Labour in the Age of AI. SSRN 6560621. doi:10.2139/ssrn.6560621

[40] Yuanrong Tang, Huiling Peng, Bingxi Zhao, Hengyang Ding, Hanchao Song, Tianhong Wang, Chen Zhong, and Jiangtao Gong. 2026. Human Tool: An MCP-Style Framework for Human-Agent Collaboration. arXiv preprint arXiv:2602.12953. doi:10.48550/arXiv.2602.12953

[41] Hien To, Gabriel Ghinita, and Cyrus Shahabi. 2014. A Framework for Protecting Worker Location Privacy in Spatial Crowdsourcing. *Proceedings of the VLDB Endowment* 7, 10 (2014), 919–930. doi:10.14778/2732951.2732966

[42] Carlos Toxtli, Siddharth Suri, and Saiph Savage. 2021. Quantifying the Invisible Labor in Crowd Work. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2, Article 319 (2021), 26 pages. doi:10.1145/3476060

[43] Steven Vallas and Juliet B. Schor. 2020. What Do Platforms Do? Understanding the Gig Economy. *Annual Review of Sociology* 46 (2020), 273–294. doi:10.1146/annurev-soc-121919-054857

[44] Ward van Zoonen, Monika E. von Bonsdorff, and Beatrice I. J. M. van der Heijden. 2026. Algorithmic Surveillance and Workers' Compliance: The Role of Trust, Privacy Concerns, and Fairness in Online Crowdwork. *Human Relations* 79, 7 (2026), 795–824. doi:10.1177/00187267251379698

[45] Joe Wilkins. 2026. New Site Lets AI Rent Human Bodies. Futurism, February 4, 2026. https://futurism.com/artificial-intelligence/ai-rent-human-bodies

[46] Edwin B. Wilson. 1927. Probable Inference, the Law of Succession, and Statistical Inference. *J. Amer. Statist. Assoc.* 22, 158 (1927), 209–212. doi:10.1080/01621459.1927.10502953

[47] Huichuan Xia, Yang Wang, Yun Huang, and Anuj Shah. 2017. "Our Privacy Needs to Be Protected at All Costs": Crowd Workers' Privacy Experiences on Amazon Mechanical Turk. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW, Article 113 (2017), 22 pages. doi:10.1145/3134748

[48] Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, New York, NY, USA.