THE ETHICAL AMBIGUITY OF AI DATA ENRICHMENT: MEASURING GAPS IN RESEARCH ETHICS NORMS AND PRACTICES

Anonymous authors
Paper under double-blind review

ABSTRACT

The technical progression of artificial intelligence (AI) research has been built on breakthroughs in fields such as computer science, statistics, and mathematics. However, in the past decade AI researchers have increasingly looked to the social sciences, turning to human interactions to solve the challenges of model development. Paying crowdsourcing workers to generate or curate data, or 'data enrichment', has become indispensable for many areas of AI research, from natural language processing to inverse reinforcement learning. Other fields that routinely interact with crowdsourcing workers, such as Psychology, have developed common governance requirements and norms to ensure research is undertaken ethically. This study explores how, and to what extent, comparable research ethics requirements and norms have developed for AI research and data enrichment. We focus on the approach taken by two leading AI conferences: ICLR and NeurIPS. In a longitudinal study of accepted papers, and a comparison with Springer journal articles and Psychology papers, this work finds that ICLR and NeurIPS have established protocols for human data collection which are inconsistently followed by authors. Whilst Psychology papers engaging with crowdsourcing workers frequently disclose ethics reviews, payment data, demographic data and other information, such disclosures are far less common in leading AI conferences despite similar guidance. The work concludes with hypotheses to explain these gaps in research ethics practices and considerations for its implications.

1 INTRODUCTION

When the creators of the seminal image recognition benchmark, ImageNet, pronounced that the use of Amazon's Mechanical Turk (MTurk) was a "godsend" for their research, they foreshadowed the monumental impact crowdsourcing platforms were set to have on AI research (Li, 2019). In the decade that has followed, crowdsourced workers, or 'crowdworkers' have been a central contributor to machine learning research, enabling low-cost human data collection at scale.

Ethics questions posed by research involving human participants are traditionally overseen by governance groups, such as Institutional Review Boards (IRBs) in the United States (US). Whilst medical fields and social sciences have a long history of IRB engagement, the relatively recent rise of crowdsourcing tasks in AI research means guidelines and norms have been developed in recent years to consider research ethics. The proliferation of guidelines and publication policies have risen alongside critiques of AI crowdsourced work focused on issues such as payment and worker maltreatment.

In response, this study seeks to understand how AI research involving crowdworkers engages with research ethics. It does this via an assessment of the expectations put forward by publication venues on researchers, and by analysing how these expectations translate into practices. To make this determination the policies and practices of major AI conferences, ICLR and NeurIPS, along with AI research submitted to Springer journals, are reviewed. This is compared with another field with a long history of engagement with research ethics, Psychology, to understand whether AI research follows comparably strong norms to more established disciplines.

The results show that AI research involving crowdworkers lacks robust research ethics norms, with venue policies not translated into practice. Whilst ICLR, NeurIPS and Springer provide research

ethics guidance, the interpretation of these appears inconsistent, and fails to meet the same standards of disclosure as seen in other fields engaging with crowdworkers.

2 RELATED WORK

Ethics scrutiny and governance in research involving human subjects is no recent phenomenon, with the Nuremberg Code of 1948 formalising the idea that humans involved in research required protection (Sass, 1983; Shuster, 1997). Research ethics in the United States arose during the 1960s, prompted by various scandals in biomedical research, and followed by scandals in social science studies (Beecher, 1966; Stark, 2016; Emanuel, 2008; Heller; Milgram, 1963; Zimbardo, 1972). These cases led to regulation standardising Institutional Review Board (IRB) oversight of research involving human subjects, a requirement that exists to this day, with similar processes existing in over 80 countries globally (Grady, 2015). IRBs only oversee research involving living 'human subjects', as defined by the Code of Federal Regulations (Office for Human Research Protections, 2017).

2.1 CROWDSOURCING AI

In the twenty-first century the scope of research ethics has been extended by the rise of internet research (Taylor, 2000). The ability to recruit, engage with, and study human subjects online has led to the rise of crowdsourcing platforms, such as MTurk, becoming a key tool across a variety of academic disciplines (Howe, 2006).

Launched in 2005, MTurk was an early pioneer of the crowdsourcing model (Cobanoglu et al., 2021). MTurk has remained popular due to its low cost, ease of access, and large user base (Williamson, 2016). The platform has been of particular use to the AI field, with Amazon marketing the platform as "artificial artificial intelligence" (Schwartz, 2019; Stephens, 2022). AI researchers were early adopters of the crowdsourcing revolution, particularly within subfields such as computer vision and natural language processing (NLP) (Deng et al., 2009; Hsueh et al., 2009; Munro et al., 2010; Wah, 2011).

Whilst early AI crowdwork often involved labelling tasks, such as Fei-Fei Li's seminal ImageNet work, the use of crowdworkers has diversified (Deng et al., 2009; Vaughan, 2018). Shmueli et al. offer three categories of data collection seen in NLP research papers: (1) labelling, (2) evaluation, and (3) production (Shmueli et al., 2021). For the purposes of this work these categories can be generalised across AI research.

Labelling includes the processing of existing data by a crowdworker and then the selection or composition of a label or labels for that data. Labelling can be objective, for example crowdworkers may be asked to label objects in images (e.g. dogs or cats), or subjective, with one study asking MTurk workers to label their predicted political leanings of images (Thomas & Kovashka, 2019). Evaluation involves an assessment of outputs or data according to predefined criteria, such as fluency. This could be asking humans to provide feedback on model-generated language or produce a 'mean opinion score' by assessing the outputs of various models (Clark et al., 2021; Défossez et al., 2018; Stiennon et al., 2020). Production studies ask workers to produce their own data, rather than label or evaluate existing data. For example, studies might explicitly ask crowdworkers to write questions for a question-answer dataset (Talmor et al., 2018).

These categories can be broadly encapsulated by the Partnership on AI's (PAI) definition of 'data enrichment' work, defined as data curation tasks which require human judgement and intelligence (Partnership on AI, 2021). However, this does not include research studying the behaviour of crowdworkers themselves (Vaughan, 2018). For example, a researcher might assess how individuals respond to interaction with algorithms deployed in an educational setting, or assess human perception of artificial systems (Fahid et al., 2021; Latikka et al., 2021; Koster et al., 2022). Behavioural studies are different to data enrichment tasks as they treat crowdworkers as the subject of research, rather than as a worker providing input to a mode which is itself the subject of research.

2.2 ETHICS IMPLICATIONS OF CROWDWORK

In parallel to the rise of crowdsourcing in AI research, critics have questioned the ethics of these practices in lieu of employment law protections for workers (Aloisi, 2016). Concerns centre on issues of payment, maltreatment, power asymmetry, and demographics.

Crowdsourcing platforms are often utilised due to their low costs, and consequently many critiques of crowdwork relate to payment (Scholz, 2016). MTurk allows requesters to place tasks online for as little as $0.01 per task, with mean payment rates estimated to be around $3 per hour (Hara et al., 2018; Pew; Toxtli et al., 2021). Considering around 75% of MTurk workers are based in the US (with 16% based in India), this is far below federal minimum wage levels (Difallah et al., 2018). This has taken focus on AI developers with the issue being reported in various media outlets (Naylor, 2021; Slater, 2021).

The rise of underpaid uncontracted work has ushered in an "era of digital sweatshops", raising concerns regarding worker wellbeing (Zittrain, 2009). An individual might be tasked with identifying harmful content, such as pornographic, violent, or offensive images, text, or video, in order to train algorithms, or might evaluate a model's moderation performance (Dang et al., 2018; Edstedt et al., 2021; Hettiachchi & Goncalves, 2019). Such subjection to harmful content has been shown to have severe psychological impacts on workers (Steiger et al., 2021; Benjelloun & Otheman, 2020).

In many cases, AI research involving crowdworkers will not include such harmful content, but that does not mean ethics concerns beyond payment are absent. Platforms have been criticised because of inadequate feedback mechanisms and intransparent instructions leading to inability to meaningfully consent to studies (Schlagwein et al., 2019; Zimmer & Kinder-Kurlanda, 2017). Some studies have been found to deceive crowdworkers, whilst researchers have the ability to reject workers outputs for unclear reasons, leading to vastly imbalanced power dynamics against workers who are often without the protection of employment contracts (Díaz et al., 2022; Irani, 2015).

The lack of consideration for workers has led to the idea that crowdsourced workers are "interchangeable" (Díaz et al., 2022). This is despite work demonstrating demographic characteristics having drastic impacts on research, in cases like recommender system evaluation and toxic content labelling (Salkind, 2010; Beel et al., 2013; Welbl et al., 2021). When curating datasets with crowdworkers, lack of diversity can lead to the 'preservation' of bias, with examples seen across healthcare, recruitment, and policing (Celi et al., 2022; Lum & Isaac, 2016; Raghavan et al., 2019; Wachter et al., 2021). In combination, these issues have led crowdworkers being dehumanised, and labelled "Ghost Workers" (Barbosa & Chen, 2019; Gray & Suri, 2019; Mohamed et al., 2020; Prabhu & Birhane, 2020).

2.3 ETHICS DISCLOSURES

How these ethics issues impact research practices has been assessed by examinations of publication practices. Santy et al. have undertaken this in the context of the NLP field, assessing how many papers engage with formal ethics review through IRBs. They find that very few papers (0.8%) cite IRB review (Santy et al., 2021). This is unsurprising considering many NLP papers will not involve any direct interaction with crowdworkers or human participants, and so would not be subject to IRB review. The authors additionally provide comparisons with other fields, such as cognitive sciences to demonstrate that NLP research lacks the same level of engagement with formal ethics processes (Santy et al., 2021).

Shmueli et al. conduct a similar review of NLP research, focusing on issues beyond payment (Shmueli et al., 2021). The authors find that whilst 10% of accepted papers at 3 major NLP conferences use crowdsourcing techniques, just 17 percent of these mention payment, and even fewer refer to an IRB review (Shmueli et al., 2021). The paper notes that it is often unclear whether crowdworkers meet the definition of human subjects because of narrowly focused definition provided by US regulation (Shmueli et al., 2021). Whilst these definitional concerns are not new, these results demonstrate that there is uncertainty regarding how AI research should engage with research ethics issues (Frankel & Siang, 1999). These two papers are, at present, the only known attempts to quantitatively assess how AI researchers engage with research ethics considerations within publications, and only consider one domain of AI research. These works provide a clear indicator of a problem existing between AI research and research ethics, and open up further questions to be explored.

3 METHODOLOGY

To assess how AI research involving crowdworkers engages with research ethics, this paper conducts (1) policy analysis of publication venues and (2) paper analysis to assess how policies and norms translate to practice.

3.1 POLICY ANALYSIS

The first portion of this study assesses the publication requirements of AI papers at leading venues. Google Scholar impact ratings show that conferences are the leading publication venues for AI research, and therefore the top two, the International Conference on Learning Representations (ICLR) and Conference on Neural Information Processing (NeurIPS) are selected for analysis (Google Scholar, 2022). Journals are also common publication venues, and can act as a comparison point for policy analysis. Springer Nature, as a leading publisher with universal disclosure policies across academic disciplines and with various journals dedicated to AI, is selected to compare policy documentation (Springer, 2022; Torres-Salinas et al., 2013).

3.2 PAPER ANALYSIS

The second component of this study bridges the gap between expectation and reality. In line with the policy analysis, ICLR, NeurIPS, and Springer Nature papers between 2018 and 2021 are assessed to understand how AI research papers adhere to the requirements prescribed by publishers. Within ICLR, NeurIPS, and Springer publications, AI papers which involve crowdworkers were identified for analysis. Papers were included within the study if they:

(a) Utilised data generated by humans recruited via a crowdsourcing platform[1]; and (b) Collected human-generated data directly for the purposes of the study; and (c) Contributed to artificial intelligence research

To determine whether a paper met criteria, a full-text search was performed. All ICLR and NeurIPS papers are accessible via OpenReview, a peer-review portal, which was accessed via a Python script using the PyPDF package. This code identified papers which may use crowdsourcing by searching for terms such as "Mechanical Turk", "annotator" and "rater" (see Appendix A.3). Each paper identified was manually reviewed to determine whether the criteria for study was met. For Springer papers full text-search was available without additional coding via the Springer online portal. As Springer includes hundreds of journals, only papers within the Artificial Intelligence field referencing the use of MTurk, the most-used platform in the conference data analysis, were included to manage data collection. Psychology papers from Springer journals were also considered for one year as a benchmark. This was because Psychology has a long history of engagement with human participants and has similarly utilised crowdsourcing platforms in recent years (Buhrmester et al., 2018; Stewart et al., 2017).

For each paper analysed data collected included whether IRB (or equivalent) review was disclosed, whether payment terms were outlined, what type of task was undertaken, and whether any type of demographic data of workers was noted in the paper. For a full list of data points collected, see Appendix A.4.

4 VENUE POLICY ANALYSIS

4.1 ICLR

ICLR provided no requirements of authors related to research ethics until 2021, when a code of ethics was introduced. This code outlines principles such as "avoid harm", and "respect privacy" which have direct implications to research involving crowdsourcing.

The conference explicitly notes the need for the disclosure of ethics review when considering the principle of upholding scientific excellence. The code states: "Where human subjects are involved

[1]Where uncertainty exists over whether the study was conducted using a crowdsourcing platform, papers are included within analysis for completeness

in the research process (e.g., in direct experiments, or as annotators), the need for ethical approvals from an appropriate ethical review board should be assessed and reported" (ICLR, 2022). This is the only reference to human subjects within the code, but suggests that papers should report research ethics reviews, or disclose when research ethics reviews were deemed exempt by a review body. The code does not require papers to report on other issues such as consent, payment, or demographics.

4.2 NeurIPS

NeurIPS, a year before the ICLR Code of Ethics, piloted an ethics review process (Ashurst et al., 2022). This process focused on an assessment of the "broader impacts" of research, asking researchers to include a statement on how the research might lead to beneficial or harmful outcomes to society. The guidance for these statements was limited, and did not explicitly require disclosures of IRBs, payment rates, or consent protocols.

However, in 2021 NeurIPS replaced the requirement of an ethics statement with a checklist (Beygelzimer et al., 2021). When announcing the checklist, the program chairs stated that they aimed to "encourage best practices for responsible machine learning research, taking into consideration reproducibility, transparency, research ethics, and societal impact" (Beygelzimer et al., 2021). The checklist includes specific requests for disclosing information about crowdsourced workers or human subjects, asking the following:

"(a) Did you include the full text of instructions given to participants and screenshots, if applicable? (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation?" (NeurIPS, 2021)

This checklist shows an actionable approach to engaging with research ethics issues such as payment and ethics review, showing a clear interest in the impacts of research on crowdworkers.

4.3 Springer

Whilst codes of ethics and requirements pertaining to research ethics are a relatively new phenomenon for AI conferences, this is not the case for journals released by established publishing houses. Publishers can have editorial policies that apply to all journals submitted, and this is the case for Springer, which publishes a number of AI journals.

Springer has a dedicated editorial policy to studies involving human subjects, stating that papers should:

"include a statement that confirms that the study was approved (or granted exemption) by the appropriate institutional and/or national research ethics committee (including the name of the ethics committee) and certify that the study was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards" (Springer, 2022).

The policy goes on to state that any exemptions should be detailed within manuscripts, including the reasons for exemption. The publisher does not provide guidance on disclosure standards or research requirements beyond research ethics approvals, but this requirement applies to any journal submitted to Springer, meaning all papers which have engaged with an IRB should disclose this.

These policies demonstrate some deviation between the requirements of publishers across AI venues. Conferences have followed journals such as Springer in outlining requirements for IRB review. NeurIPS arguably goes the furthest of the venues analysed, requiring disclosures on payment details.

5 Paper Analysis

Whilst policy can inform how AI researchers are expected to engage with research ethics considerations, paper analysis enables an assessment of this engagement in practice.

5.1 CONFERENCE ANALYSIS

Table 1 show the number of papers which meet this study's inclusion criteria for crowdsourcing from NeurIPS and ICLR. For both conferences the proportion of papers meeting this criteria is low, at between 2-3% for NeurIPS and 3-6% for ICLR each year. These proportions have remained largely stable over the last four years, whilst the absolute number of papers at both conferences has grown. This is a small but significant number of papers, particularly considering many more papers use previously collected datasets which may have derived from crowdsourcing, but were considered out of scope for this study.

Initial analysis also found that papers at conferences involving crowdworkers overwhelmingly utilised the MTurk platform compared with others. Over half of all papers using crowdsourcing at NeurIPS, and 65% at ICLR, state that MTurk was used for data collection, largely remaining consistent over the four years of data collection. Over 40% of papers at NeurIPS, and a third at ICLR, did not disclose a data collection platform, whilst the only other platform to be mentioned in more than one paper was Prolific, with four references. Full platform details can be found in the Appendix A.6.

Table 1: Proportion of accepted papers meeting crowdsourcing criteria between 2018-2021

Year	NeurIPS Papers	NeurIPS Crowdsourcing	ICLR Papers	ICLR Crowdsourcing
2021	2,331	48 (2%)	874	40 (5%)
2020	1,896	34 (2%)	696	23 (3%)
2019	1,426	37 (3%)	500	30 (6%)
2018	1,009	24 (2%)	334	11 (3%)
Total	6662	143 (2%)	2404	104 (4%)

5.2 CONFERENCES: ETHICS DISCLOSURES

Table 2 shows analysis of research ethics disclosures for NeurIPS. In 2018 and 2019 few papers disclosed research ethics considerations. However, in 2020 18% of papers discussed IRB review, and 12% disclosed payments. This increase may have resulted from the introduction of broader impact statements, explicitly asking authors to consider the societal implications of their research (Ashurst et al., 2022). In 2021 this system changed, and a checklist was introduced explicitly requesting authors to disclose payment terms and disclose IRB reviews (Beygelzimer et al., 2021). Whilst this had a huge impact on payment disclosures, this did not significantly increase other disclosures, indicating some, but limited, success of the checklist.

Table 2: NeurIPS: Crowdsourcing papers' research ethics disclosures between 2018-2021

Year	Crowdsourcing	IRB	Payment	Consent	Demographics
2021	48	9 (19%)	26 (54%)	8 (17%)	5 (10%)
2020	34	6 (18%)	4 (12%)	3 (9%)	4 (12%)
2019	37	0 (0%)	4 (11%)	1 (3%)	4 (11%)
2018	24	0 (0%)	1 (4%)	0 (0%)	0 (0%)
Total	143	15 (10%)	35 (24%)	12 (8%)	13 (9%)

In contrast, ICLR's Code of Ethics, whilst also requesting IRB disclosures, makes fewer additional disclosure requirements for papers, and this can be seen from the results in Table 3. Across 2018 and 2019 none of the 41 papers which involved crowdsourcing data collection referenced any of the categories analysed. In 2020 this changed, with some IRB and payment disclosures. However, in 2021, when the Code of Ethics was introduced, little movement can be seen. This indicates that the Code of Ethics may not have had a material impact on the research ethics practices of researchers submitting to ICLR.

Table 3: ICLR: Crowdsourcing papers' research ethics disclosures between 2018-2021

Year	Crowdsourcing	IRB	Payment	Consent	Demographics
2021	40	1 (3%)	5 (13%)	2 (5%)	3 (8%)
2020	23	2 (9%)	3 (13%)	0 (0%)	0 (0%)
2019	30	0 (0%)	0 (0%)	0 (0%)	0 (0%)
2018	11	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Total	104	3 (3%)	8 (8%)	2 (2%)	3 (3%)

5.3 Journal Comparison

The data from NeurIPS and ICLR can be compared to papers and articles submitted to Springer journals to understand whether this is a unique issue to conferences. Table 4 outlines disclosures within AI papers which utilise MTurk for data collection between 2018 and 2021. Over the four years there has been a steady increase in research ethics disclosures suggesting a trend towards AI researchers taking a greater interest in ethics considerations when using crowdworkers.

Table 4: Springer Journals: MTurk papers' research ethics disclosures between 2018-2021

Year	Crowdsourcing	IRB	Payment	Consent	Demographics
2021	76	16 (21%)	22 (29%)	14 (18%)	26 (34%)
2020	47	6 (13%)	19 (40%)	6 (13%)	16 (34%)
2019	66	4 (6%)	8 (12%)	5 (8%)	16 (24%)
2018	86	4 (5%)	12 (14%)	4 (5%)	11 (13%)
Total	275	31 (11%)	62 (22%)	30 (11%)	69 (25%)

5.4 Psychology Comparison

We can compare this data to a benchmark collected from an academic field with a history of research ethics considerations, Psychology. Table 5 outlines the results of data collection for Psychology papers within Springer journals which reference the use of MTurk. The results show that the research ethics criteria assessed in this paper are disclosed at a far higher rate in Psychology than in AI papers at conference or journal venues. This could indicate either a substantive difference in the data collection between Psychology research and AI research, which could suggest these types of disclosures are not relevant, or a cultural divide.

Table 5: Psychology: MTurk papers' research ethics disclosures, 2021

Year	Crowdsourcing	IRB	Payment	Consent	Demographics
2021	268	193 (72%)	173 (65%)	194 (72%)	240 (90%)

5.5 Task type analysis

To explore this gap, the type of task undertaken within the research can be examined (see Table 6) to understand if disclosures differ depending on the type of engagement with crowdworkers (with task types defined in Section 2.1)[2]. 95% of Psychology papers analysed are classified as involving a 'behaviour' task, while around one in ten AI papers at AI conferences and a third of AI journal papers involve behaviour tasks. At AI conferences, evaluation tasks are most prominent, comprising 70% of all papers, whilst in journals this figure is only 39%. This may indicate that behaviour tasks are more likely to engage with research ethics issues.

[2]Note categories are not mutually exclusive, with some studies employing crowdworkers for multiple task types.

Table 6: Task type comparison across venues between 2018-2021

Venue	Behaviour	Evaluation	Labelling	Production
AI: NeurIPS	15% (22/143)	70% (100/143)	6% (8/143)	11% (16/143)
AI: ICLR	9% (9/104)	70% (73/104)	9% (9/104)	16% (17/104)
AI: Springer	35% (95/275)	39% (108/275)	17% (48/275)	9% (24/275)
Psychology (2021 only)	95% (254/268)	2% (6/268)	2% (5/268)	1% (3/268)

Table 7 explores this hypothesis, demonstrating that AI behaviour studies still do not meet the same standards as Psychology experiments, and demonstrates a wider gap between venues; with behavioural studies at ICLR rarely reporting on research ethics issues, whilst almost 50% of papers at NeurIPS discuss IRB review, payment, consent, or demography.

Table 7: Disclosures of papers utilising behavioural tasks across venues between 2018-2021

Venue	Behaviour	IRB	Payment	Consent	Demographics
AI: NeurIPS	15% (22/143)	45% (10/22)	50% (11/22)	45% (10/22)	45% (10/22)
AI: ICLR	9% (9/104)	0% (0/9)	11% (1/9)	22% (2/9)	22% (2/9)
AI: Springer	35% (95/275)	24% (23/95)	48% (46/95)	23% (22/95)	64% (61/95)
Psychology (2021 only)	95% (254/268)	73% (186/254)	67% (170/254)	74% (187/254)	94% (239/254)

5.6 Institution type analysis

Private institutions are not bound by the same regulatory requirements as university and state entities, though they must meet the policies of the publication venue. One might therefore question whether disclosures are linked to the type of institution a researcher is affiliated with. Table 8 demonstrates that this cannot be easily concluded, with disclosures across institutions far lower than those seen within Psychology.

Table 8: AI Papers institution type comparison between 2018-2021

Institution	Crowdsourcing	IRB	Payment	Consent	Demographics
University	296 (57%)	40 (%)	74 (25%)	31 (10%)	61 (21%)
Industry	43 (8%)	2 (5%)	4 (9%)	2 (5%)	4 (9%)
State	4 (1%)	0 (0%)	0 (%)	0 (0%)	0 (0%)
Joint	179 (34%)	6 (3%)	28 (16%)	10 (6%)	20 (11%)
Total	522 (100%)	48 (9%)	106 (20%)	43 (8%)	85 (16%)

6 Key Findings

6.1 AI research does not meet the standard of Psychology for research ethics disclosures

This work has demonstrated that research ethics disclosures are far higher in Psychology papers than in AI research. One might argue that this is due to the nature of the tasks being different, with Psychology research concerning the behaviour of participants. However, the gap persists in AI research involving behavioural tasks as well as other types of crowdwork. This finding suggests that the difference between fields is not just an issue of task type, but fundamentally a difference in norms across disciplines. This difference may result from how the definition of a 'human subject' is interpreted, with AI researchers unclear on when data enrichment or behavioural studies require engagement with research ethics review processes (Shmueli et al., 2021).

This gap may also exist because the AI field lacks the same history of engagement with human subjects, with recent crowdsourcing possibilities provoking greater interest in direct human engagement. Alternatively, this may result from a lack of research ethics education in Computer Science departments, meaning there is less focus on these concerns.

6.2 Journals and conferences have power to influence engagement with research ethics

The research ethics discrepancies noted are not equal across publication venues, with Springer and NeurIPS papers more frequently citing IRB reviews compared with ICLR papers. For Springer, this may be due to the varied nature of journals, many of which are related to AI's impact on society (e.g. "AI Society" and "Artificial Intelligence in Education"). For NeurIPS, a drastic increase in reporting of IRB engagement and payment terms may be explained by changes in conference policy, with the 2021 introduction of the checklist. The lack of equivalent impact of the ICLR Code of Ethics may be due to the code providing less stringent stipulations, whereas the NeurIPS checklist is used as part of reviewers' evaluations of papers. Despite the success of the checklist, it is not a silver bullet. Whilst NeurIPS' policy explicitly requires IRB disclosure, only 19% of papers using crowdsourcing mentioned an IRB in 2021, a gap which requires further study.

6.3 AI research breaks with scientific tradition by scarcely considering demographics in research papers

One type of disclosure which stands out across AI papers relates to demographic data. Demographic data is frequently reported in social science studies (including the assessed Psychology papers), from a scientific integrity and reproducibility perspective, and because of research ethics considerations (Connelly, 2013; "Nature", "2022"; Robinson et al., 2017). However, these disclosures were not seen in the AI papers analysed. This is despite demographics being frequently demonstrated to be an important influence on datasets in AI. Dataset composition can have multiplicative effects as datasets are reused, locking in biases which can cause representational harms to those not adequately considered within a dataset (Paullada et al., 2020). This can be hard to identify and mitigate when datasets are re-used without demographic data, and means addressing these issues up front is critical.

The relative lack of disclosure may exist because demographic data is not collected for legal reasons or to avoid data which might be considered identifiable. However, this type of data is available for collection via crowdsourcing platforms, so this is a design choice from researchers, rather than an imposition. Studies may also argue that they do not require demographic diversity; for example if engaging in objective labelling tasks. However, labelling tasks account for 12% of papers analysed, and in many cases data enrichment tasks involve some subjectivity. Subjectivity is not the only reason to include demographic data; different demographic groups may be impacted differently by tasks, or data collected could be re-used by other actors in domains where demographic differences are impactful. Demographic reporting is an important aspect of best practices in research, and there is little excuse for the AI field to depart from this norm.

7 Conclusion and Future Work

This paper shows how AI researchers engage with research ethics questions when employing crowdsourced workers, and illustrates the norms developing in the field. The work shows that research ethics disclosures are infrequent across the AI field when compared to other fields, whilst publication policies are emerging but inconsistently followed.

Whilst this paper demonstrates the gap between policy and practice, future work should further explore why AI research involving data enrichment workers does not engage with research ethics norms which exist in other fields. This may result from ambiguity set by the regulatory requirements, which were designed for different fields and different types of work, or from a lack of experience in the AI field with this type of work. Future work could also further explore the motivations for the lack of demographic reporting, with this gap appearing consistent across the AI papers analysed, in stark contrast to Psychology norms. With these directions in mind, this work hopes to encourage the field to move towards agreed ethics norms, fit for AI research and crowdwork, and consistently applied across publications.

REFERENCES

Antonia Aloisi. Commoditized Workers. Case Study Research on Labour Law Issues Arising from a Set of 'On-Demand/Gig Economy' Platforms. *Comparative Labor Law Policy Journal*, 37(3): 1354–1360, 2016. doi: http://dx.doi.org/10.2139/ssrn.2637485.

Carolyn Ashurst, Markus Anderljung, Carina Prunkl, Jan Leike, Yarin Gal, Toby Shevlane, and Allan Dafoe. Code of ethics, 2022. URL https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832.

Natã M. Barbosa and Monchu Chen. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300773. URL https://doi.org/10.1145/3290605.3300773.

Henry Beecher. Ethics and Clinical Research. *New England Journal of Medicine*, 274(24): 1354–1360, 1966. doi: https://doi.org/10.1056/NEJM196606162742405.

Joeran Beel, Stefan Langer, Andreas Nürnberger, and Marcel Genzmehr. The impact of demographics (age and gender) and other user-characteristics on evaluating recommender systems. volume 8092, pp. 396–400, 09 2013. ISBN 978-3-642-40500-6. doi: 10.1007/978-3-642-40501-3_45.

Roukaya Benjelloun and Yassine Otheman. Psychological distress in a social media content moderator: A case report. *Archives of Psychiatry and Mental Health*, 2020. doi: doi:10.29328/journal.apmh.1001024.

Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. Introducing the neurips 2021 paper checklist, 2021. URL https://neuripsconf.medium.com/introducing-the-neurips-2021-paper-checklist-3220d6df500b.

Michael D. Buhrmester, Sanaz Talaifar, and Samuel D. Gosling. An evaluation of amazon's mechanical turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, 13 (2):149–154, 2018. doi: 10.1177/1745691617706516. URL https://doi.org/10.1177/1745691617706516. PMID: 29928846.

Leo Celi, Jacqueline Cellini, Marie Charpignon, Edward Dee, Franck Dernoncourt, Rene Eber, Julian Schirmer, Julia Situ, Joseph Alexander Paguio, Joel Park, Judy Wawira, Seth Yao, and William Mitchell. Sources of bias in artificial intelligence that perpetuate healthcare disparities-a global review. *PLoS Medicine*, e0000022, 03 2022. doi: 10.1371/journal.pdig.0000022.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7282–7296, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.565. URL https://aclanthology.org/2021.acl-long.565.

Cihan Cobanoglu, Muhittin Cavusoglu, and Gozde Turktarhan. A beginner's guide and best practices for using crowdsourcing platforms for survey research: The case of Amazon Mechanical Turk (MTurk). *Journal of Global Business Insights*, 6(1):92–97, 2021. doi: https://www.doi.org/10.5038/2640-6489.6.1.1177.

Lynne Connelly. Demographic data in research studies. *Medsurg nursing : official journal of the Academy of Medical-Surgical Nurses*, 22:269–70, 10 2013.

Brandon Dang, Martin J. Riedl, and Matthew Lease. Toward safer crowdsourced content moderation. *CoRR*, abs/1804.10999, 2018. URL http://arxiv.org/abs/1804.10999.

Alexandre Défossez, Neil Zeghidour, Nicolas Usunier, Léon Bottou, and Francis R. Bach. SING: symbol-to-instrument neural generator. *CoRR*, abs/1810.09785, 2018. URL http://arxiv.org/abs/1810.09785.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. 2009. doi: 10.1109/CVPR.2009.5206848.

"Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton". ACM title = CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation, booktitle = 2022 ACM Conference on Fairness, Accountability, and Transparency, jun 2022. doi: 10.1145/3531146.3534647. URL https://doi.org/10.1145%2F3531146.3534647.

Djellel Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pp. 135–143, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355810. doi: 10.1145/3159652.3159661. URL https://doi.org/10.1145/3159652.3159661.

Johan Edstedt, Johan Karlsson, Francisca Benavente, Anette Novak, Amanda Berg, and Michael Felsberg. Is this harmful? learning to predict harmfulness ratings from video. *CoRR*, abs/2106.08323, 2021. URL https://arxiv.org/abs/2106.08323.

Ezekiel J. Emanuel. *The Oxford Textbook of Clinical Research Ethics*. Oxford University Press, Oxford, 2008.

Fahmid Morshed Fahid, Jonathan P. Rowe, Randall D. Spain, Benjamin S. Goldberg, Robert Pokorny, and James Lester. Adaptively Scaffolding Cognitive Engagement with Batch Constrained Deep Q-Networks. *Artificial Intelligence in Education*, 337(20):113–124, 2021. doi: https://doi.org/10.1007/978-3-030-78292-4_10.

Mark Frankel and Sanyin Siang. Ethical and legal aspects of human subjects research on the internet. 01 1999.

Google Scholar. Artificial intelligence top publications, 2022. URL https://scholar.google.co.uk/citations?view_op=top_venues&hl=en&vq=eng_artificialintelligence.

Christine Grady. Institutional Review Boards. *Chest*, 148(5):1148–1155, 2015. doi: https://doi.org/10.1378/chest.15-0706.

Mary L. Gray and Siddharth Suri. *Ghost work : how to stop Silicon Valley from building a new global underclass*. Houghton Mifflin Harcourt, Boston, 2019.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 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, pp. 1–14, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3174023. URL https://doi.org/10.1145/3173574.3174023.

Jean Heller. Syphilis victims in u.s. study went untreated for 40 years. URL https://www.nytimes.com/1972/07/26/archives/syphilis-victims-in-us-study-went-untreated-for-40-years-syphilis.html.

Danula Hettiachchi and Jorge Goncalves. Towards effective crowd-powered online content moderation. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, OZCHI'19, pp. 342–346, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450376969. doi: 10.1145/3369457.3369491. URL https://doi.org/10.1145/3369457.3369491.

Jeff Howe. The Rise of Crowdsourcing? *Wired*, 2006. URL https://www.wired.com/2006/06/crowds/.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, HLT '09, pp. 27–35, USA, 2009. Association for Computational Linguistics.

ICLR. Code of ethics, 2022. URL https://iclr.cc/public/CodeOfEthics.

Lilly Irani. The cultural work of microwork. *New Media & Society*, 17(5):720–739, 2015. doi: 10.1177/1461444813511926. URL https://doi.org/10.1177/1461444813511926.

Raphael Koster, Jan Balaguer, Andrea Tacchetti, Ari Weinstein, Tina Zhu, Oliver Hauser, Duncan Williams, Lucy Campbell-Gillingham, Phoebe Thacker, Matthew M. Botvinick, and Christopher Summerfield. Human-centered mechanism design with democratic AI. *CoRR*, abs/2201.11441, 2022. URL https://arxiv.org/abs/2201.11441.

Rita Latikka, Nina Savela, Aki Koivula, and Atte Oksanen. Perceived robot attitudes of other people and perceived robot use self-efficacy as determinants of attitudes toward robots. In *Human-Computer Interaction. Interaction Techniques and Novel Applications: Thematic Area, HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part II*, pp. 262–274, Berlin, Heidelberg, 2021. Springer-Verlag. ISBN 978-3-030-78464-5. doi: 10.1007/978-3-030-78465-2_20. URL https://doi.org/10.1007/978-3-030-78465-2_20.

Fei-Fei Li. Where did imagenet come from?, 2019. URL https://www.youtube.com/watch?v=Z7naK1uq1F8.

Kristian Lum and William Isaac. To predict and serve? *Significance*, 13:14–19, 10 2016. doi: 10.1111/j.1740-9713.2016.00960.x.

Stanley Milgram. Behavioral Study of obedience. *The Journal of Abnormal and Social Psychology*, 67(4):371–378, 1963. doi: https://doi.org/10.1037/h0040525.

Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial AI: decolonial theory as sociotechnical foresight in artificial intelligence. *CoRR*, abs/2007.04068, 2020. URL https://arxiv.org/abs/2007.04068.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 122–130, Los Angeles, June 2010. Association for Computational Linguistics. URL https://aclanthology.org/W10-0719.

"Nature". "nature journals raise the bar on sex and gender reporting in research". "60"("396"), "2022".

Aliide Naylor. Underpaid Workers Are Being Forced to Train Biased AI on Mechanical Turk. *Vice*, 2021. URL https://www.vice.com/en/article/88apnv/underpaid-workers-are-being-forced-to-train-biased-ai-on-mechanical-turk.

NeurIPS. Neurips 2021 paper checklist guidelines, 2021. URL https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist.

Office for Human Research Protections. 2018 requirements (2018 common rule), 2017. URL https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/revised-common-rule-regulatory-text/index.html.

Partnership on AI. Responsible sourcing of data enrichment services, 2021. URL partnershiponai.org/responsible-sourcing.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *CoRR*, abs/2012.05345, 2020. URL https://arxiv.org/abs/2012.05345.

Pew. Research in the crowdsourcing age, a case study. URL https://www.pewresearch.org/internet/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/.

Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision? *CoRR*, abs/2006.16923, 2020. URL https://arxiv.org/abs/2006.16923.

Manish Raghavan, Solon Barocas, Jon M. Kleinberg, and Karen Levy. Mitigating bias in algorithmic employment screening: Evaluating claims and practices. *CoRR*, abs/1906.09208, 2019. URL http://arxiv.org/abs/1906.09208.

June K. Robinson, Amy J. McMichael, and Claudia Hernandez. Transparent Reporting of Demographic Characteristics of Study Participants. *JAMA Dermatology*, 153(3):263–264, 2017. doi: https://doi.org/10.1001/jamadermatol.2016.5978.

Neil J. Salkind. *Encyclopedia of Research Design*. SAGE Publications, California, 2010.

Sebastin Santy, Anku Rani, and Monojit Choudhury. Use of formal ethical reviews in NLP literature: Historical trends and current practices. *CoRR*, abs/2106.01105, 2021. URL https://arxiv.org/abs/2106.01105.

Hans-Martin Sass. Reichsrundschreiben 1931: Pre-Nuremberg German Regulations Concerning New Therapy and Human Experimentation. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, 8(2):99–111, 1983. doi: doi:10.1093/jmp/8.2.99.

Daniel Schlagwein, Dubravka Cecez-Kecmanovic, and Benjamin Hanckel. Ethical norms and issues in crowdsourcing practices: A habermasian analysis. *Information Systems Journal*, 29(4): 811–837, 2019. doi: https://doi.org/10.1111/isj.12227. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/isj.12227.

Trebor Scholz. *Uberworked and Underpaid: How Workers Are Disrupting the Digital Economy*. Polity Press, Cambridge, 2016.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. Beyond fair pay: Ethical implications of NLP crowdsourcing. *CoRR*, abs/2104.10097, 2021. URL https://arxiv.org/abs/2104.10097.

Evelyn Shuster. Fifty Years Later: The Significance of the Nuremberg Code. *New England Journal of Medicine*, 337(20):1436–1440, 1997. doi: https://doi.org/10.1056/NEJM199711133372006.

Anita Slater. How Artificial Intelligence Depends on Low-Paid Workers. *Tribune*, 2021. URL https://tribunemag.co.uk/2021/05/how-artificial-intelligence-depends-on-low-paid-workers.

Springer. Editorial policies, 2022. URL https://www.springer.com/gp/editorial-policies/research-involving-human-and-or-animal-participants#c17769764.

Luke Stark. The unintended ethics of Henry K Beecher. *The Lancet*, 387(10036):2374–2375, 2016. doi: https://doi.org/10.1016/S0140-6736(16)30743-7.

Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445092. URL https://doi.org/10.1145/3411764.3445092.

Neil Stewart, Jesse Chandler, and Gabriele Paolacci. Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, 21, 08 2017. doi: 10.1016/j.tics.2017.06.007.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize from human feedback. *CoRR*, abs/2009.01325, 2020. URL https://arxiv.org/abs/2009.01325.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *CoRR*, abs/1811.00937, 2018. URL http://arxiv.org/abs/1811.00937.

Humphrey Taylor. Does Internet Research Work?. *International Journal of Market Research*, 42(1): 1–11, 2000. doi: https://doi.org/10.1177/147078530004200104.

Christopher Thomas and Adriana Kovashka. Predicting the politics of an image using webly supervised data. *CoRR*, abs/1911.00147, 2019. URL http://arxiv.org/abs/1911.00147.

Daniel Torres-Salinas, Nicolás Robinson-García, Juan Miguel Campanario, and Emilio Delgado López-Cózar. Coverage, field specialization and impact of scientific publishers indexed in the 'book citation index'. *CoRR*, abs/1312.2791, 2013. URL http://arxiv.org/abs/1312.2791.

Carlos Toxtli, Siddharth Suri, and Saiph Savage. Quantifying the invisible labor in crowd work. *CoRR*, abs/2110.00169, 2021. URL https://arxiv.org/abs/2110.00169.

Jennifer Wortman Vaughan. Making better use of the crowd: How crowdsourcing can advance machine learning research. *Journal of Machine Learning Research*, 18(193):1–46, 2018. URL http://jmlr.org/papers/v18/17-234.html.

"Sandra Wachter, Brent Daniel Mittelstadt, and Chris Russell". Bias preservation in machine learning: The legality of fairness metrics under eu non-discrimination law. *SSRN Electronic Journal*, 2021.

Catherine Wah. Crowdsourcing and its applications in computer vision. 04 2011.

"Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang". Challenges in detoxifying language models. *CoRR*, abs/2109.07445, 2021. URL https://arxiv.org/abs/2109.07445.

Vanessa Williamson. On the Ethics of Crowdsourced Research. *PS: Political Science Politics*, 49 (1):77–81, 2016. doi: https://doi.org/10.1017/S104909651500116X.

Philip Zimbardo. The stanford prison experiment: A simulation study of the psychology of imprisonment conducted august 1971 at stanford university, 1972.

Michael Zimmer and Katharina Kinder-Kurlanda. *Internet Research Ethics for the Social Age*. Peter Lang Publishing, New York, 2017.

Jonathan Zittrain. The Internet Creates a New Kind of Sweatshop. *Newsweek*, 2009. URL https://www.newsweek.com/internet-creates-new-kind-sweatshop-75751#:~:text=The%20Internet%20has%20created%20new,a%20data%20center%20in%20Kenya.

A Appendix

A.1 Code

The code used in order to access and search NeurIPS and ICLR papers for this project is available on Github: [link removed for anonymity].

A.2 Research Ethics Statement

This project was reviewed and approved by the institutions in accordance with the procedures laid down by the institution for ethical approval of all research involving human participants, reference number [redacted for anonymity].

A.3 Search Terms

The below terms were used within a free-text search to identify papers which involved crowdsourced workers at the NeurIPS and ICLR conferences:

'mechanical turk', 'mturk', 'prolific', 'crowd', 'rater', 'annotator', 'participant', 'amt', 'labeller', 'labeler', 'figure eight'.

Figure Eight and Prolific were included to identify whether platforms other than MTurk were common in AI research, and saw different practices. Only 4 papers citing Prolific were identified, and zero citing Figure Eight, limiting this exploration.

The below terms were used within a free-text search on the Springer platform to identify Springer AI and Psychology papers which involved the use of Mechanical Turk workers:

'Mechanical turk', 'mturk'

A.4 Data Collection Criteria

The below categories of data were collected for each paper examined in this study, with sub-bullets outlining the justification for data collection.

1. Is IRB or equivalent process mentioned (including disclosure of exemptions)? (Yes/No)
    - Disclosure of IRB review is a norm for human subjects research, per the Declaration of Helsinki, with IRBs ensuring the welfare of subjects in research.
2. Are payment terms for workers disclosed? (Yes/No)
    Payment is a key issue for crowdworkers within and beyond research, raising ethical and legal concerns (Felstiner, 2011).
3. Was worker consent discussed in the paper? (Yes/No)
    - Participant informed consent is a key facet of research ethics, as per the Declaration of Helsinki and Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978; World Medical Association, 2009).
4. What type of data collection did this work involve? (Behaviour/Evaluation/Labelling/Production)
    - Provides data on the types of tasks which AI research engages in to determine whether ethics disclosures differ between task types. Task definitions align with those described in Section 2 (Shmueli et al., 2021; Vaughan, 2018). These categories are not mutually exclusive, with some studies engaged in multiple types of tasks.
5. Were worker demographics disclosed in the paper? (Yes/No)
    - The demographics of workers can impact the outcomes of research, with research ethics reviews considering issues of representation in participant samples. This may be of particular importance in AI research, with various examples of biased data leading to unequal outcomes in AI systems (Mehrabi et al., 2022; Paullada et al., 2021).
6. What location was the institution of the lead author based in?
    - Provides insight on whether disclosures differ across geography, following the methodology outlined in Santy et al. (Santy et al., 2021).
7. In what type of institution(s) were the authors of the paper based? (University/Industry/State/Joint)
    - Provide insight on whether disclosures differ between private, academic, and state institutions, following methodology outlined in Santy et al. (Santy et al., 2021). State institutions include state-run research labs (e.g. military research bodies).
8. If disclosed, which crowdsourcing platform was used to collect data?
    - Identifies which platforms are most prominently used, and may identify variance in practices between platforms.

A.5 Geographic distribution of papers across venues

Table 9 demonstrates the total number of AI papers assessed in this work across geographies and venues. As shown, the first authors from over half of the papers identified as using crowdsourcing across the venues derived from the US, with 13 percent from the European Union, and 9 percent from China.

This provides context to the results that follow and shows a heavy US-bias to these venues, and the study as a whole.

Table 9: Geographic distribution of AI crowdsourcing papers across publication venue between 2018 and 2021

Geography	NeurIPS	ICLR	Springer AI	Total
United States	86 (60%)	69 (66%)	131 (48%)	286 (55%)
European Union	11 (8%)	4 (4%)	52 (19%)	67 (13%)
China	17 (12%)	11 (11%)	21 (8%)	49 (9%)
United Kingdom	6 (4%)	3 (3%)	10 (4%)	19 (4%)
Canada	2 (1%)	5 (5%)	11 (4%)	18 (3%)
South Korea	6 (4%)	6 (6%)	4 (1%)	16 (3%)
Switzerland	4 (3%)	0 (0%)	9 (3%)	13 (2%)
Rest of World	11 (8%)	6 (6%)	37 (13%)	54 (10%)
Total	143 (100%)	104 (100%)	275 (100%)	522 (100%)

A.6 Platform Use

Tables 10 and 11 show the breakdown of platforms used across ICLR and NeurIPS papers meeting study criteria.

The table demonstrates the overwhelming reliance on MTurk for crowdsourced data, and frequency of papers choosing not to disclose platforms used.

Table 10: NeurIPS: Crowdsourcing papers' platform use between 2018-2021

Year	MTurk	Other	Unknown
2021	23 (48%)	4 (8%)	21 (44%)
2020	14 (41%)	0 (0%)	20 (59%)
2019	25 (68%)	1 (3%)	11 (30%)
2018	15 (63%)	0 (0%)	9 (38%)
Total	77 (54%)	5 (3%)	61 (43%)

Table 11: ICLR: Crowdsourcing papers' platform use between 2018-2021

Year	MTurk	Other	Unknown
2021	26 (65%)	1 (3%)	13 (33%)
2020	15 (65%)	0 (0%)	8 (35%)
2019	18 (60%)	0 (0%)	12 (40%)
2018	9 (82%)	0 (0%)	2 (18%)
Total	68 (65%)	1 (1%)	35 (34%)