# Reliance and Automation for Human-AI Collaborative Data Labeling Conflict Resolution

MICHELLE BRACHMAN, IBM Research, USA
ZAHRA ASHKTORAB, IBM Research, USA
MICHAEL DESMOND, IBM Research, USA
EVELYN DUESTERWALD, IBM Research, USA
CASEY DUGAN, IBM Research, USA
NARENDRA NATH JOSHI, Adobe, USA
QIAN PAN, IBM Research, USA
AABHAS SHARMA, IBM Research, USA

Human data labeling with multiple labelers and the resulting conflict resolution remains the norm for many enterprise machine learning pipelines. Conflict resolution can be a time-intensive and costly process. Our goal was to study how human-AI collaboration can improve conflict resolution, by enabling users to automate groups of conflict resolution tasks. However, little is known about whether and how people will rely on automation during conflict resolution. Currently, automation commonly uses labelers' majority vote labels for conflict resolution, as the top chosen label by most labelers is often correct. We envisioned a system where an AI would assist in finding cases where the labeler majority vote was wrong and where automation is supported for batches or groups of conflicts. In order to understand whether humans could use labeler and AI information effectively, we investigated how and when users rely on labeler and AI information and on automated group conflict resolution. We ran a study with 144 Mechanical Turk workers. We found that automation increased users' accuracy/time, use of automated conflict resolution was relatively similar regardless of whether the automation was based on labeler or AI selected labels, and providing labeler and AI selected labels may reduce inappropriate reliance on automation.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **Collaborative and social computing**.

Additional Key Words and Phrases: data labeling, conflict resolution, reliance, trust, automation

**ACM Reference Format:**
Michelle Brachman, Zahra Ashktorab, Michael Desmond, Evelyn Duesterwald, Casey Dugan, Narendra Nath Joshi, Qian Pan, and Aabhas Sharma. 2022. Reliance and Automation for Human-AI Collaborative Data Labeling Conflict Resolution. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 321 (November 2022), 27 pages. https://doi.org/10.1145/3555212

Authors' addresses: Michelle Brachman, michelle.brachman@ibm.com, IBM Research, 314 Main Street, Cambridge, MA, 02139, USA; Zahra Ashktorab, zahra.ashktorab1@ibm.com, IBM Research, 1101 Kitchawan Rd PO Box 218, Yorktown Heights, NY, 10598, USA; Michael Desmond, mdesmond@us.ibm.com, IBM Research, 1101 Kitchawan Rd PO Box 218, Yorktown Heights, NY, 10598, USA; Evelyn Duesterwald, duester@us.ibm.com, IBM Research, 1101 Kitchawan Rd PO Box 218, Yorktown Heights, NY, 10598, USA; Casey Dugan, cadugan@us.ibm.com, IBM Research, 314 Main Street, Cambridge, MA, 02139, USA; Narendra Nath Joshi, joshinarendranath@gmail.com, Adobe, 345 Park Ave, San Jose, CA, 95110, USA; Qian Pan, qian.pan@ibm.com, IBM Research, 314 Main Street, Cambridge, MA, 02139, USA; Aabhas Sharma, aabhas.sharma@ibm.com, IBM Research, 314 Main Street, Cambridge, MA, 02139, USA.


2573-0142/2022/11-ART321 $15.00
https://doi.org/10.1145/3555212

## 1 INTRODUCTION

With the expanding use of AI and Machine Learning comes the increasing need for labeled data in both enterprise and research contexts. In many cases, this data must be labeled by humans manually, a time-consuming and error-prone process, which frequently leads to conflicts [17, 35, 54]. *Conflicts* are instances where labelers have selected different labels for the same data item. Conflicts in data labeling or annotation are often the result of unclear boundaries between categories, ambiguity of the data being labeled, and differences in individuals' understandings of concepts or data [38, 79]. These issues make conflicts challenging to resolve, leading to manual conflict resolution being a lengthy and costly process.

Some argue that conflict resolution isn't always necessary or possible, due to a lack of a singular correct label, high subjectivity, or ambiguity of the task [26, 28, 65]. However, many contexts still depend on an accurate and singular set of ground truth data labels. Researchers have studied ways to resolve conflicts automatically, such as simply using the *majority vote* of the labelers, or more complex algorithms [13, 45, 64, 73, 80]. Yet, conflict resolution is often still performed manually by humans, due to the fact that the hardest conflicts to resolve are often on boundaries, making them sub-optimal for machines. Recent work has found that human-machine teaming can have improved performance over either individually [9].

Our goal was to understand whether human-AI collaboration could improve efficiency and accuracy of conflict resolution, by combining automated approaches with the expertise of human manual labeling. Ideally, we wanted to: 1) provide a way to resolve groups of conflicts automatically and help the user to decide when to use this automation, and 2) support users in making manual decisions for the remaining more challenging conflicts.

Automated group conflict resolution could enable faster resolution of more straightforward conflicts so that users can focus their time and energy on the more challenging conflicts. Not all conflicts are equally hard to resolve. For example, some may have a 75%/25% split between labeler votes for two labels (likely more straightforward), while others could have four labels each with 25% of labeler votes (likely more challenging). Yet, creating heuristic cutoffs for hard and easy conflicts may not be valid in all contexts. Thus, our tool groups conflicts and enables users to choose to automatically or manually resolve groups of tasks.

Even when most labelers agreed on a label, the labelers do not always get it right. However, due to psychological effects like the *bandwagon effect* [69], there is a risk that users might go along with the majority of labelers for a conflict regardless of correctness. Having another opinion might spur users to think more critically. Thus, we wanted to explore whether the combination of human and AI information could reduce inappropriate reliance on majority vote during conflict resolution. However, there are also concerns about trust and inappropriate reliance in automation [14, 16, 27, 55, 56, 81]. We wanted to understand the interplay between these two sources of information for conflict resolution.

In order to better understand whether automation could enable faster conflict resolution for easier conflicts and manual resolution for more challenging conflicts, as well as whether providing AI information alongside labeler majority vote could reduce inappopriate reliance, we had three research questions:

- RQ1: How do automation and labeler-AI information affect $\frac{accuracy}{time}$, time, and accuracy?
- RQ2: How do users rely on labelers, AI predictions, and the combination of labelers and AI predictions for conflict resolution?
- RQ3: How do users decide whether to rely on automated conflict resolution?

We designed, developed, and evaluated a conflict resolution tool that enables manual and automated group conflict resolution with labeler and/or AI suggested labels. We ran a study with

144 Amazon Mechanical Turk workers, who resolved a set of 100 real data labeling conflicts that resulted from a crowdsourced data labeling task. Our study had two factors: 1) only manual conflict resolution vs. the choice of manual and automated conflict resolution, and 2) the type of suggested labels and information available (labeler majority vote, AI predicted probability, or both). We asked users about their trust in the labelers and the AI, as well as how they made decisions about whether to use the optional automated group conflict resolution. We found that users who were offered automated group resolution had higher accuracy/speed than those with only manual resolution. Users with both labeler and AI suggested labels relied on the system less when the system's label choice was wrong, but also relied on the system less when it would have accurately resolved a conflict. The reasons users chose to resolve manually and by group provide insight into how they made decisions about reliance.

Our contributions are:

- An empirical investigation of the impact of human, AI, and the combination of human and AI information on data labeling conflict resolution
- A study that provides insight into accuracy, speed, reliance, and user decision making for automated conflict resolution
- A set of design recommendations for improving appropriate reliance on automation in conflict resolution and more broadly: 1) balance reliance and time by providing system direction, and 2) support interactive hedging

## 2 RELATED WORK

Our work builds upon and contributes to existing research on resolving and reducing conflicts, support for data labeling, and human-AI interaction.

### 2.1 Resolving and Reducing Conflicts

Researchers have explored ways to support reducing and resolving conflicts in contexts where work is completed by multiple workers and aggregated, like data labeling and data annotation for machine learning algorithms, crowdwork, and decision making.

Much of the work surrounding conflict resolution and crowdwork response aggregation has focused on discussion, explanation, and argumentation. Structured adjudication, a workflow in which users give an explanation with reference to related guidelines, supported users in reaching consensus for medical data [59]. Similarly, having crowdworkers provide rationales for decisions in a tool called Microtalk improved worker accuracy by 20% [24]. Contextual argumentation even further improved accuracy over the basic one-shot argumentation of Microtalk and enabled workers to reach near-perfect accuracy [20]. This aligns with the research finding that the amount of discussion can end up affecting the resolvability of a conflict [61]. Further, providing explanations rather than just numerical confidence can even prompt users to resolve conflicts more efficiently and appropriately [60].

One way to reduce conflicts in the first place could be to improve labelers' consistency and label use. A structured labeling system enabled people to form groups as they labeled, which led to improved consistency with reduced speed [43]. In the Revolt system, crowdworkers work collaboratively to deal with conflicts in real-time, by giving explanations in cases where other crowdworkers disagreed and generating new labels [19]. Revolt reduced the upfront cost of creating label guidelines. Another way to avoid needing conflict resolution is to accept multiple responses and use a distribution, which reduced time and improved accuracy [22]. Accepting multiple labels is especially useful when data are ambiguous, making applying a single label highly error-prone

and potentially invalid. However, a distribution of labels doesn't work for all machine learning contexts.

Much of the work around conflict reduction and resolution has focused on human discussion and workflows. Our aim was to reduce time spent resolving conflicts by enabling users to focus their time manually resolving the hardest conflicts and use automation for the rest. Little work has addressed decision making for automated human-AI conflict resolution.

### 2.2 AI-Supported Data Labeling and Qualitative Coding

Our work closely relates to AI-supported tools and methods for labeling data (the process that leads to conflict resolution), as well as qualitative coding, a qualitative research method that involves applying codes, like labels, to textual data.

*2.2.1 Data Labeling.* Researchers have looked at several ways to improve and speed up data labeling more generally with human-AI collaboration, such as with active learning and visual information. Active learning is one way to increase intelligent support for data labeling, in which a machine learning model determines which data needs to be labeled next [51]. Researchers have also begun to study explanations in the context of Active Learning and found that explanations supported trust calibration [30]. However, visual-interactive labeling, in which users select which data to label, can have better outcomes than active learning and the two could potentially be used together [11]. Label and Learn provides a visualization that shows how labels are affecting the classifier, which improves the labeling experience and helps users understand the effect of labeling [66]. Researchers also found that highlighting words to bring attention to important parts for labelers increased the efficiency of data labeling [21]. We build upon the ideas of automation from human-AI collaborative data labeling and explore how users can leverage automation for conflict resolution.

*2.2.2 Qualitative Coding.* Like data labeling for machine learning, qualitative coding can be a tedious and lengthy process. Though automation may be helpful and desired for qualitative coding, it is rarely available [48]. Users want to be able to automate the extension of coding to large amounts of data while maintaining transparency and control over the process [48]. Several systems have recently been developed that can make accurate qualitative coding suggestions to annotators [40, 48], though we still know little about how people use them. The Cody tool, which provides semi-automated support for qualitative coding through the use of rule-based matching and machine learning, improved inter-rater reliability [58]. Both rule-based and machine learning methods can provide high accuracy code suggestions, though both have drawbacks: rules require extra effort and machine learning requires labeled data [23]. Rather than directly automating qualitative coding, researchers have also worked to develop ways to support qualitative coding, such as using visual analytics [25]. Similar to data labeling, we know little about users' use of human and/or AI information, and how they might use that information to automate groups of tasks at once.

### 2.3 Human-AI Interaction

Our work fits broadly in to the context of human-AI interaction [2, 70, 74], from which we drew inspiration. In particular, our work relates most closely to research focusing on optimizing human-AI collaboration, and trust and reliance on AI.

Researchers have looked at how to optimize and understand human-AI teams [5, 7, 8]. They found that having the highest accuracy AI is not necessarily the best for human teammates, because the whole team needs to be taken into account when modeling expected accuracy [5]. Additionally, updating the AI based on new information that improves the AI's performance can negatively affect the overall team performance [8]. Our work also looks at how we can utilize a combination

of human and machine intelligence to optimize accuracy and workload in the context of conflict resolution.

One way to help people cooperate better with an AI teammate is to improve their mental model. A study of mental model formulation in a collaborative game showed that overestimation of an AI's abilities can lead to inappropriate mental model formulation [29]. One way to improve users' mental model is by increasing the simplicity and reducing the number of features needed to explain the error boundary of a model [7]. Though we don't specifically address mental models, we asked users about how they decided to use automation, as well as about their perceptions of the accuracy of themselves and the system, which can give us some information about how users perceived the system.

Often, human-AI collaboration and teaming focuses on a system that is entirely algorithmically determined, but we consider automation based on labeler and AI content. Several papers have begun to investigate the differences in how people react to humans compared to an AI within automated systems. In a study of a movie recommender system, human explanations were compared to system-generated explanations and found to be rated higher and generated more trust in the recommendations [44]. In a cooperative game setting, people had more positive views of a human partner compared to an AI partner, though it didn't affect outcomes [4]. This paper contributes further to the understanding of perceptions of human- and AI-based automation, as well as the combination of both human and AI content.

### 2.4 Decisions with Human and AI Influence

Prior work has looked at how people trust and/or rely on AI, as well as how people are influenced by other people (the bandwagon effect).

*2.4.1 Relying on an AI: Trust and Reliance.* Trust and reliance are important issues in human-AI interaction, as trust affects users' interactions with intelligent systems. Trust and reliance are of special concern in AI-assisted decision making, in which people need to make important decisions in areas like medicine, finances, flight, or law. Appropriately evaluating the performance of systems and calibrating reliance correctly is an essential and non-trivial task for humans [32].

Early work suggested that explanations might address issues users had with proper calibration of trust [31]. However, explanations do not always reduce overreliance [9, 37]. Instead, cognitive forcing functions, which encourage users to actively engage, may be more helpful in reducing overreliance on AI compared to explanations [15]. Relatedly, updating pilots with confidence information improved trust calibration over static reliability information [49].

Others have found factors that affect trust, such as attractiveness of the AI agent [78], perceived system performance [75, 76], and transparency [34, 39, 71, 77]. Surprisingly, reliability didn't have more of an impact on trust than attractiveness [78]. Yet, the reliability and accuracy of a system can have a significant impact on users' trust, as users are able to perceive accuracy of systems and moderate their trust accordingly [75, 76]. Further, surprises in the accuracy of a system can erode trust [75], but transparency can moderate the effects of expectation mismatch on trust [39].

Most work has looked at trust and reliance on individual tasks and focused on full automation. In this work, we explore how two factors affect trust and reliance on the system for conflict resolution: 1) having labeler, AI, or a combination of labeler and AI information, and 2) manual conflict resolution vs. a choice of manual or automated conflict resolution.

*2.4.2 Relying on others: The Bandwagon Effect.* The *bandwagon effect* is a well-studied psychological effect in which people are impacted by the opinions of others. This effect has been widely studied in a variety of contexts, such as voting [10], branding [63], and more relevantly, computer-supported collaborative interfaces [36, 41, 67–69]. For example, explicit ratings of news articles impacted how

Table 1. A summary of the 6 study conditions.

| Resolve | Group Resolve Options | Support Information | Condition |
|---|---|---|---|
| Choice (Manual or Group) | Labelers' top voted label | Labeler agreement | L-C |
| | AI top prediction | AI prediction | AI-C |
| | Labelers' top voted label & AI top prediction | Labeler agreement & AI prediction | LAI-C |
| Manual | | Labeler agreement | L-M |
| | | AI prediction | AI-M |
| | | Labeler agreement & AI prediction | LAI-M |

many news articles people read and how long they spent reading them [41]. Bandwagon influence also increased users' intentions to purchase items and perceptions of credibility and quality [69]. When considering the impact of human influence on credibility, the source of the information may matter. Crowdsourcing reduced people's perceptions of the trustworthiness and completeness of health information [36]. When users have access to both human influence and AI influence, an AI can be as influential as a human [72]. Our work extends prior work on the bandwagon effect by studying the influence of human and AI choices in the conflict resolution space. Further, we compare not only the impact of having human *or* AI information, but what it means to have both information at the same time.

## 3 CONFLICT RESOLUTION TOOL

The Conflict Resolution Tool (CRT) (Figure 1) is designed to facilitate the resolution of labeling conflicts which naturally stem from group labeling activities. The CRT enables us to study the impacts of automation and labeler and/or AI support information on conflict resolution. Our study had six variants of the tool for our six conditions, varying along two axes: 1) labeler, AI, or labeler and AI support information, and 2) manual or choice of manual or group resolve, as shown in Table 1.

### 3.1 Support Information

CRT presents a sequence of labeling conflicts and each conflict consists of an example (Figure 1, box 1) and a set of proposed labels (Figure 1, box 2). The user's goal is to select the correct label from the set, thus resolving the conflict. The system helps the user to choose the correct label by presenting a representation of labeler agreement, and /or an AI predicted probability of the correct label, depending on the condition.

*3.1.1 Labeler Agreement.* Providing labeler vote agreement is the most basic way to perform conflict resolution. CRT analyzes the labels applied to each conflict and generates a distribution of applied labels which is displayed alongside the set of proposed labels (Figure 1, box 3). The distribution shows the percentage of labelers that selected each label (agreement) and can help the user to understand how the overall set of labelers interpreted the example. In two conditions, participants only had access to this labeler agreement information (L-C and L-M) and in two conditions the labeler agreement information was presented alongside an AI prediction (LAI-C and LAI-M).

*3.1.2 AI Prediction.* We decided to provide AI predicted labels with the aim of reducing overreliance on the labelers by providing another perspective. A label probability distribution, predicted by an

AI trained on all non-conflicting labeled data, is also displayed alongside each example (Figure 1, box 4). This distribution presents the probability of each label being correct from the point of view of a statistical model. The AI prediction is converted to a percentage value assigned to each label for consistency. The AI prediction was explained with this text: *The probability the AI calculated that each label is correct.* In two conditions, participants had access to both the labeler agreement and the AI prediction (LAI-C and LAI-M). Two AI conditions included only the AI predicted probability information as a comparison to increase generalizability of our outcomes (AI-C and AI-M), even though we would not expect a conflict resolution system to leave out labeler votes. The AI training is described in Section 4.4.

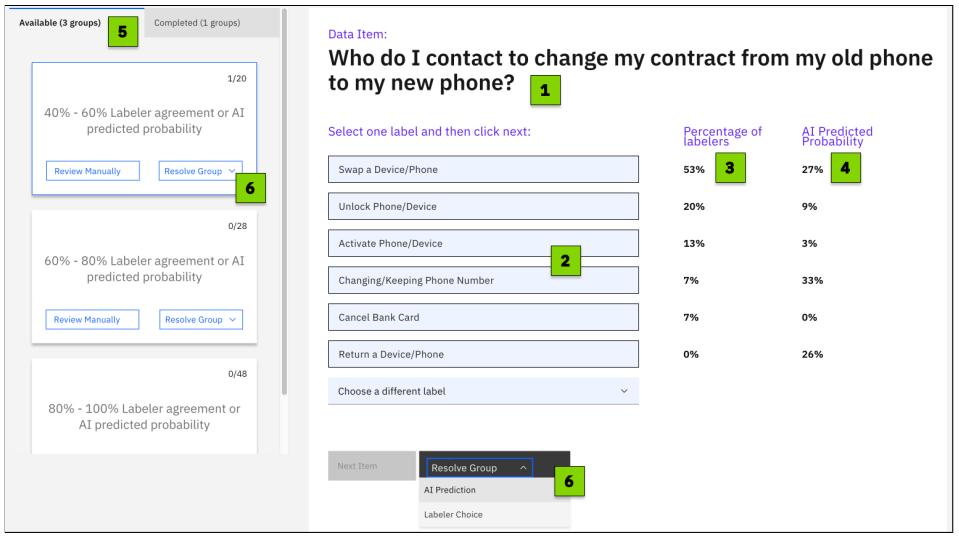


Fig. 1. Conflict Resolution tool for the LAI-C condition . (1) The example that is labeled in a conflicting manner. (2) The set of labels that define the conflict. (3) The distribution of labelers that applied each label. (4) AI probability distribution label correctness. (5) Groups of conflicts characterized by difficulty. (6) The group resolution feature, which enables automatic resolution of conflict groups by applying system labels.

## 3.2 Grouping

In order to enable automated conflict resolution, in which users can resolve sets of conflicts, conflicts are organized into groups (Figure 1, box 5). Each group is characterized by thresholds of aggregated labeler agreement, AI predicted probability, or a combination of both metrics depending on system configuration. Intuitively, the group system partitions conflicts into sets of varying difficulty. Groups with relatively high labeler agreement and AI predicted probability are likely to be easily resolvable, while those with low labeler agreement and AI predicted probability are likely to be more difficult to resolve and may require greater manual scrutiny.

## 3.3 Group Resolution

A central feature of CRT is *Group Resolve*. When Group Resolution is available, groups of conflicts can be resolved by manually resolving each conflict individually, or by resolving the entire group automatically using group resolve (Figure 1, box 6). Once a user chooses to group resolve, a *group resolve modal* appears, as shown in Figure 2. This view was designed to allow the user to check

all of the AI predicted and/or labeler top choice labels in a group before agreeing to the system resolving all of the conflicts in that group.

Three of our six conditions included the Group Resolution feature, where users had the *choice* to group resolve or manually resolve (L-C, AI-C and LAI-C). Group resolve applies a system selected label to all remaining unresolved conflicts in the group, and may be applied via the 'Labeler Choice' or 'AI Prediction' options, depending on which is available. In the case of group resolution using 'Labeler Choice' the system will automatically apply the label with the highest labeler agreement (available in L-C and LAI-C). Alternatively, in the case of group resolution using 'AI Prediction', the highest probability label predicted by the AI is applied (available in AI-C and LAI-C).

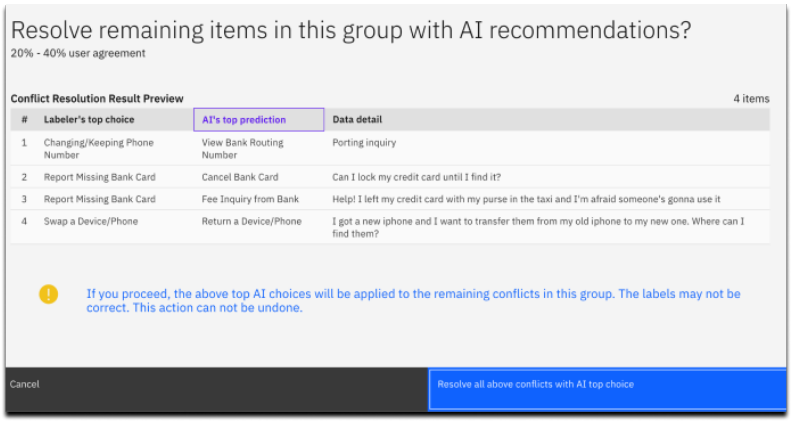


Fig. 2. The group resolve modal for LAI-C. When users selected to resolve a group, they saw this modal interface overlaid over the CRT. It lists the data items and labels that would be applied if the user confirms. For the LAI-C condition, the user can compare the Labeler and AI choices in this modal. In the AI-C and L-C conditions, this modal only showed the appropriate AI predicted label or top choice label, respectively.

We included the group resolve modal (Figure 2), that allows users to check the labeler choice and/or AI predicted label for all items in a group, based on literature on swift trust and the benefits of being able to verify the results of trust [33, 52]. Swift trust is trust that needs to be established quickly, such as between teammates put together for a project. Trusting automation often has the same potential problems as in swift trust between humans- there isn't necessarily the time to gain trust through experiences. One way to support swift trust is to enable people to 'hedge', meaning that they can *trust but verify*. Our review step provides users the option to hedge in CRT.

## 4 STUDY DESIGN

We wanted to understand when and why users trusted human labels and/or AI predictions to resolve groups of labeling conflicts. We also wanted to understand how the ability to automatically resolve groups of conflicts affected accuracy and efficiency compared to manual resolution. To this end, we designed a 2x3 between-subjects study which employed 144 Mechanical Turk workers. Each worker resolved a set of 100 labeling conflicts using variants of the conflict resolution tool described in section 3.

### 4.1 Conditions

The study consisted of six conditions: three where participants had the option to resolve groups of labeling conflicts automatically using group resolve (See 3.3), and three where participants were limited to resolving conflicts manually one at a time.

*4.1.1 Choice Conditions.* The choice conditions gave participants the option to automatically resolve conflicts using the 'Resolve Group' button (Figure 1, box 6), in addition to standard manual resolution. The difference in the three choice conditions was *how* users could choose to resolve groups, and what supporting information was available: L-C participants had labeler agreement information and labeler top choice group resolve, AI-C had AI predicted probability and AI group resolve, and LAI-C had both labeler and AI information and group resolve choices. The study compares the conditions **with** group-resolve to the conditions **without** group-resolve to answer whether group resolution improves speed and accuracy of conflict resolution.

*4.1.2 Manual Conditions.* Participants in the three manual conditions were limited to manually resolving all 100 conflict resolution tasks. The three conditions differed in the supporting information available: L-M had labeler agreement information, AI-M had only AI predicted probability, and LAI-M had both labeler and AI information.

### 4.2 Protocol

Participants were first presented with some general instructions and provided informed consent. They were then directed to the conflict resolution interface to resolve the 100 labeling conflicts. The particular features available in the interface were dependent on the condition to which the participant was assigned (see Table 1). Upon completion of the resolution task participants were directed to a survey. The survey collected demographic information and information about their motivations and trust in the AI predictions and human labelers. We asked participants in the choice conditions about their decision about when to use group resolve or not: "If you resolved whole groups of conflicts, how did you decide when to do so? If you only manually resolved tasks individually, please describe why you chose not to resolve groups of conflicts." We adapted our trust questions from a recently designed and validated trust survey, focusing on reliability of the system, understandability of the system, familiarity with similar systems, propensity to trust, and trust in the automation [42]. The Appendix Tables 6, 7, and 8 provide all survey questions.

### 4.3 Dataset

We used a 240 example, short text intent dataset with 21 labels to facilitate the study. Originally used to train a banking and telecommunications chatbot, we considered the dataset to be both a realistic natural language dataset that a corporation might need labeled and accessible to Amazon Mechanical Turk workers (see Table 2 for examples of items and labels in the dataset). Further, for this dataset, a prior study [3] of data labeling resulted in conflicts across 109 labelers for all 240 items, demonstrating that labelers are prone to conflicting opinions in this dataset. The data from this prior study allows us to study conflict resolution on a dataset of **real data labeling conflicts**, rather than ones we have artificially created for this study.

We chose a random set of 15 labelers from the original set of 109 to use in this study as this is a more realistic number of labelers who might work on a real dataset. Of the random 15 labelers, there were 240 labeling conflicts, again demonstrating the high level of conflict between labelers in this dataset. We needed to limit the number of labeling conflicts participants would need to resolve due to the time constraints of this study. To do so, we selected the 100 conflicts of the 240 with the lowest labeler agreement i.e. where the fewest number of labelers agreed on a label. As such we

Table 2. Example items and labels from the dataset. Note, as discussed in other literature [53], ground truth labels are also subjective. The first item may be an easier conflict, as 7/15 labelers chose a single label. The second may be more difficult, as the labelers are more spread across labels.

| Item | Labels provided by labelers | Ground Truth |
|---|---|---|
| What steps can I take when I lose my credit card? | Report Missing Bank Card (7), Replace Bank Card (3), Cancel Bank Card (2),Activate Phone/Device (1), Changing/Keeping Phone Number (1), View Bank Routing Number (1) | Report Missing Bank Card |
| Will my text messages come over to my new phone when I migrate to it? | Swap a Device/Phone (4), Activate Phone/Device (4), Changing/Keeping Phone Number (2), Report Missing Bank Card (2), Activate Roaming for Device (2), Activate Bank Card (1) | Swap a Device/Phone |

chose the 100 most challenging labeling conflicts for this study. Within this set of conflicts, the label chosen by the majority of labelers was correct 81% of the time.

### 4.4 AI Predictions

After pulling out the 100 items with the most labeler conflict as the conflict resolution tasks for this study, we used the remaining 140 data examples that were not included in the study data set and their ground truth labels as training data for the AI within the CRT. The idea was that these 140 data examples and their ground truth labels would simulate a scenario where a set of labelers had agreed upon labels for a set of data. The 140 items we chose for this training data were the items from the previous data labeling tasks in which the highest number of labelers agreed. Hence, we expect that if there were to be full agreement on a set of items, this set would approximate that scenario. The 140 training data examples were encoded into fixed size vectors using the Universal Sentence Encoder [18]. We then trained a neural network [1] on the 140 training examples, and used it to predict label probabilities for the 100 study examples. We tried LabelSpreading (MA(mean accuracy)=0.78), LogisticRegression (MA=0.87), RandomForestClassifier (MA=0.82), and MLPClassifier (MA=0.88), finding that the MLP classifier was the most accurate using 3-fold cross validation. The model was 71% accurate at predicting the correct label on the 100 conflicts participants saw as tasks. Thus, our AI predicted probabilities were designed to be realistic, rather than artificially set. The model was kept stable throughout the study and not retrained or updated based on conflicts resolved by users during the study task. This ensured that AI predictions were consistent across all users.

### 4.5 Groups

Regardless of condition, all 100 labeling conflicts were organized into four groups in the interface. Depending on a particular condition, groups of conflicts were split on either thresholds of labeler agreement (L-C, L-M), AI prediction (AI-C, AI-M), or a combination of both (LAI-C, LAI-M). A total of four groups were chosen to make it worthwhile for participants to reason about whether to resolve conflicts automatically using the group resolve feature. Our aim was for the groups to not be so large that users would be dissuaded from automating them, but also generally not so small that they were not worth automating. For the purposes of consistency, we also aimed for thresholds

[1] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

Table 3. Group segmentation for the different study conditions.

| Condition | Split on | Group0 | Items | Group1 | Items | Group2 | Items | Group3 | Items |
|---|---|---|---|---|---|---|---|---|---|
| L-C | Labeler agreement | 20-30% | 11 | 30-40% | 32 | 40-50% | 22 | 50-60% | 35 |
| AI-C | AI predicted probability | 20-40% | 9 | 40-60% | 15 | 60-80% | 28 | 80-100% | 48 |
| LAI-C | max(AI predicted probability, labeler agreement) | 20-40% | 4 | 40-60% | 20 | 60-80% | 28 | 80-100% | 48 |

that provided somewhat similar group sizes across conditions, while still using the real data. To implement the groupings and distribute the conflicts somewhat evenly, the range of available values were split into four, as shown in Table 3. Within each group, participants saw conflicts in increasing order of labeler agreement and/or AI predicted probability, such that any unresolved conflicts in a partially resolved group would all be more likely to have correct automated results than those previously seen. Otherwise, resolving by group would have not been a sensible action to take.

### 4.6 Participants

We recruited 144 participants from Amazon Mechanical Turk, 24 per condition. The study was conducted with Mechanical Turk workers so we could support a bonus structure to create vulnerability, a necessary component in order for trust and reliance to be relevant [46]. Furthermore, for our context, banking and telecommunications chatbot data, the average person should be able to make reasonable decisions to resolve data labeling conflicts. We required that participants had completed at least 500 HITs and had a 95% acceptance rate for their HITs. We paid all participants 4 USD for completing the HIT, which exceeds federal minimum wage. Participants could also receive a bonus based on time and accuracy: if participants had at least 80% accuracy, they could receive a 50 cent bonus for completion in less than 14 minutes, 25 cent bonus for completion in 14-17 minutes, and a 10 cent bonus for completion in 17-20 minutes.

*4.6.1 Pilot.* We had 6 participants in our pilot, two in each of the choice conditions to test the functionality and incentive structure. Based on the pilot, we modified our incentives from a bonus only for accuracy to a bonus based on accuracy and time. This accuracy and time-based bonus aligns better with the goals and context of our research. Real users are likely short on time, but still need high accuracy on their data labeling.

## 5 METRICS AND ANALYSIS

We collected and analyzed accuracy/time, reliance, user behaviors and user attitudes.

### 5.1 Behavioral Data

Due to the quality variations in Amazon Mechanical Turk work, we filtered out users with outlier time and/or accuracy. To accomplish this, we ran our analysis for behavioral data only with those users' data where time and accuracy were within two standard deviations of the mean. This removed 1 data point that was an outlier in time and accuracy, 2 that were time outliers, and 13 data points that were accuracy outliers, primarily from the manual resolution conditions. After this process, our analysis of behavioral data included 128 participants: 23 from condition L-C, 23 from AI-C, 23 from LAI-C, 18 from L-M, 20 from AI-M, and 21 from LAI-M.

*Accuracy/Time*: We collected $\frac{accuracy}{time}$ to capture users' ability to balance the two goals.

*Reliance:* We captured users' reliance on the system by computing how often they agreed with the labels suggested by the system. This included both explicit manual agreement or implicit agreement through group resolve.

*Behaviors*: We captured users' interactions with the system, including their use of group resolve, canceling group resolve, and time spent completing these actions.

We analyzed users' reliance, accuracy/time, behaviors, and survey Likert scale responses quantitatively. Where relevant, we used Fisher's exact test, ANOVA, and Kruskal-Wallis with post-hoc Dunn test and Holm correction for multiple comparisons.

### 5.2 Survey Data

In our survey, we collected participants' self-reported responses about their trust in the support on Likert Scales from 1-5, their goals in completing the tasks, and their decisions about auto-resolving. For our survey Likert Scale responses, we found that only questions about trust in the automation (AI trust: $\alpha = 0.65$, labeler trust: $\alpha = 0.86$) and familiarity ($\alpha = 0.77$) with the system were reliable using Cronbach's alpha, so we only discuss those responses.

We used thematic analysis [12] to analyze users' survey responses to when they chose to manually or group resolve in the choice conditions (70 responses due to two missing survey responses). We chose to use thematic analysis because it is a flexible method that can be used once all of the data is already collected. Two authors started by reading the responses and performing open coding. The authors then discussed and agreed upon a first version of the codebook. They then re-coded individually based on the agreed upon codebook. Finally, they discussed disagreements and came to agreement on application of the codes.

## 6 RESULTS

Our goal was to better understand human-AI collaboration for automated conflict resolution. Towards this goal, we report on our three research questions.

### 6.1 RQ1: How do group resolve and different information types affect $\frac{accuracy}{time}$, time, and accuracy?

In an ideal scenario, users would balance time and accuracy, such that they save time, but also maintain or increase accuracy. To understand whether our participants succeeded in this trade-off, we compared $\frac{accuracy}{time}$ across the conditions. Due to the non-normality of data, we used a root transform, followed by a two-way ANOVA to look at the effects of our two factors on $\frac{accuracy}{time}$: 1) manual vs. choice of manual or group resolve and 2) AI, labeler, or labeler and AI information. Participants with the choice of manual or group resolve had significantly higher $\frac{accuracy}{time}$ ($M = 24, SD = 35$) than those with only manual resolve ($M = 5, SD = 2.4, F(1, 122) = 41, p < 0.001$), as shown in Figure 3. We did not find a significant difference between participants who had the Labeler, AI, or Labeler+AI information for the interaction effect. While it is not surprising that those who could resolve whole groups of conflicts at once completed the tasks faster, this result could have been different if users rarely relied on group resolve due to lack of trust in the AI or if they relied on group resolve in ways that had much lower accuracy than manual resolution.

We found a significant difference in time between those who had group resolve ($M = 10\ mins, SD = 7.8\ mins$) compared to those who only had manual resolve ($M = 19.1\ mins, SD = 8.7\ mins$) ($F(1, 122) = 44, p < 0.001$) using an ANOVA with a root transform. We did not find other significant differences in time based on having Labeler, AI, or Labeler+AI information, or in the interaction between the availability of group resolve and type of information available. While we expected

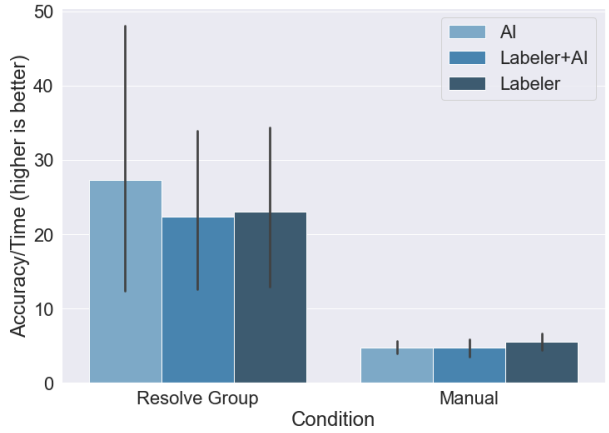


Fig. 3. Accuracy/Time across conditions.

group resolve to have an impact on time, we also expected group resolve to have an impact on accuracy. However, time was reduced by nearly half with the use of group resolve **without** a significant reduction in accuracy.

We found a significant overall difference in accuracy between Labeler, AI and Labeler+AI groups ($F(2, 122) = 8.8, p < 0.001$), as well as the interaction of information type with the availability of group resolve ($F(2, 122) = 3.1, p < 0.05$) using an ANOVA with a cube transform. We used a TukeyHSD follow-up for the main and interaction effects. Those who had labeler information ($M = 82\%, SD = 8\%$) had higher accuracy than both those with AI information ($M = 76\%, SD = 9\%, p < 0.01$) and those with Labeler and AI information ($M = 74\%, SD = 13\%, p < 0.001$). We found three interaction effects: LAI-M participants performed worse than L-M participants ($p < 0.001$) and than L-C participants ($p < 0.05$), and L-M participants performed better than AI-C participants ($p < 0.05$). We did not find a significant effect on accuracy between those who had group resolve ($M = 78\%, SD = 8\%$) and those who only had manually resolve ($M = 77\%, SD = 13\%$). The differences in accuracy are likely due to the fact that the Labeler information had higher accuracy (81%) than AI predictions (71%). Thus, when participants relied heavily on the provided information, they performed better in the Labeler condition and worse in the AI condition. Our next sections explore participants' reliance on the provided information, especially in the cases where participants had both labeler and AI information.

### 6.2 RQ2: How do users rely on labelers and AI predictions for conflict resolution?

Correctly calibrating reliance is critical for effective use of human-AI systems. We first report overall use of group resolve, which tells us at a high level whether users were willing to rely on the system, followed by appropriate and inappropriate reliance.

*6.2.1 Resolve Group Use.* Somewhat surprisingly, we did not find differences in group resolve use between conditions for the 69 participants who had access to group resolve. We analyzed the average number of items participants resolved using group resolve and did not find a significant difference based on the type of information available ($p > 0.1$).

Overall, 74 % of participants (51/69) resolved by group at least once in the L-C, AI-C, and LAI-C conditions. Of the remaining participants, 14% (10/69) selected group resolve but cancelled, and 12% (8/69) manually resolved all items without ever selecting group resolve. Of the 51 participants

Table 4. Number of participants who exhibited each of the three resolve group behaviors.

| Condition | 100% group resolve | 100% manual resolve | Manual & group |
|---|---|---|---|
| L-C | 5 | 9 | 9 |
| AI-C | 4 | 5 | 14 |
| LAI-C | 5 | 4 | 14 |

who resolved groups, 14 were in the L-C condition, 18 in the AI-C condition, and 19 in the LAI-C condition, which was not significantly different ($p = 0.18$).

Among all choice conditions (L-C, AI-C, and LAI-C), we noticed three distinct behaviors: 1) one set of users chose to use group resolve for all tasks, 2) one set of users chose to resolve all tasks manually, and 3) one set of users applied a combination of manual and group resolution. Participants in each of the three conditions exhibited each of the behaviors (see Table 4). Our thematic analysis of survey responses provides some insight into why this happened (see Table 9). Fifteen participants described that they wanted to resolve the tasks manually because they felt that their own responses would be more accurate. On the other hand, seven participants responded that they wanted to resolve by group because it was faster. Sixteen participants described using a combination of manual and group conflict resolution and the reasons why, which we discuss in Section 6.3.

*6.2.2 Reliance.* We wanted to understand whether participants appropriately relied on group resolve and suggested labels. By reliance, we mean user agreement with, or selection of, the presented system suggestion, which is either the labelers' top voted label and/or the AI top predicted label. When manually resolving conflicts, users explicitly rely on the system when they select the suggested label. During group resolve, the reliance on system suggestions is more implicit by accepting the suggested label for each item in the group. Users can examine the system suggested labels before confirming, enabling them to make an informed decision about whether to rely when resolving groups.

We further distinguish between *appropriate* and *inappropriate* reliance by taking into account the accuracy of a label suggestion, as follows:

$$Appropriate\ reliance = \frac{\#\ items\ user\ agreed\ with\ correct\ label\ suggestion}{\#items\ with\ correct\ label\ suggestion} \tag{1}$$

$$Inappropriate\ reliance = \frac{\#\ items\ user\ agreed\ with\ incorrect\ label\ suggestion}{\#\ items\ with\ incorrect\ label\ suggestion} \tag{2}$$

Overall, participants in the LAI conditions had lower inappropriate reliance, but they also had lower appropriate reliance.

*Reliance on Manual and Group Resolve:* To explore appropriate reliance with group resolve, we analyzed data from the 38 participants who resolved conflicts both manually and automatically by group (using the group resolve feature). The data from these participants tells the story of reliance when participants are choosing when to resolve by group and when to resolve manually. The participants we filtered out would have either 100% appropriate reliance and 100% inappropriate reliance, or they resolved all conflicts manually like those in the manual resolution only conditions.

We found a difference across the three conditions in inappropriate reliance ($H(2) = 14.5, p < 0.001$) as well as appropriate reliance ($H(2) = 19.2, p < 0.001$). Participants in the LAI-C condition

Table 5. Appropriate and inappropriate reliance for each condition. The -C conditions include the subset of users who resolved manually and by group. *The LAI conditions had significant differences from the other conditions.

| Condition | Appropriate Reliance | Inappropriate Reliance |
|---|---|---|
| L-C | $M = 94\%, SD = 12\%$ | $M = 84\%, SD = 14\%$ |
| AI-C | $M = 97\%, SD = 4\%$ | $M = 73\%, SD = 20\%$ |
| LAI-C | $M = 86\%, SD = 5\%*$ | $M = 51\%, SD = 13\%*$ |
| L-M | $M = 95\%, SD = 6\%$ | $M = 58\%, SD = 16\%$ |
| AI-M | $M = 94\%, SD = 9\%$ | $M = 50\%, SD = 23\%$ |
| LAI-M | $M = 75\%, SD = 18\%$ | $M = 37\%, SD = 14\%$ |

had significantly lower inappropriate reliance than those in the L-C condition ($p < 0.01$) and the AI-C condition ($p < 0.05$). However, participants in the LAI-C condition also had significantly lower appropriate reliance than those in the L-C condition ($p < 0.01$) and the AI-C condition ($p < 0.001$). See Table 5 for the values for reliance on group for the different conditions.

Due to the differences in the LAI-C condition, we wanted to understand more about how participants appropriately and inappropriately relied on the AI and labelers when they had access to both types of information. We used a Generalized Linear Mixed-Effects Model with participant as the random effect and which information was correct (labeler, AI, or labeler+AI) as the independent variable to compare the impact of these measures within-subjects. We did not find significant differences. However, on average, participants inappropriately relied about twice as much when both the labelers and AI were both wrong ($M = 39\%, SD = 23\%$), compared to when just the AI was wrong ($M = 13\%, SD = 19\%$) or when just the labelers were wrong ($M = 13\%, SD = 15\%$). Similarly, participants in the LAI-C condition appropriately relied most when both the labelers and AI were correct: on average 99% of the time ($SD = 1\%$). They relied appropriately about half as much when either only the labelers were correct ($M = 58\%, SD = 27\%$) or only the AI was correct ($M = 56\%, SD = 30\%$).

*Reliance with only Manual Resolve:* We also wanted to know if participants relied on AI or labeler suggested labels more appropriately when resolving conflicts manually. To study this, we compared reliance in the manual conditions. Across the three manual conditions, we found a significant difference for inappropriate reliance ($H(2) = 12.9, p < 0.01$) and a significant difference for appropriate reliance ($H(2) = 26.3, p < 0.001$). Participants in the LAI-M condition had significantly less inappropriate reliance on the suggested labels than participants in the L-M condition ($p < 0.01$). Participants in the LAI-M condition also appropriately relied on suggested labels significantly less than participants in the L-M ($p < 0.001$) or AI-M conditions ($p < 0.001$). We expected that providing AI and labeler information might reduce inappropriate reliance, by helping to bring users' attention to system errors. However, we did not expect that having both information would also reduce appropriate reliance.

We again wanted to better understand why those with both labeler and AI information (LAI-M) differed in reliance and whether that reliance was based on a particular type of information. For inappropriate reliance, there was an overall significant effect of whether the labelers, AI or both labelers and AI were wrong ($\chi^2(2, 21) = 17.4, p < 0.001$). When both labelers and the AI were wrong, participants were significantly more likely to inappropriately rely ($M = 65\%, SD = 29\%$)

than when just the labelers ($M = 22\%, SD = 15\%, p < 0.01$) or the AI were wrong ($M = 29\%, SD = 20\%, p < 0.01$). We found an overall significant difference across appropriate reliance on the three forms of information (labeler, AI, and both) ($\chi^2(2, 21) = 7.7, p < 0.05$). Post-hoc tests comparing the estimated marginal means with Tukey adjustment for multiple comparisons shows that participants appropriately relied significantly more ($p < 0.05$) when both the labelers and AI were right ($M = 86\%, SD = 19\%$), than when only the labelers were right ($M = 49\%, SD = 26\%$). When only the AI was right, participants relied appropriately 61% of the time ($SD = 23\%$).

### 6.3 RQ3: How do users decide whether to rely on automated conflict resolution?

Users' decisions about when to rely on automated conflict resolution can begin to explain how they ended up appropriately and/or inappropriately relying on the AI. In this section, we focus on the 69 participants who had the choice of manual or group resolve (L-C, AI-C and LAI-C). We interweave our statistical results with the results of the thematic analysis (see Table 9) to illustrate how users made decisions about group and manual resolve. Participants used several methods to decide whether to use manual and/or group resolve: interaction, check high-level performance, check individual, or using their own perceptions.

*6.3.1 Decision Making: Interaction. Confidence through doing* was our highest prevalence decision making code, with 19% of participants' statements. Additionally, 10% of participants chose to manually resolve items in order to better understand the system or the task. These participants talked about wanting to complete some tasks on their own first, before feeling comfortable to resolve tasks using the system. For example, one participant wrote:

> *I felt the first few groups were not quite accurate enough to ignore my own human intelligence while the last group I felt (after completing about 10) that it's accuracy was adequate enough to use.* - P23 (AI-C, 80% accuracy)

Another participant focused more on learning about the system:

> *I started solving tasks manually because I wanted to understand the logic of the task, when I finished the second group I decided to use the other form of resolution, because I wanted to understand what it was like and when I realized that my answers tended to follow the majority, I decided that this method was the most effective.* - P12 (L-C, 85% accuracy)

To explore whether participants used this strategy differently across conditions, we analyzed the percentage of tasks completed manually per group. For the users who both manually and group resolved, we analyzed the percentages of the groups they resolved manually (see Figure 4). Since the groups increase in likelihood of having accurate system labels, a desirable outcome would be for users to manually resolve more conflicts in groups with lower expected system label accuracy and leverage group resolve for groups with higher expected accuracy. We can see that this trend occurs much more in the AI-C and LAI-C conditions than in the L-C conditions. In condition L-C, even in the lower groups, participants still resolved most of the items by group. We performed a Kruskal-Wallis test to compare the average percentage of groups resolved manually by these users and found a high-level significance across the three conditions ($H(2) = 7.7, p < 0.05$). Participants in the L-C condition manually resolved significantly smaller percentages of the groups ($M = 0.27, SD = 0.39$) than those in the AI-C condition ($M = 0.5, SD = 0.45, p < 0.05$) and the LAI-C condition ($M = 0.5, SD = 0.45, p < 0.05$). These results show that one strategy participants used to determine when to rely on group resolve was to manually resolve some items before resolving a group, though participants in the L-C condition may have done this less than others.

*6.3.2 Decision Making: High-Level Performance.* Our thematic analysis showed that some users leveraged high-level performance predictors as a way to decide when to manually or group resolve.

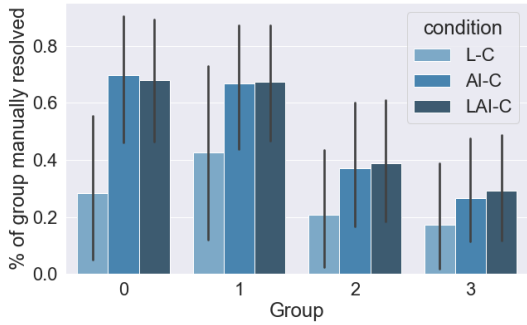


Fig. 4. Percentage of groups resolved manually by participants who used both manual and group resolve.

Eleven percent of responses referenced using percentages to decide whether to use group resolve and six percent of responses mentioned specific thresholds of percentages where participants felt they should manually or group resolve conflicts. Fourteen percent of users referenced the performance of the labelers and/or AI as a reason for their decision, but didn't describe what information they used to assess performance. Participants wrote about selecting groups where they felt comfortable resolving with the system:

> *I resolved whole groups of conflicts when the people were in agreement more than 60% of the time I believe.* - P37 (LAI-C, 77% accuracy)
>
> *I decided to manually do the tasks when the probability of the AI being right was lower. I trusted AI over myself in 3 of the 4 tasks [groups] due to it's high accuracy probability rate.*- P26 (AI-C, 71% accuracy)

The number of users who resolved each group can provide insight into whether users were appropriately using percentages and thresholds to determine when to resolve manually or by group. Figure 5 shows the number of users who resolved each group for the users who used manual and group resolution. Ideally, fewer users would resolve by group in the lower groups and more would resolve by group in the higher groups. From group 0 to group 3, the labelers had increasing majority votes and the AI had increasing predicted probability, indicating it was likely a better decision to use group resolve in the higher groups. We do see an encouraging trend of resolving higher groups more than lower groups. However, across conditions, participants did also use group resolve for even the lowest percentage group.

*6.3.3 Decision Making- Check Individual.* Participants also described checking the results of the AI and/or labelers in order to make decisions about whether to manually or group resolve (see Table 9 for the breakdown of codes). In some cases, checking caused users to feel more confident, for example if they only found a few errors:

> *It looked like the AI got them correct except for maybe one or two that didn't seem to fit so I went with resolve groups of conflicts.* -P18 (AI-C, 71% accuracy)

While others decided that finding errors was a reason not to resolve a group of conflicts:

> *I thought I might resolve the whole group but when I chose to do so it showed me the whole list and I quickly found a couple wrong which bothered me so I continued manually.* -P7 (L-C, 91% accuracy)

In the LAI-C condition where users had both the labeler top voted label and AI top predicted label, checking the correctness was more effort due to needing to check both options:

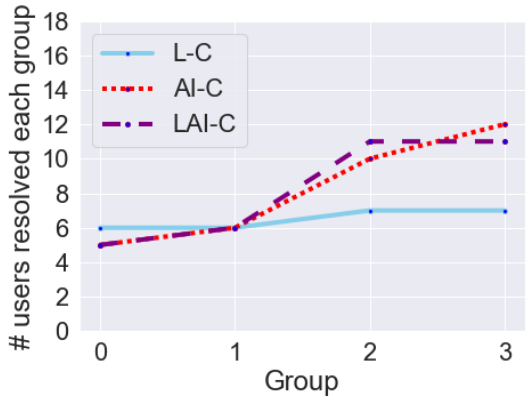


Fig. 5. Number of users who resolved each group per condition.

> *I looked at the AI response, the labeler response and what I thought the correct response would be. In each group, if it looked like either the AI or the labelers were correct in the vast majority of the cases, I chose that option in order to save time.* -P45 (LAI-C, 80% accuracy)

When users chose to resolve a group, they saw a group resolve modal that showed the list of data items and the labels that would be applied with resolve group. Users could choose to cancel at that point. The time users spent on the group resolve modal and their cancel actions can provide more insight into whether they were valuable features and whether people may have been using them to try to calibrate their trust. Canceling after viewing the group resolve modal may indicate that users were engaging with the displayed content in an attempt to make informed decisions.

We found evidence that participants spent more time on the group resolve modal than it would take to merely open it and click submit. We again analyzed the data for participants who resolved with a combination of manual and group resolve. Participants spent on average 13s ($SD = 21.4s$) on the group resolve modal. Across the three conditions, we found a significant difference in time spent on the group resolve modal ($H(2) = 10.7, p < 0.01$). Participants in LAI-C ($M = 15.6s, SD = 17.8s$) spent significantly longer on the group resolve modal than those in L-C ($M = 10.9s, SD = 18.5s$) ($p < 0.01$) and than those in AI-C ($M = 12.3s, SD = 25.9s$) ($p < 0.05$). The LAI-C Group Resolve Modal list showed both the AI and the labeler top choices, which the user could compare, since they had the option to use either AI or labeler group resolve. Participants spending more time on the modal in the LAI-C condition may mean that participants engaged with both sets of labels.

We also found that participants occasionally cancelled group resolve actions after selecting them. Of the 69 participants who had the choice to group resolve, 32 participants (46%) cancelled a resolve group action. The number of participants who cancelled Resolve Group actions was relatively evenly spread across conditions ($p > 0.05$), with 12 participants canceling in L-C, 11 participants canceling in AI-C, and 9 participants cancelling in LAI-C. This likely means that at least some of the participants were checking the results of the group resolve before submitting, a behavior that could be linked to improved reliance.

*6.3.4 Perceptions.* Fifteen participants based their decisions about whether to manually or group resolve on perceptions (see Table 9). Some mentioned that they did trust, while others did not, which aligns with our previous finding that participants' use of group resolve was split across the conditions:

*I decided to trust the AI, based on what I saw.* - P46 (LAI-C, 74% accuracy)

*I do not 100% trusting AI choices* - P50 (LAI-C, 85% accuracy)

Participants rated their trust in the labelers and/or the AI on a Likert scale from 1-5. Participants trended toward trusting both the labelers and AI and there was not a significant difference between participants' trust in the labelers ($M = 3.9, SD = 0.8$) and the AI ($M = 3.7, SD = 0.8, p = 0.13$).

Participants also wrote about using their own judgement (10%) or wanting to check the answers (11%) as reasons for using manual resolution primarily. Participants reported how accurate they believed themselves, the labelers, and/or the AI were on scales from 0-100. On average, participants believed that they were 82% accurate ($SD = 13\%$), while they perceived the AI to be 77% accurate ($SD = 14\%$) and the labelers to be 79% accurate ($SD = 14$). In actuality, participants were 72% accurate ($SD = 18$), while the labelers were 81% accurate and the AI was 71% accurate. These differences in participants' perceptions of their own and the system's accuracy may have impacted their lack of appropriate reliance and/or their inappropriate reliance.

## 7 DISCUSSION AND DESIGN RECOMMENDATIONS

### 7.1 Balancing reliance and time: Provide more system direction

In an ideal human-AI collaboration to resolve labeling conflicts, the user makes near-perfect decisions about when to manually resolve in order to best balance using their intelligence to correct the hardest conflicts, while allowing the system to resolve easier conflicts to save time. We did see impressive gains in accuracy/time for those who could choose to use group resolve. However, study participants exhibited high inappropriate reliance in the L-C and AI-C conditions (84% and 73%, respectively). The use of heuristics to improve reliance on an AI may be less effective than actively engaging with the content [15]. Accordingly, some of our participants' strategies, like looking at the percentages of agreement, may be a heuristic that can lead to inappropriate reliance. Although users with the option to group resolve finished much quicker without a significant reduction in accuracy, ideally, the accuracies would be much better than the human or automation alone [6].

*Recommendation:* Provide more system direction about how to check automation, when to use automation, and when to manually resolve conflicts. Systems could selectively offer automation when the risk is low, or require users to perform checks on a certain amount of data before automating. These recommendations are similar to cognitive forcing functions that have been found to support appropriate reliance, like waiting to introduce automation or waiting until a user asks for automation [15]. To maintain user control, another option would be to provide nudges [50], that aim to improve users' decisions without limiting their options.

### 7.2 Trust: Support Interactive Hedging

We designed our system to enable users to hedge, or trust and verify. Participants could view the list of examples and proposed system labels for group resolve to check before deciding to confirm or cancel the operation [52]. For the LAI-C condition, participants even had two pieces of information with which to make a decision. We found that many participants took advantage of the ability to hedge, with 47% of participants canceling a resolve group action and participants spending on average over 10 seconds using the group resolve modal, with the most amount of time spent in the LAI-C condition. Our survey results further support that participants were checking 'the list' on the group resolve modal and finding errors. This aligns with prior work finding that users wanted to check over individual suggested labels in qualitative coding [58]. We expand this finding, showing that when automating groups of conflict resolutions, users especially want to check the quality of automation.

One risk associated with our tool's hedging feature is that users can find errors and lose trust in the system all-together. Other work has found that errors can erode trust, especially trust formed quickly [47, 75]. This was especially a concern in the LAI-C condition, where participants could see both labeler and AI information, likely leading to the significantly lower appropriate reliance in LAI-C. Further, in cases where either the labeler or the AI was correct but not both, participants had approximately half as much appropriate reliance as when they were both correct. While explanations have been used to mediate trust issues in other AI contexts, research has found that explanations for labeling may not be as important as correct suggestions [58]. Our current system requires users to either accept all system-provided conflict resolutions in a group or to go back and manually complete all, which likely reduced the amount and accuracy of group resolve actions.

*Recommendations:* Give users more control over the groups of conflict resolutions they want to automate. This can manifest in several ways, such as allowing users to filter and explore groups of items based on users' own selections of labeler agreement or AI predicted probability, allowing users to remove items from an automation group, or enabling uses to modify system-selected labels. Visualization, which has been used effectively in other interactive machine learning systems [1, 57], may also be helpful in enabling users to effectively explore and control automation. Allowing users to hedge in these ways could reduce the amount of trust a user needs to have in the system, as they have more control, while also increasing engagement and accuracy [15]. Modifying automated labels would have even further benefits for systems where the machine learning model continuously learns from the user, such as active learning systems [62].

### 7.3 Limitations

We chose to run this study with Mechanical Turk workers, who may not be entirely representative of the population, such as subject matter experts, who would normally complete this task. However, we found that we had a spread of participants who prioritized time, accuracy, and accuracy and time equally, which we believe is likely to approximate workers who have differing time constraints and desires to obtain the highest accuracy possible. In our scenario, participants only automated up to 48 tasks at once. Their behaviors, trust and the systems needed to support may differ on a much larger scale, like automating thousands of tasks at once. Future work should validate our findings in different contexts.

## 8 CONCLUSION

In this work, we explored how users resolved conflicts manually and by group with AI and/or labeler information. We wanted to understand how users made decisions about when to rely on the system manually and by group and whether they could rely appropriately on the system. Participants with the option to group resolve had significantly higher accuracy/time than those without, suggesting that this direction is worth further investigation. We did not find differences in trust or in how many participants were willing to resolve by group across conditions. However, participants who had labeler and AI information relied inappropriately on the system less than those with just labeler or just AI information. Participants used a variety of strategies to decide whether to manually or group resolve, ranging from using heuristics like percentages to deeply comparing individual labeler and AI suggested labels. Finally, our support for hedging shows promise for enabling users to trust and verify the system's labeling, but future work should explore ways to ensure errors don't prevent appropriate reliance that could save users time. Our findings and design recommendations apply not only for human-collaborative conflict resolution systems, but across a variety of related domains that could support multi-user input with AI, like qualitative research, data labeling and annotation, and interactive machine learning.

## REFERENCES


[1] Saleema Amershi, Bongshin Lee, Ashish Kapoor, Ratul Mahajan, and Blaine Christian. 2011. Cuet: human-guided fast and accurate network alarm triage. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 157–166.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[3] Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T Wolf, et al. 2021. AI-Assisted Human Labeling: Batching for Efficiency without Overreliance. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

[4] Zahra Ashktorab, Q. Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. 2020. Human-AI Collaboration in a Cooperative Game Setting: Measuring Social Perception and Outcomes. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 96 (Oct. 2020), 20 pages. https://doi.org/10.1145/3415167

[5] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2020. Optimizing AI for Teamwork. *arXiv preprint arXiv:2004.13102* (2020).

[6] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2021. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11405–11414.

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[8] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[9] Gagan Bansal, Tongshuang Wu, Joyce Zhu, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S Weld. 2020. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *arXiv preprint arXiv:2006.14779* (2020).

[10] Matthew Barnfield. 2020. Think twice before jumping on the bandwagon: Clarifying concepts in research on the bandwagon effect. *Political studies review* 18, 4 (2020), 553–574.

[11] Jürgen Bernard, Marco Hutter, Matthias Zeppelzauer, Dieter Fellner, and Michael Sedlmair. 2017. Comparing visual-interactive labeling with active learning: An experimental study. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 298–308.

[12] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. (2012).

[13] Alexander Braylan and Matthew Lease. 2020. Modeling and Aggregation of Complex Annotations via Annotation Distances. In *Proceedings of The Web Conference 2020*. 1807–1818.

[14] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.

[15] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *arXiv preprint arXiv:2102.09692* (2021).

[16] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[17] Chris Callison-Burch. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1* (Singapore) *(EMNLP '09)*. Association for Computational Linguistics, USA, 286–295.

[18] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).

[19] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2334–2346.

[20] Quanze Chen, Jonathan Bragg, Lydia B Chilton, and Dan S Weld. 2019. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[21] Minsuk Choi, Cheonbok Park, Soyoung Yang, Yonggyu Kim, Jaegul Choo, and Sungsoo Ray Hong. 2019. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[22] John Joon Young Chung, Jean Y Song, Sindhu Kutty, Sungsoo Hong, Juho Kim, and Walter S Lasecki. 2019. Efficient elicitation approaches to estimate collective crowd answers. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[23] Kevin Crowston, Xiaozhong Liu, and Eileen E Allen. 2010. Machine learning and rule-based automated coding of qualitative data. *proceedings of the American Society for Information Science and Technology* 47, 1 (2010), 1–2.
[24] Ryan Drapeau, Lydia Chilton, Jonathan Bragg, and Daniel Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 4.
[25] Margaret Drouhard, Nan-Chen Chen, Jina Suh, Rafal Kocielnik, Vanessa Pena-Araya, Keting Cen, Xiangyi Zheng, and Cecilia R Aragon. 2017. Aeonium: Visual analytics to support collaborative qualitative coding. In *2017 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 220–229.
[26] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing ambiguity in crowdsourcing frame disambiguation. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 6.
[27] Neta Ezer, Sylvain Bruni, Yang Cai, Sam J Hepenstal, Christopher A Miller, and Dylan D Schmorrow. 2019. Trust engineering for Human-AI teams. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 63. SAGE Publications Sage CA: Los Angeles, CA, 322–326.
[28] Arthur Flexer and Thomas Grill. 2016. The problem of limited inter-rater agreement in modelling music similarity. *Journal of new music research* 45, 3 (2016), 239–251.
[29] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376316
[30] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable Active Learning (XAL) Toward AI Explanations as Interfaces for Machine Teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.
[31] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. 2008. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces*. 227–236.
[32] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
[33] Robert R Hoffman, Matthew Johnson, Jeffrey M Bradshaw, and Al Underbrink. 2013. Trust in automation. *IEEE Intelligent Systems* 28, 1 (2013), 84–88.
[34] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 164–168.
[35] Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. 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* (Boulder, Colorado) *(HLT '09)*. Association for Computational Linguistics, USA, 27–35.
[36] Yan Huang and S Shyam Sundar. 2020. Do We Trust the Crowd? Effects of Crowdsourcing on Perceived Credibility of Online Health Information. *Health Communication* (2020), 1–10.
[37] Maia Jacobs, Melanie F Pradier, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. *Translational psychiatry* 11, 1 (2021), 1–9.
[38] Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1637–1648.
[39] René F Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2390–2395.
[40] Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. 5–9.
[41] Silvia Knobloch-Westerwick, Nikhil Sharma, Derek L Hansen, and Scott Alter. 2005. Impact of popularity indications on readers' selective exposure to online news. *Journal of broadcasting & electronic media* 49, 3 (2005), 296–313.
[42] Moritz Körber. 2018. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association*. Springer, 13–30.
[43] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured labeling for facilitating concept evolution in machine learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3075–3084.
[44] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. https://doi.org/10.1145/3290605.3300717

[45] Cheng-Yu Lee and Von-Wun Soo. 2006. The conflict detection and resolution in knowledge merging for image annotation. *Information processing & management* 42, 4 (2006), 1030–1055.
[46] J. D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of Human Factors and Ergonomics Society* 46 (2004), 50 – 80.
[47] Poornima Madhavan, Douglas A Wiegmann, and Frank C Lacson. 2006. Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human factors* 48, 2 (2006), 241–256.
[48] Megh Marathe and Kentaro Toyama. 2018. Semi-automated coding for qualitative research: A user-centered inquiry and initial prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.
[49] John M McGuirl and Nadine B Sarter. 2006. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human factors* 48, 4 (2006), 656–665.
[50] Cristina Mele, Tiziana Russo Spena, Valtteri Kaartemo, and Maria Luisa Marzullo. 2021. Smart nudging: How cognitive technologies enable choice architectures for value co-creation. *Journal of Business Research* 129 (2021), 949–960.
[51] Ana Elisa Méndez Méndez, Mark Cartwright, and Juan Pablo Bello. 2019. Machine-crowd-expert model for increasing user engagement and annotation quality. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.
[52] Debra Meyerson, Karl E Weick, Roderick M Kramer, et al. 1996. Swift trust and temporary groups. *Trust in organizations: Frontiers of theory and research* 166 (1996), 195.
[53] Michael Muller, Christine T Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, et al. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
[54] Stefanie Nowak and Stefan Rüger. 2010. How Reliable Are Annotations via Crowdsourcing: A Study about Inter-Annotator Agreement for Multi-Label Image Annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval* (Philadelphia, Pennsylvania, USA) *(MIR '10)*. Association for Computing Machinery, New York, NY, USA, 557–566. https://doi.org/10.1145/1743384.1743478
[55] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *PloS one* 15, 2 (2020), e0229132.
[56] Vlad L Pop, Alex Shrewsbury, and Francis T Durso. 2015. Individual differences in the calibration of trust in automation. *Human factors* 57, 4 (2015), 545–556.
[57] Donghao Ren, Saleema Amershi, Bongshin Lee, Jina Suh, and Jason D Williams. 2016. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 61–70.
[58] Tim Rietz and Alexander Maedche. 2021. Cody: An AI-Based System to Semi-Automate Coding for Qualitative Research. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[59] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019. Understanding expert disagreement in medical data analysis through structured adjudication. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.
[60] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. Ambiguity-aware AI Assistants for Medical Data Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[61] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–19.
[62] Burr Settles. 2009. Active learning literature survey. (2009).
[63] Shayan Shaikh, Aneela Malik, MS Akram, and Ronika Chakrabarti. 2017. Do luxury brands successfully entice consumers? The role of bandwagon effect. *International Marketing Review* (2017).
[64] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*. 254–263.
[65] Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2020. Would you describe a leopard as yellow? Evaluating crowd-annotations with justified and informative disagreement. In *Proceedings of the 28th International Conference on Computational Linguistics*. 4798–4809.
[66] Yunjia Sun, Edward Lank, and Michael Terry. 2017. Label-and-Learn: Visualizing the Likelihood of Machine Learning Classifier's Success During Data Labeling. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. 523–534.
[67] S Shyam Sundar. 2008. *The MAIN model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative.
[68] S Shyam Sundar, Silvia Knobloch-Westerwick, and Matthias R Hastall. 2007. News cues: Information scent and cognitive heuristics. *Journal of the American society for information science and technology* 58, 3 (2007), 366–378.

[69] S Shyam Sundar, Anne Oeldorf-Hirsch, and Qian Xu. 2008. The bandwagon effect of collaborative filtering technology. In *CHI'08 extended abstracts on Human factors in computing systems*. 3453–3458.

[70] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–6.

[71] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 211 (Nov. 2019), 24 pages. https://doi.org/10.1145/3359313

[72] Jinping Wang, Maria D Molina, and S Shyam Sundar. 2020. When expert recommendation contradicts peer opinion: Relative social influence of valence, group identity and artificial intelligence. *Computers in Human Behavior* 107 (2020), 106278.

[73] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems* 22 (2009), 2035–2043.

[74] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[75] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[76] Kun Yu, Shlomo Berkovsky, Dan Conway, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2016. Trust and reliance based on system accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. 223–227.

[77] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do i trust my machine teammate? an investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 460–468.

[78] Beste F Yuksel, Penny Collisson, and Mary Czerwinski. 2017. Brains or beauty: How to engender trust in user-agent interactions. *ACM Transactions on Internet Technology (TOIT)* 17, 1 (2017), 1–20.

[79] Himanshu Zade, Margaret Drouhard, Bonnie Chinh, Lu Gan, and Cecilia Aragon. 2018. Conceptualizing disagreement in qualitative coding. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2018-April. https://doi.org/10.1145/3173574.3173733

[80] Jing Zhang, Xindong Wu, and Victor S. Sheng. 2016. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review* (2016). https://doi.org/10.1007/s10462-016-9491-9

[81] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

## APPENDIX

Received April 2021; revised November 2021; accepted March 2022

Table 6. Demographic questions in the post-study survey

| Conditions | Question | Response Options |
|---|---|---|
| All | What is your primary spoken language? | English, Arabic, Chinese, French, Hindi, Italian, Spanish, Tamil, Other |
| All | What is your highest level of formal education? | Less than high school, High school, Undergraduate College Degree Graduate Degree (Master's, PhD, JD, etc) |
| All | What is your age? | 18-29, 30-39, 40-49, >50 |
| All | How long have you been doing Mechanical Turk HIT tasks? | Less than 6 months, 6 months - 1 year, 1-2 years, 2-3 years,More than 3 years |
| All | What kind of exposure have you had to the topic of Artificial Intelligence (AI)? | I have never heard of AI., I have heard about AI in the news, friends, or family., I closely follow AI-related news., I have some work experience and/or formal education related to AI., I have significant work experience related to AI |

Table 7. Custom and open-ended questions in the post-study survey

| Conditions | Question | Response Options |
|---|---|---|
| All | What was your priority in completing the tasks? | I prioritized time more than accuracy, I prioritized accuracy more than time., I prioritized time and accuracy approximately equally. |
| All | Approximately how accurate do you believe your responses were across all tasks? | slider 0-100% |
| L and LAI | Approximately how accurate do you believe the labelers were across all tasks? | Slider 0-100% |
| AI and LAI | Approximately how accurate do you believe the AI was across all tasks? | Slider 0-100% |
| C | If you resolved whole groups of conflicts, how did you decide when to do so? If you only manually resolved tasks individually, please describe why you chose not to resolve groups of conflicts. | Open-ended |

Table 8. Trust Questions in the post-study survey. All questions were given to all participants, with variations in whether the questions referred to the AI or the labelers. In the LAI conditions, Participants saw these questions twice, once for the AI and once for Labelers. All questions had response options: Strongly disagree, Disagree, Neutral, Agree, Strongly Agree

| Question |
|---|
| The [AI, Labelers'top choices] is/are capable of resolving conflicts. |
| I am familiar with similar tools. |
| One should be careful with unfamiliar [AI systems/labelers]. |
| The [AI's conflict resolution is/ human labeler choices are] reliable. |
| The [AI's conflict resolution is/ human labeler choices are] unpredictable. |
| I trust the [AI's conflict resolution/human labeler choices]. |
| The [AI's conflict resolution is/human labeler choices are] likely to be wrong. |
| I was able to understand why the [AI resolved conflicts the way it did/ labelers chose labels]. |
| I'm more likely to trust [an AI system/human labelers] than mistrust [it/them]. |
| I can rely on the [AI/ human labelers] to resolve conflicts. |
| The [AI/ human labelers] might make sporadic errors in resolving conflicts. |
| [Automated and intelligent systems/human labelers] generally [work/perform] well. |
| I am confident about the [AI's/human labelers'] capability to resolve conflicts. |
| It's difficult to identify which labels the [AI/human labelers] will select. |
| I have used similar systems |
| The [AI/human labelers] are capable of resolving challenging data labeling conflicts. |

Table 9. Thematic Analysis codes and prevalence for the question: *If you resolved whole groups of conflicts, how did you decide when to do so? If you only manually resolved tasks individually, please describe why you chose not to resolve groups of conflicts.** 17 responses were filtered out for being irrelevant or lacking rationale, leaving 53 responses.

| | | | % Users in Method Strategy | | | |
|---|---|---|---|---|---|---|
| Theme | Code | Description | Manual | Group | Both | Unknown |
| Strategy-Reason | Accuracy | Considered correctness | 28 | 6 | 15 | 6 |
| | Time | Considered length of task | 0 | 13 | 0 | 0 |
| | Effort | Considered easiness | 4 | 2 | 0 | 0 |
| Decision Making-Interaction | Confidence through doing | Manually resolved to gain confidence in group resolve | 0 | 0 | 17 | 2 |
| | Try it | Wanted to understand how the system works as a whole | 6 | 0 | 4 | 0 |
| Decision Making-Check Individual | Found errors | Saw cases where AI and/or labelers were wrong | 6 | 2 | 2 | 2 |
| | Checked List | Checked group resolve modal | 2 | 2 | 0 | 0 |
| | Compared | Checked both AI and labeler to see comparison | 2 | 2 | 4 | 4 |
| Decision Making-High-Level Performance | Verify | Broadly based on how AI and/or labelers performed | 4 | 2 | 2 | 6 |
| | Percentage | Used of labeler vote % and/or AI probability % | 0 | 0 | 9 | 2 |
| | Threshold | Used a specific threshold of time and/or accuracy | 0 | 2 | 4 | 0 |
| Decision Making-Perceptions | Trust | Either specifically mentioned trust or believing that one way was better | 2 | 2 | 2 | 6 |
| | Control | Critical that they checked all tasks | 11 | 0 | 0 | 0 |
| | Own Judgement | Believed strongly in their own judgement | 8 | 0 | 2 | 0 |