Comparing Zealous and Restrained AI Recommendations in a Real-World Human-AI Collaboration Task

Chengyuan Xu
University of California, Santa Barbara
Santa Barbara, USA
cxu@ucsb.edu

Kuo-Chin Lien
Appen
Sunnyvale, USA
klien@appen.com

Tobias Höllerer
University of California, Santa Barbara
Santa Barbara, USA
holl@cs.ucsb.edu

Figure 1: In video anonymization, face annotation and blurring is a high-stakes task that requires humans to check every frame. It demands high recall because one missed face can reveal a person's identity in the entire video. We can improve recall and reduce task completion time by forming a human-AI team. We may have two AIs with the same (F1) performance as shown in (c) but provide different sets of recommendations (a & b). A "zealous" AI would prioritize recall by suggesting more detections, even low-confidence ones. A "restrained" AI would only provide high-precision recommendations. Which AI teammate can help the human annotators finish in less time and with higher recall?

ABSTRACT

When designing an AI-assisted decision-making system, there is often a tradeoff between precision and recall in the AI's recommendations. We argue that careful exploitation of this tradeoff can harness the complementary strengths in the human-AI collaboration to significantly improve team performance. We investigate a real-world video anonymization task for which recall is paramount and more costly to improve. We analyze the performance of 78 professional annotators working with a) no AI assistance, b) a high-precision "restrained" AI, and c) a high-recall "zealous" AI in over 3,466 person-hours of annotation work. In comparison, the zealous AI helps human teammates achieve significantly shorter task completion time and higher recall. In a follow-up study, we remove AI assistance for everyone and find negative training effects on annotators trained with the restrained AI. These findings and our analysis point to important implications for the design of AI assistance in recall-demanding scenarios.

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581282

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; **Collaborative interaction**; • **Computing methodologies** → *Computer vision*; *Machine learning*.

KEYWORDS

human-AI team, AI-assisted decision making, precision and recall, real-world application, empirical study, computer vision, face detection, video annotation

ACM Reference Format:
Chengyuan Xu, Kuo-Chin Lien, and Tobias Höllerer. 2023. Comparing Zealous and Restrained AI Recommendations in a Real-World Human-AI Collaboration Task. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3544548.3581282

1 INTRODUCTION

Machine-learning-based artificial intelligence (AI) systems have exceeded human performance in certain applications. But in high-stakes domains where fully-autonomous AI is not at peak performance or not permitted, such as in clinical decision-making [7, 9, 49, 53, 56] or driver assistance [11, 13, 22], forming a human-AI team is a viable strategy to improve both efficiency and accuracy. AI can provide recommendations while human users maintain agency and control over the final decisions. Studies have shown the human-AI team is expected to achieve "complementary team performance"

– the team performance being better than either one alone [2, 5, 61]. But there are more questions than answers on which exact factors in the AI system affect the team performance and how.

Bansal et al. recently showed in simplified binary classification problems that **the most accurate AI is not necessarily the best teammate unless it helps to improve the team utility** [2]. But how about in more complex problems where the AI teammate is not simply better or worse for its accuracy? For example, in many computer vision problems, people determine the best-performing algorithms based on combination metrics such as the F1 score [14, 47], which can be broken down into two metrics – precision and recall [16, 33, 34]. Researchers can either balance the two metrics or prioritize one over the other to identify the best model for their application [20, 40]. Two AI systems can have the same F1 score but provide very different recommendations with different measures of recall (see a, b in Figure 1). The tradeoff between precision and recall puts them on different parts of the same F1 isoline (see Figure 1 c). Without additional context, one might argue that there is no better or worse between these two AIs.

In order to capitalize on complementary strengths of humans and AI when presented with tradeoffs in AI precision and recall, we need to be able to answer two questions: **1) for a given task, can we clearly identify if either precision or recall is more important than the other**, and **2) independent of importance, is it vastly easier or harder for humans to improve either precision or recall.**

Consider for example a pedestrian detection task in a driver assistance system: prioritizing the detection model towards either precision or recall will hurt the other. Human instinct tells us the risk of a missing detection could be lethal, so we should tune the AI system to prioritize recall, i.e., towards a "zealous" AI that provides more detections (recommendations), even the low-confidence ones, at the risk of more false positive errors. In this context, the opposite "restrained" AI would only provide high-confidence detections and prioritize precision, but at the risk of more false negative errors.

In this work, we investigate how a high-recall zealous AI and a high-precision restrained AI can affect human-AI team performance in a real-world scenario. Compared to, say testing pedestrian detectors on the road, video anonymization is a similar but easier-to-test recall-demanding task. We set up a face annotation task for personally identifiable information (PII) protection that blurs human faces in a real-world video dataset [23]. PII protection is a critical task with increasing demand for both ethical research and abiding by regulatory requirements[1]. Similar to pedestrian detection, where the cost of a missing detection is very high, one unlabeled face in a single frame can reveal a person's identity in the entire video, if not the entire dataset.

This paper focuses on the common yet critical human-AI collaboration setting, in which recall is more important than precision. As for our second question, "is it vastly easier or harder for humans to improve either precision or recall?", an in-depth analysis of the video annotation workflow shows that improving recall is more costly than precision in this task since it is much harder for human annotators to draw a bounding box accurately than rejecting an incorrect one (see Section 3.2 & 3.3 for a full discussion).

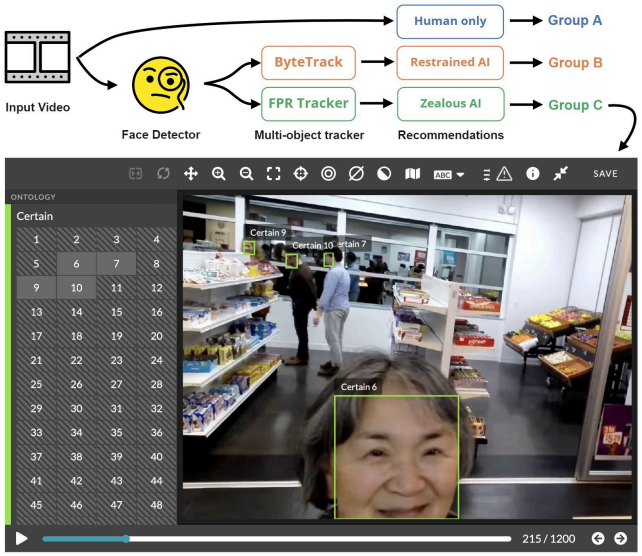


Figure 2: Data processing workflow for Part 1 of the study and the annotation tool user interface. The two AI teammates share the same face detector, which generates bounding box face detections for each frame independently. The ByteTrack tracker [64] and our proposed false-positive-robust (FPR) tracker define the restrained or zealous AI recommendations – they track the per-frame detections temporally to pre-annotate the videos as shown above. For the human-only workflow, annotators must manually draw a box and adjust its size and location across many frames.

The answers to our two questions for our task reveal **an optimization opportunity: the AI recommendation tradeoff between precision and recall can be used to exploit complementary strengths of the human and the AI in such collaborative tasks**. We posit that similar optimization opportunities exist for many other human-AI collaboration tasks. In addition, locating faces is a human instinct[2] that requires no specific training or domain expertise to get started, making face detection a good candidate task to study the effects of different AI recommendations. The relatively small inter-personal differences also make the task a good representative of recall-demanding human-AI collaboration tasks.

Our large-scale empirical study had 78 professional data annotators spend over 3,466 person-hours[3] to submit a total of 2780 annotated 30-second videos. The between-subjects study split the annotators into three 26-people treatment groups. Detailed worker profiles ensured similar average experience between the groups (details in Section 4.2). Each participant annotated human faces in 36 real-world videos of a variety of activities (see examples in

[1] E.g., The General Data Protection Regulation (EU) or The California Consumer Privacy Act of 2018 (CCPA)

[2] Here we refer to the ability to find human faces in a given image. We do not refer to recognizing people by face, which can be affected by Prosopagnosia (face blindness).

[3] Our system logged 3,466 person-hours of annotation work, which does not include pilot studies, training sessions, and answering multiple questionnaires. On average it took the 78 annotators three to four weeks to finish the entire study.

Figure 4). We measure each group's annotation quality and task completion time. Any improvement in time is very meaningful for annotation tasks not only because of the cost. Fatigue induced by long working hours may also cause a decline in quality.

In Part 1 of the **two-part study**, the three groups of annotators processed the same 24 videos, each with a) no AI assistance, b) pre-annotated bounding boxes recommended by the restrained AI, or c) the zealous AI. Figure 2 summarizes the treatment groups and shows the annotation tool's interface. In Part 2, the three groups annotated another 12 videos but all without the AI's help. This design **allows us to learn how prior human-AI collaboration experience can affect user skills**, should they lose access to AI recommendations in the future. The two-part experiment aims to answer the following research questions:

- Q1 Can the human-AI teams achieve "complementary team performance" in this task?
- Q2 Which AI helps annotators be more efficient, i.e. save time?
- Q3 Which AI helps annotators achieve higher recall?
- Q4 Will collaborating with an AI improve or hurt user skills?

We will answer each of the research questions in Section 5. Here we summarize this work's contributions:

- We propose the concept of restrained and zealous AI recommendations to compare the tradeoff between precision and recall in tuning AI-assisted decision-making systems and investigate how they affect human-AI team performance in high-stakes recall-demanding tasks.
- We design a large empirical study to compare the restrained and zealous AI on a face annotation task for video anonymization with 78 professional data annotators. The two-part experiment yielded significant findings to inform future AI assistance design for recall-demanding tasks.
- The analysis of 3,466 person-hours of annotation work reveals significant findings:
  - Our study serves as a real-world case study of complementary team performance (cf. [25, 31, 42, 44]).
  - Identifying the complementary strengths of both human and AI teammates for a task is key to better team performance. The recall-demanding task and the higher cost of improving recall motivated us to propose the zealous AI, which provides high-recall recommendations and leads to significantly better task completion time and recall.
  - The follow-up study demonstrates that naively pairing humans with an AI system designed for autonomous settings without optimizing it for the task at hand or for the human-AI workflow could potentially have a negative training effect on the users.

2 RELATED WORK

Factors affecting human-AI team performance. While human-AI teams have been studied extensively from various perspectives like in crowdsourcing settings [25, 37], computer vision tasks [25, 44, 53], high-stakes tasks [3, 4, 44, 63], and real-world tasks [1, 25, 31, 42, 44, 49, 55], we still have more questions than answers on exactly which factors affect team performance and how. Researchers have looked into factors like users' mental models [3, 10], user expectations [28, 58, 62], cognitive biases [45], model updates during collaboration [4], model accuracy [2, 58], model interpretability or explanations [5, 6, 21, 26, 36, 46, 54], as well as the tradeoff between accuracy and interpretability [9]. Studying user's trust and appropriate or inappropriate reliance on AI [7, 30, 35, 41, 59, 63] is another important direction.

This paper is aligned with works that focused on the tradeoff between precision and recall in AI recommendations and its effect on team performance. Kay et al. [28] introduced the acceptability of accuracy as a new measure and survey instrument to connect classifier evaluation to users' subjective perception of accuracy. Kocielnik et al. [29] compared two 50%-accurate AI-powered scheduling assistants – one avoids false positive errors, and one avoids false negative. This is a similar design as for our restrained and zealous AIs – their study found that false positive errors are more acceptable by participants, which corroborates the overall better performance we observed in the zealous AI group, who also dealt with more false positive errors.

Balancing precision and recall to compare two real-world AI systems in a human-AI collaboration task is not easy, previous works derived insight from hypothetical systems or manually balanced recommendations [28, 29]. In this work, we provide a real-world user study by observing how 78 professional users would interact with two high-performance face tracking AI systems that are tuned to truthfully portray the realistic tradeoff between high-precision and high-recall on a recent egocentric video dataset.

Face detection. The annotation platform we used has a built-in face detector, RetinaFace [18], integrated for autonomous workflows. Our literature search found RetinaFace remains a top-ranking method on the WIDER FACE benchmark [57]. Because more recent methods do not provide significant performance improvement, we continue to use RetinaFace as a consistent baseline to compare with our algorithmic improvements in tracking.

Multi-object tracking. In the AI-assisted face annotation task, the AI teammate provides annotation recommendations for users to review. Conventionally a face detector provides per-frame face bounding boxes and a multi-object tracking (MOT) algorithm produces continuous tracks of the same object across frames. This is known as tracking-by-detection. Recent MOT methods like TransTrack [48], DETR [8], Deformable DETR [65], TrackFormer [38], and TransMOT [12] etc. all move toward the end-to-end Transformer-based [50] architecture. However, these black-box MOTs share the same drawback as they are designed for fully-autonomous settings. Similar to Caruana et al.'s observation that modular system provides better transparency [9], the two-part tracking-by-detection frameworks actually provide us the interpretability and flexibility to steer the output recommendations as needed, so we can produce restrained and zealous AI recommendations for comparison. We reviewed state-of-the-art methods in related multi-object tracking benchmarks [15, 39, 51] in search of a multi-object tracker suitable for a human-in-the-loop annotation workflow. ByteTrack [64] is a conventional tracker that outperforms numerous Transformer-based trackers mentioned earlier.

Video annotation. While there are various public video annotation platforms or tools to choose from [19, 27, 52, 60], we use a proprietary video annotation tool to gain access to professional

data annotators who are already familiar with the specific tool from their past project experience. This tool has Linear Interpolation [52] activated by default, which provides semi-automatic assistance by linearly interpolating a box between two manually annotated key frames. In this study, all participants, including annotators who review AI's annotation recommendations have access to this functionality. Linear Interpolation is also an ideal baseline as all participants have sufficient experience using it. We will refer to this basic setup as human only, the baseline method, or the manual method in the rest of the paper.

3 ALGORITHM CHOICES AND PILOT STUDIES

3.1 Precision and recall in multi-object tracking

Precision, recall, and F1 are important performance metrics that can describe the characteristics of a model and are central concepts in this work and other human-AI research [28, 29]. Specifically, in the context of annotating and tracking faces with bounding boxes in videos:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{Face boxes correctly drawn}}{\text{All boxes drawn by the user (or the AI)}} \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{Face boxes correctly drawn}}{\text{All ground truth face boxes}} \quad (2)$$

$$\text{F1} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

where the TPs are true positives, face boxes that were correctly drawn. The FPs are false positives, boxes drawn by the AI or user which did not match real faces properly. The FNs are false negatives, where there is a real face, but the box is missing.

The F1 score is the harmonic mean of the precision and recall (Equation 3). We visually introduced the concept of this function using three methods that have the same F1 score in Figure 1 (c). Put simply, a video pre-annotated by a high-recall method (zealous AI) would have more false-positive boxes – the user will make more rejections but add fewer missing boxes. A video pre-annotated by the high-precision method (restrained AI) would provide mostly correct boxes but the user will need to add more missing boxes.

We are interested in how users will perform differently given restrained or zealous AI recommendations in an AI-assisted face annotation task. While it is easy to generate high-precision annotations by simply avoiding low-confidence detections, it is hard for trackers to produce high-recall results while maintaining a similarly high F1 score at the same time. This motivates us to propose a tracking algorithm that pushes recall to the limit, but aims to maintain a similar level of F1 score. We take advantage of the fact that **our tracking results will be reviewed by human annotators, allowing us to make targeted optimizations**. We test our ideas of a user-friendly tracker with professional annotators through pilot studies. Observing how users work with trackers allows us to further improve the algorithm.

3.2 Pilot studies

We conducted two pilot studies to observe how professional data annotators work with AI recommendations. Annotators were tasked to draw bounding boxes around potentially moving or blurred faces of any size in a 1,200-frame video sequence of a busy shopping scene in both sessions (similar to hard videos in the formal study). We provided training material on how to review recommendations from the AI for the face annotation job. The annotation tool user interface is shown in Figure 2. With their consent, we recorded their screens to keep track of mouse movements and other user habits. Each session included ten different users with above-average experience. Both pilot studies concluded with a survey about experiment design and their experience. The two pilot sessions were spaced two weeks apart to test algorithm and design improvements.

Users' screen recordings helped us observe the following user habits and behaviors that are not possible to be identified solely from the results:

- *Certain bad recommendations cost most of the human review efforts.* Following the Pareto Principle [24], annotators in fact spent most of their time and effort amending a small fraction of AI recommendations. The tiny bounding boxes (see examples of three tiny faces in Figure 2), duplicate detections (often clustered), and temporally sparse detections (short tracks) are the most costly recommendations. Addressing these issues allows annotators to have better continuity in their workflow.
- *Model explanation should not increase task complexity.* Initially, we offered model explanations using "Certain" and "Uncertain" labels based on the face detector's confidence, hoping this can assist users' decision-making. But video recordings and user feedback revealed that the extra information in fact increased the task complexity and caused unnecessary confusion. This design was eventually not considered in the formal experiment.

Observing how human annotators review AI recommendations (bounding box pre-annotations) in multi-object detection and tracking tasks inspired us to **break the complex workflow into three fundamental user actions: *accept***, ***reject***, or ***solve***, each coming with a higher cost in time. Figure 3 explains each action's time complexity. We can connect these three actions with our two main objectives (time and recall) to make **a simple deduction to identify the human-AI complementary strengths** in this task:

1 ***reject*** improves precision and ***solve*** improves recall. A correct ***accept*** improves both.
2 It takes the AI constant time to ***solve*** additional cases (give more recommendations) with a downside of more false-positive boxes for humans to reject.
3 Humans are faster at ***reject***ing a false-positive (incorrect) box than to ***solve*** a false-negative (missing) box.
4 We also know recall is more important than precision in video anonymization tasks.
5 Thus, **a clear path to better human-AI team performance is to delegate more *solve* actions to the AI, so the human's overall effort is reduced by doing more easy *rejecting* and only *solving* the most challenging faces**.

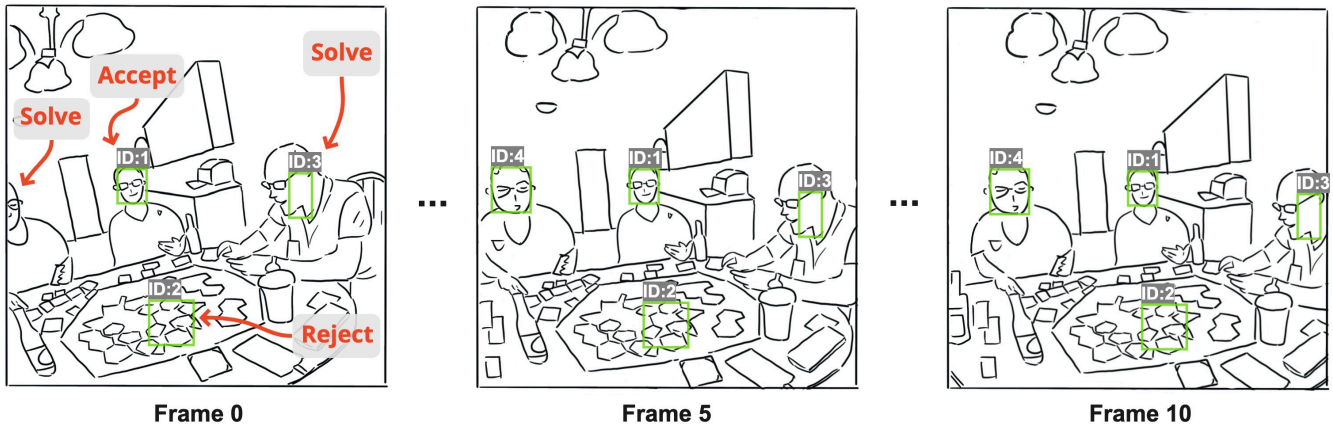


Figure 3: When reviewing the AI teammate's recommendations (green bounding boxes), a user takes one of the three actions for each box: *accept*, *reject*, or *solve*. In video annotation, because the boxes are temporally tracked across many frames, each action's time complexity is drastically different, note the two types of *Solve* in frame 0 can come at different cost, too.

ID:1 – A user can *accept* the true-positive track ID:1 boxes without any action.

ID:2 – The entire false-positive ID:2 track can be rejected with two mouse clicks by deleting the ID in any of the frames, which is $O(1)$ in time complexity.

ID:3 – False-positive recommendations, like track ID:3, are the most time-consuming to *solve*: the user can delete and redo this face, or manually adjust every frame until the AI's pre-annotation becomes acceptable with $\geq n$ mouse clicks, $O(n)$.

ID:4 – In frame 0, to *solve* the false-negative missing box for the left-most person, a user needs to manually draw a box and adjust its location and size until the AI-suggested box ID:4 comes in with $\leq n$ mouse clicks, $O(n)$ where n is the number of frames.

3.3 The false-positive-robust (FPR) tracker

We adopted a tracking-by-detection system to produce face pre-annotations (Section 2), the two-part system design allows us to feed the same per-frame face detection from RetinaFace [18] to different downstream multi-object trackers like the ByteTrack [64] or our own designs for a fair comparison. Learning from our pilot studies observations, we propose the false-positive-robust (FPR) tracker that specifically provides user-friendly annotation recommendations. We use **the following unconventional strategies** to design the FPR tracker that can take overwhelmingly noisy detections with a high false positive rate as input but outputs "clean" tracks for a human-in-the-loop workflow:

- To improve the AI's recall, we apply an **extremely low threshold ($t \geq 0.01,\ t \in [0, 1]$) on the face detector's confidence score** to keep any potentially useful detected boxes. This is not a viable solution for Autonomous AI systems but we are working in conjunction with a human.
- The consequence of such a low face detector threshold is **clusters of overlapping boxes** on small faces. Our solution: for each cluster, we perform non-maximum suppression [43] by only keeping the single bounding box with the highest confidence score because in most cases they are duplicate detections on one true face. This step also improves the AI recommendations' precision.
- Finally, based on our observation that **the majority of temporally sparse detections are false positives** induced by the low threshold, we remove any tracks that are shorter than m consecutive frames so they do not interrupt users' continuity. We used $m = 10$ in the FPR tracker. Although some true-positive faces are also removed, users are much faster at solving an unlabeled face from scratch than filling the gaps between temporally sparse detections.

To design the experiment, we also need a restrained AI that generates recommendations of similar performance (F1 score) but with high precision. This is done by using only the high-confidence ($t \geq 0.8, t \in [0, 1]$) face detections with ByteTrack. To ensure fair comparison and reduce moving parts in our systems, we use the same face detection model RetinaFace [18] for both AI teammates. It is the two different (fully transparent) trackers we apply that push the AI recommendations towards either high-precision or high-recall (Figure 2).

Note that we were only able to optimize the FPR tracker and ByteTrack through pilot studies because the ground truth data was not available for the 36 videos used in the user study. After the study, we aggregated the annotations from all 78 participants (2,780 submissions in total) to form an expert-reviewed consensus to serve as the ground truth. It turns out the zealous AI recommendations (FPR tracker) yielded an F1 score of 90.9% and the restrained AI (ByteTrack) had an F1 score of 93.4%. While the two AIs did not provide identical initial performance for their human teammates, we achieved the goal of two distinctive high-recall and high-precision AIs (Figure 5). The performance gap also provided us additional evidence to support our previous deduction on the zealous AI being the superior choice for this task, which we will discuss in Section 5.1.

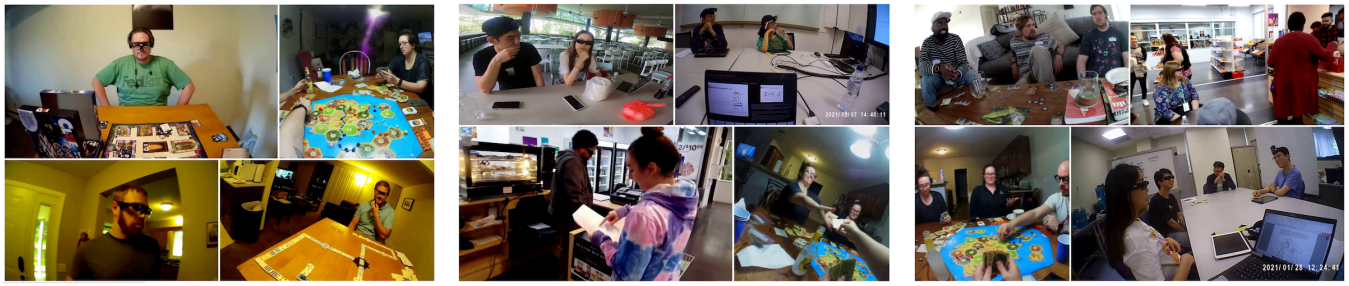


Figure 4: Screenshot examples of Ego4D videos [23] used in our face annotation experiment. Easy videos include about one face to annotate in each non-empty frame. Medium videos include about two faces. Hard videos include three or more faces. Videos with more faces are expected to take longer time to finish. The study results show shorter to longer completion times for Easy, Medium, and Hard videos in both parts (see Figure 6 and Figure 7), demonstrating that our video difficulty categorization is reasonable and performed as expected. We also considered scene diversity, box size (smaller faces are harder), and camera movement intensity (more movement is harder) to ensure a balanced difficulty distribution in selecting the specific videos.

4 EXPERIMENTS

In this work, we aim to investigate how restrained and zealous AI recommendations will affect human-AI team performance. We are also curious if the collaboration experience with an AI teammate can affect users' skills, should they lose access to AI assistance in the future. We design a two-part empirical study to test the restrained and zealous AIs in a recall-demanding high-stakes task.

4.1 The task and data.

Face annotation for video anonymization is a perfect example of recall-demanding tasks – a missing face in a single frame can reveal a person's identity in the entire video. The high-stakes nature requires humans to annotate or verify every frame, yet the manual process will become the throughput bottleneck. The tedious process and long hours may also fatigue annotators and cause a decline in quality. In addition, **because the task of locating faces requires no specific training or domain expertise, it should help the generalizability of our observations** to other AI-assisted annotation tasks or even to other recall-demanding human-AI collaboration tasks.

In our human-AI collaboration setting, the AI teammate provides recommendations in the form of bounding boxes (see examples in Figure 2), and a user reviews each of the AI's pre-annotations to make one of the three decisions shown in Figure 3. We evaluate users' performance on the two most important metrics for face anonymization: **task completion time** and **recall**.

To test different AI recommendations in a real-world setting, we curate 36 first-person videos from a large-scale egocentric video dataset Ego4D [23]. Privacy has always been a major concern for datasets collecting human activities so first-person videos are ideal for this study. The videos we selected include various indoor social activities that are suitable for benchmarking face detection and annotation tasks. Each video clip is 30 seconds long, or 900 frames. We estimate each video takes about 30 minutes to one hour to fully annotate, depending on its difficulty.

The different annotation methods (without or with different AI recommendations) adopted by the three treatment groups are the first level of independent variables that we will discuss in the next section. The second level of independent variables that can affect users' performance is the difficulty of the videos. We divide the videos into Easy, Medium, and Hard categories based on the average number of people one needs to track simultaneously in non-empty frames (see examples in Figure 4). We also considered factors like scene diversity, bounding box size, and camera movement intensity that affect the annotation difficulty in a more subtle way. Based on this overall difficulty ranking distribution, we ensure Part 1 and Part 2 videos are not only similar in content but also consistent in annotation difficulty.

We generate the bounding box ground truth by aggregating the crowd's annotations to reach a consensus, which is further reviewed and refined by a domain expert. We used an equal number of manual and AI-assisted submissions for each video to generate an unbiased ground truth.

On task completion time, **annotators are advised to finish each video without taking breaks longer than five minutes** but we still need to reject outlier video completion times caused by a known limitation of the annotation tool – the timer continues if an ongoing task window was left idle, or the timer will reset if the annotator continues from previously saved progress. We adopted median absolute deviation (MAD) [32] by comparing each video's completion time within each group to reject 420 out of 2780 (15.11%) completed videos, including completion times that are less than six minutes (the minimum time needed to verify each frame) or longer than median + 3 * MAD. The rejected videos also include all 36 submissions from one particular problematic user, see Section 6.2.

4.2 Participants and three treatment groups.

A total of 78 in-house professional data annotators completed our study. It is important to note that in this project **they are paid at their regular hourly rate, so participants are not motivated by compensation to work faster**.

In the between-subjects experiment, participants were evenly split into three 26-people treatment groups to annotate identical sets

Group	Novice	Veteran	Part 1 method	Submissions	Part 2 method	Submissions
A	11	14	Human only	602	Human only	299
B	14	12	Restrained AI + Human	619	Human only	304
C	13	13	Zealous AI + Human	621	Human only	299

Table 1: In the two-part study, the three treatment groups use different methods in Part 1, but we remove all AI assistance in Part 2. The novice and veteran workers represent a balance of different user expertise in each group. The submission numbers are the 30-second annotated videos each group finished. Note that Group A is one user short as a particular worker was later rejected because of repeated bad submissions.

of videos. The annotators' profiles ensure similar average experience between the groups. The assignments also considered people's day/night shifts and computer setup to ensure a fair comparison.

The participants have at least two months or up to five years of data annotation experience, with an average experience of 20.9 months. We use the median experience of 17 months to split the user expertise factor so each group has about half novice and half veteran workers (see Table 1). All annotators were aware of participating in a study testing new AI-assisted annotation algorithms and were free to leave the study at any time. The Human Subjects Committee (HSC) approved our procedure and each participant was provided a consent form during the survey session.

Group A servers as the baseline, they use an efficient annotation tool that supports linear interpolation [52] but solely relies on manual annotation in both parts of the study. Groups B and C work with their AI teammates in Part 1 of the study. They use the same tool as Group A but the AI will have pre-annotated the videos (see example in Figure 2). Group B reviews the restrained AI recommendations that prioritize precision. Group C reviews the zealous AI recommendations that prioritize recall (see a, b in Figure 1). The treatment groups are summarized in Table 1 or Figure 2. We informed the participants in Groups B and C that they are working with an AI that provides recommendations to assist their annotation work, but they do not know the difference between the two human-AI groups.

4.3 Experiment procedure of the two-part study.

Before beginning the study, we organized a video conference training session with each treatment group to calibrate the task background and requirements. All participants were also asked to review the instruction text and a training video on the landing page. Previous pilot study users become supervisors in each group to ensure all participants have finished the training and the surveys before processing to the next step. We also created three instant messaging (IM) groups to answer questions and send out reminders when necessary. The overall procedure can be summarized as follows:

$$\text{Training} \rightarrow \text{Survey 0} \rightarrow$$
$$\text{Part 1 (24 videos, different methods)} \rightarrow \text{Survey 1} \rightarrow$$
$$\text{Part 2 (12 videos, same method)} \rightarrow \text{Survey 2}$$

In **Part 1**, all participants from Groups A, B, and C each annotated 24 videos using different methods. For each annotator, the videos were assigned in random order by the annotation platform. We also reminded all participants to avoid taking breaks longer than five minutes before finishing a video, so the timing is more accurate. Depending on the method and individual pace, it took all groups on the order of two to three weeks to finish Part 1. In **Part 2**, all participants annotated another 12 videos from similar scenes. But we took away the AI assistance from the two human-AI teams B and C in order to find out if their previous human-AI collaboration experiences trained them in any way so that they would perform differently on manual annotations from here on out.

A post-task survey was administered after each part of the study. **Survey 0** was set to "repeat until perfect", this was to verify that the participants were clear about the task requirements before they could start the actual annotation. **Survey 1** focused on getting people's immediate feedback on their experience working with the AI they were paired with. Questions include the correctness and consistency of the AI recommendations, and if the AI made their job easier. This allows us to compare if participants' subjective feelings match the different AI recommendations' underlying personae (high-precision vs. high-recall). **Survey 2** focused on comparing the annotators' preference between AI-assisted and human-only methods after they had experienced both workflows on the same task.

5 RESULTS

In this section, we present our study results and analysis by answering each research question presented in Section 1. For statistical analysis, we ran one-way ANOVA or one-way Welch ANOVA tests, depending on the underlying assumptions being satisfied, followed by Pairwise Tukey-HSD or Games-Howell post-hoc tests, respectively. To examine interactions between factors, we conducted two-way ANOVAs followed by Pairwise Tukey-HSD or Bonferroni-corrected post-hoc tests. We adopted Type III sums of squares in ANOVA to address unbalanced data.

Research questions **Q1**, **Q2**, and **Q3** focus on results from Part 1 of the study (Figures 5, 6, 8, and 10a), in which Groups B and C collaborated with restrained and zealous AIs. Question **Q4** focuses on results from Part 2 (Figures 7, 9, and 10b) to examine how the prior human-AI collaboration experience could affect the users.

5.1 Q1: Can the human-AI teams achieve "complementary team performance" in this task?

Bansal et al.[5] defines *complementary team performance* as the human-AI team performance exceeding both the human-only and AI-only performance.

Figure 5 shows the two human-AI teams B & C reached comparable F1 scores of 96.9% & 96.8%, respectively, significantly better than the human-only Group A that reached 94.5% (Welch $F_{2,1151} = 18.2$,

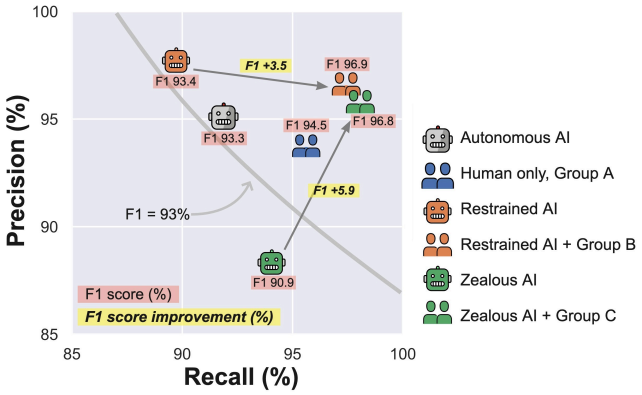


Figure 5: Visualizing each group's overall annotation quality on the precision-recall plot with F1 scores (Part 1). Group A manually annotates all videos and without surprise, they are the slowest (Figure 6) with a quality better than Autonomous AI alone but worse than the two human-AI groups' team effort. Annotators in Groups B & C had to *accept*, *reject*, or *solve* the face boxes pre-annotated by the restrained or zealous AIs to improve the human-AI team performance. The arrows show how much humans improved from the AIs' initial annotation.

$p < 0.0001$). Both human-AI teams improved F1 accuracy and recall significantly compared to their human-only counterpart.

Because the high-stakes nature of this task rules out autonomous AI as a viable option, we really only need to compare the human-AI team performance with human-only performance in Part 1 of our study. However, to verify complementary team performance, we also verify that the two human-AI teams achieved higher performance in terms of F1 scores and recall than their respective AI's initial standalone performance.

Comparing each human-AI team with their perspective AI teammates' initial performance – Group B annotators improved the restrained AI from 93.4% to 96.9% (Welch $F_{1,1228} = 178$, $p < 0.0001$), Group C annotators improved the zealous AI from 90.9% to 96.8% (Welch $F_{1,837} = 169$, $p < 0.0001$). Both human-AI teams improved significantly from their respective AI teammate's solo performance.

It is understandable that Bansal et al. only considered accuracy and did not compare task completion time in complementary performance, since the human-AI teamwork will undoubtedly add more time than AI alone. As we discussed, task completion times directly affect the operation cost as people are paid at an hourly rate, making it a critical metric for annotation tasks, so we additionally compare the human-AI teams' task completion times with the human-only team.

We saw overall significant differences between all three groups on task completion time (Welch $F_{2,1039} = 48.6$, $p < 0.0001$), as shown in Figure 6, left. As a baseline, on average it took 1.05 hours for Group A to manually annotate a 30-second video of 900 frames. Group B took a significantly shorter time of 0.91 hours (Games-Howell $p < 0.001$) to review the restrained AI recommendations. Group C only used 0.73 hours to review zealous AI's recommendations, also significantly shorter than the human-only Group A (Games-Howell $p < 0.0001$).

It is also worth noting that Group C, the zealous human-AI team, had an overall significantly worse starting point than Group B in terms of F1 score: 90.9% vs. 93.4% (Welch $F_{1,854} = 35.32$, $p < 0.0001$) as shown in Figure 5. However, annotators working with the zealous AI managed to achieve a significantly higher improvement in F1 score of +5.9% vs. +3.5% (Welch $F_{1,934} = 45.02$, $p < 0.0001$) in significantly less time! This disadvantage for Group C provided the opportunity to demonstrate that our deduction in Section 3.3 was correct – a human-AI team can do better in both time and quality (in terms of F1 improvement) by asking the human to ***reject*** more false positives and only ***solve*** the most challenging faces, i.e., the high-recall zealous AI.

In summary, we have not only verified complementary team performance on accuracy, but also showed human-AI teams could achieve significantly shorter task completions time in a real-world case study.

5.2 Q2: Which AI helps annotators be more efficient, i.e. save time?

We mentioned that the professional **annotators are paid at their fixed hourly rate in this task**, which means 1) they are not necessarily motivated to work faster, and 2) from the business perspective, their task completion time directly impacts operation costs. We discussed in Section 5.1 that overall, both human-AI teams have significantly shortened task completion time compared to the baseline Group A (Figure 6 left). Specifically, the zealous AI recommendations help annotators use 20% less time than the restrained AI recommendations with statistical significance (0.73 hours vs. 0.91 hours, Games-Howell $p < 0.0001$).

Video difficulty. Figure 6 (middle) plots task time by video difficulty and saw a significant interaction between group and video difficulty on task completion time (ANOVA $F_{4,1577} = 5.37$, $p < 0.0001$, $\eta_p^2 = 0.016$, small). Specifically, Group C which reviewed zealous AI recommendations used significantly less time than both Group A and B in medium videos (Bonferroni $p < 0.0001$ & $p < 0.0001$), as well as in hard videos (Bonferroni $p < 0.0001$ & $p < 0.01$). But no significant difference was found for easy videos among the three groups.

This observation matches very well with our expectations to different video difficulties: the built-in linear interpolation tool for manual annotation is very efficient in tracking a single face continuously, but **AI recommendations can dramatically reduce task time when tracking multiple faces simultaneously in medium and hard videos**. This finding allows the system designer to optimize efficiency further: if we know a certain portion of the data has one or fewer people in each frame, it would be reasonable to bypass the AI pre-annotation to save on the GPU budget.

User expertise. When solely considering the user expertise factor, we were surprised that veteran workers are overall significantly slower than novice workers in both parts of the study (Welch, Part 1: $F_{1,1380} = 85.6, p < 0.0001$, Part 2: $F_{1,665} = 22.2, p < 0.0001$)! However, if we consider how people are paid, this result would be a reasonable optimization given the incentives – veteran workers know the acceptable work pace, so they do not need to work

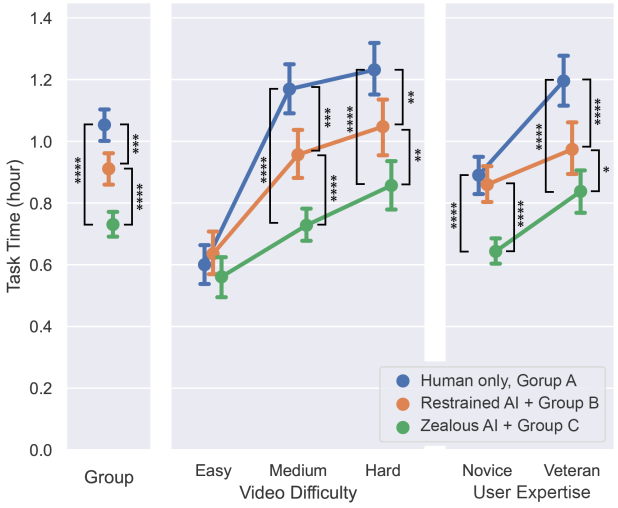


Figure 6: Average annotation time for a single video in Part 1. Lower is better. Error bars represent the 95% confidence interval. Treatment Group A used a baseline manual method and the annotators in Groups B and C reviewed restrained and zealous AI recommendations in Part 1. Groups B & C included the GPU time used to calculate the AI recommendations.

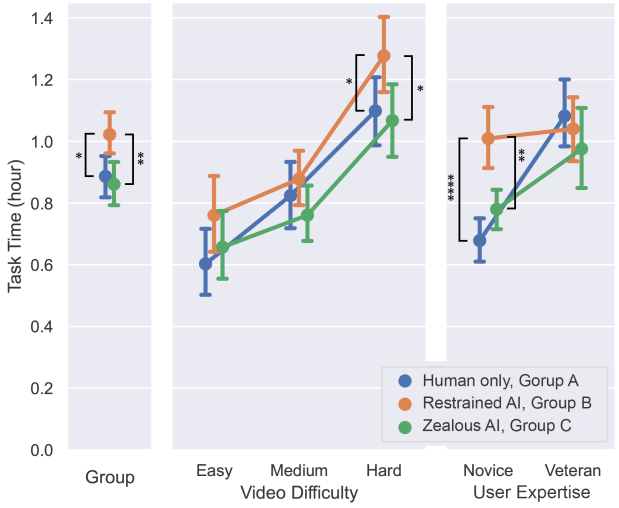


Figure 7: Average annotation time for a single video in Part 2. After working 2-3 weeks on Part 1, every worker annotated another 12 videos in Part 2 but all used the same manual tool without AI recommendations. We no longer see a significant difference between Groups A & C but Group B is now slower in hard videos, mainly caused by novice workers.

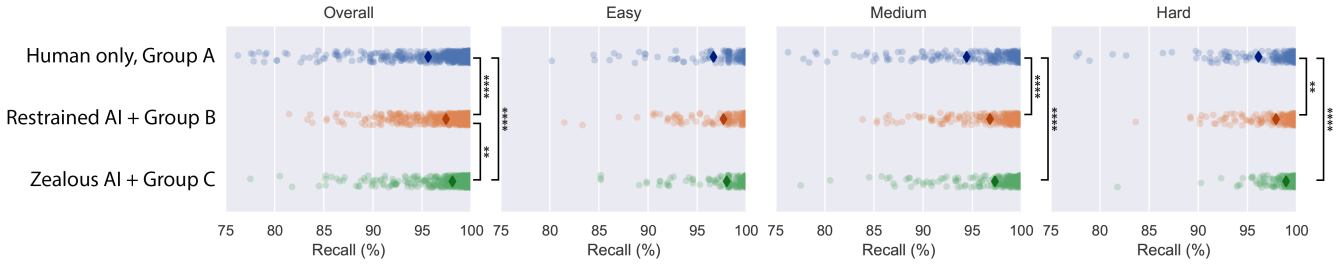


Figure 8: The recall distribution of annotated videos in Part 1. For the purpose of visualization clarity, we plot the 75-100% range in all recall distributions, which omits maximally 2% of outlier cases. Higher recalls and a "shorter tail" are better. The average recall is marked with a darker diamond. The recall distribution reveals the likelihood of having a higher quality result, an insight needed to analyze results from crowdworkers. E.g., in hard videos (right), annotations from "zealous AI + Group C" have a shorter tail than other methods, as expected, the high-recall zealous AI recommendations make it easier for more people to achieve higher recalls especially when people's attention are pushed to the limit when there are three or more faces to track across many frames simultaneously.

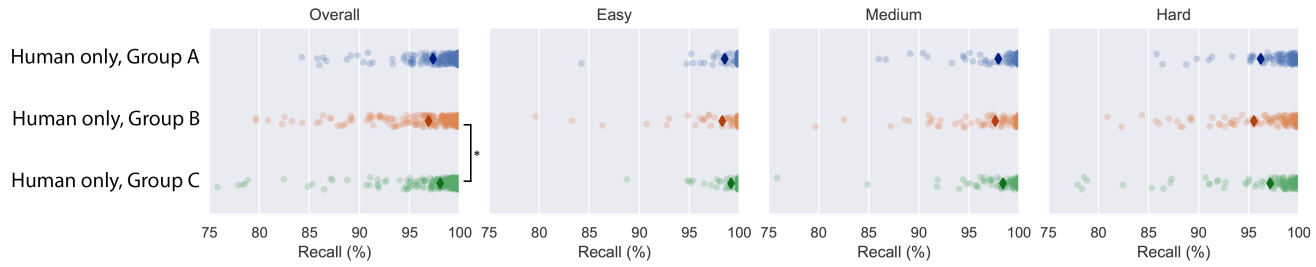
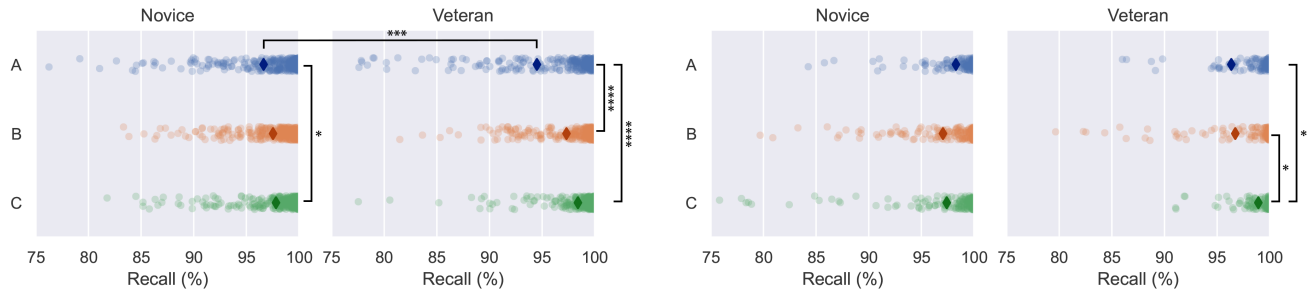


Figure 9: The recall distribution of annotated videos in Part 2. The previously human-AI collaborative Groups B & C no longer have access to the AI recommendations so they used the same manual method that Group A have been using. The overall subplot (left) shows visible longer tails from these two groups, especially Group C in hard videos (right), indicating a discrepancy in individuals' performance now without the help from AIs.



(a) Part 1 (different methods between groups) **(b) Part 2 (same method: human only)**

Figure 10: Recall distribution of annotated videos split by user expertise. Figure (a) shows both human-AI Groups B & C gained advantage over the manual method Group A mainly through veteran workers. The longer tails in Figure (b, novice) provide a new perspective to interpret Group C's long tails in Figure 9 (Overall) that the performance discrepancy is mostly caused by novice workers after they lost access to AI recommendations.

faster than necessary. We further discussed worker's incentives in Section 6.2.

When we consider the group and user expertise factors at the same time, as shown in Figure 6 (right), both novice and veteran workers in Group C who reviewed the zealous AI recommendations were significantly faster than the baseline (Bonferroni $p < 0.0001$ & $p < 0.0001$), while only the veterans in Group B finished faster (Bonferroni $p < 0.0001$). This allows us to infer that, unlike the restrained AI that helps veterans more, **the zealous AI can consistently improve user completion time for both novice and veteran annotators**.

5.3 Q3: Which AI helps annotators achieve higher recall?

From the F1 scores in Figure 5 we know that both AI-assisted methods yield significantly higher-quality annotations than the baseline method (compared in Section 5.1), yet we saw no clear winner between the two human-AI teams. Because recall is paramount in video anonymization tasks, we analyze Group B and C's recall performance in detail.

Figure 8 shows that Group C, the annotators who reviewed zealous AI recommendations, have an overall significant advantage over Group B, which reviewed restrained AI recommendations (Games-Howell $p < 0.01$). Interestingly, we noticed **a visible shorter tail** in Group C's recall distribution in hard videos (Figure 8, right). This observation matches the very nature of zealous AI – giving more recommendations, even low-confidence ones, so the human teammate is less likely to miss a face. This strategy is especially effective in hard videos because tracking too many faces simultaneously pushes the user's attention to its limit. **Zealous AI's superfluous recommendations allow the user to focus on the action of *reject*, rather than searching for missing faces and then *solve***.

Taking user expertise into account, Figure 10 (a) reveals that while both AIs improved the veterans' recall performance compared to the baseline Group A (Bonferroni A/B: $p < 0.0001$, A/C $p < 0.0001$), for novice workers, we only saw a significant advantage of Group C over Group A (Bonferroni $p < 0.048$). It corroborates our previous finding on completion time that "the zealous AI can consistently improve both novice and veteran annotators" and extends the statement to higher recalls percentages as well.

5.4 Q4: Will collaborating with an AI improve or hurt user skills?

Should the annotators lose access to their AI teammates in the future, how will they perform? While we are interested in improving human-AI team performance, we should also seriously consider how the prior human-AI collaboration experience would affect people's skills in the long run before deploying a new system.

To find out, we removed AI recommendations from Groups B and C in Part 2, so all groups now work with the manual tool that they have always been using for other projects. It took most annotators two to three weeks to complete Part 1 of the study. For the sake of interpreting the results of Part 2, we can consider this period a training period and their performance in Part 2 showcasing the effect of this medium-term training effort.

Both Groups B & C collaborated with their perspective AI teammates for 2-3 weeks, **but the restrained-AI-trained annotators in Group B performed worse than their peers in different ways** – the novice workers were significantly slower than both A & C, especially in hard videos. The veteran workers' annotations had lower recall percentages than the zealous-AI-trained workers in Group C.

Completion time. Figure 7 shows the task completion time of Part 2's 12 new videos without AI recommendations. In all video difficulties, Group C, annotators who previously worked with the zealous AI in Part 1, managed to finish as quickly as Group A, the annotators who were trained using the very manual method now in deployment for all groups. It shows that training with zealous AI recommendations does not negatively affect users' task completion time on subsequent manual tasks.

However, we were surprised to see that Group B annotators trained with the restrained AI became overall significantly slower than Groups A & C (Tukey-HSD A/B: $p < 0.021$, B/C: $p < 0.01$), and more specifically in hard videos (Bonferroni A/B: $p < 0.044$, B/C: $p < 0.013$). Figure 7 (right) shows that the effects stem mainly from the novice users (Bonferroni A/B: $p < 0.0001$, B/C: $p < 0.01$).

Recall. On annotation quality, Figure 9 shows the Groups B annotators, trained by the high-precision restrained AI now produce lower-recall annotations (Games-Howell $p < 0.05$) than Group C which was trained with the high-recall zealous AI. The user expertise breakdown shows the effect mostly comes from the veteran workers (Bonferroni $p < 0.028$).

What caused the negative training effect from the restrained AI?. We would think that annotators in Group B should perform better in Part 2 of the study now that they have to manually annotate – they practiced more on manually adding missing faces (***solve***) working with the restrained AI recommendations. In contrast, Group C which trained with the zealous AI focused on ***rejects***. However, the experiment results show otherwise. Why was only Group B negatively affected? We believe there are two main factors in play:

1) Not optimizing the AI teammate for the human-in-the-loop workflow. Despite the fact that both AIs used the same face-detection model to generate the untracked bounding boxes in each frame for the tracker to process, the restrained AI recommendations were produced by ByteTrack [64] which is designed for autonomous tasks rather than for human-AI collaboration. We observed various issues using that tracker directly in pilot studies, so we proposed the FPR tracker specifically for a human-in-the-loop workflow with many optimizations with human users in mind (discussed in Section 3.3). Given the fact that only novice users became much slower in Part 2 of our study while veterans, who are more familiar with the annotation tool, were unaffected, we strongly believe that the negative transfer effect can be linked back directly to training with the restrained AI.

2) Not optimizing the AI teammate for the task. Recommendations from the high-precision restrained AI are naturally lower in recall than the zealous AI, i.e., the restrained AI missed more faces. Users who worked with such an AI for 2-3 weeks might actually have gotten used to the AI's pre-annotated videos (in Part 1) as "acceptable quality", thus matching their annotation effort with the less optimal recall when working on their own in Part 2. On the other hand, the zealous AI recommendations – the high-recall AI more exhaustively demonstrated all faces that should be annotated, potentially raising the quality standard for the task.

In conclusion, various pieces of evidence from Part 2 of our study showed that despite decent human-AI team performance when working with the AI, naively deploying an AI system into a human-AI setting without considering the nature of the task or without optimizing it for the human teammates could lead to negative effects and potential deskilling of the users.

6 DISCUSSION

6.1 The key to forming a strong human-AI team

We propose the restrained AI and the zealous AI to depict the tradeoff between precision and recall as two characteristics that have the potential of becoming advantages in human-AI teams if used properly. By actually using the annotation tools and watching annotators' screens for many hours, we observed that annotators need much less effort in improving precision than recall in a model-assisted annotation task, i.e., rejecting an incorrect box is much easier than adding a missing box, thus we should delegate more effort in improving recall to the AI so human only handles the most difficult boxes that the AI missed (Figure 1c).

We think **an important insight from this study is that it is worthwhile to identify the complementary strengths of both human and AI teammates through an in-depth analysis of the task at hand**. While our observations can improve real-world object detection and tracking annotation tasks, in which correcting false-positive errors are easier for human, another task with a higher cost in correcting such errors could lead to different or even opposite optimizations. Working closely with end users can inspire us to decompose the AI's different properties (in our case precision and recall) and turn them into advantages to complement human skills. We hope this study can motivate fellow researchers to rethink existing AI assistance designs or at least the design for other video annotation tasks.

6.2 Can AI teammates set the quality lower bound in a crowdsourcing setting?

We identified and rejected a single veteran user who submitted the majority of the low-quality annotations. This is an unexpected yet not surprising finding in a crowdsourcing setting: when paid at a flat hourly rate, people are not necessarily motivated to work faster. When lacking a quality-based performance evaluation mechanism, people are not necessarily motivated to push for "better-than-sufficient" quality.

However, could there be other users not making an effort in Groups B or C as well but not being identified? Because the two AIs have pre-annotated the videos in decent quality ($F1 > 90\%$), it's hard to tell if someone is actually happy with the AI's recommendations or is not pushing for even better quality.

What we know for sure is that such low-quality submissions, intentional or unintentional, will certainly appear in other real-world crowdsourcing tasks. However, in absence of ground truth, we won't be able to identify them in a real-world setting. It is also very costly to identify bad submissions – ImageNet asks 10 votes for each image [17], and Microsoft COCO asks 3-5 workers to judge each segmentation [33].

Could the AI recommendations have played a critical role in preventing low-quality submissions, i.e., setting a lower bound for the annotation quality? While not verified in our study, this observation could provide yet another strong motivation for human-AI collaboration in a crowdsourcing setting. We encourage fellow researchers to consider this in future experiment designs.

6.3 Seemingly contradictory survey results

Figure 11 shows user responses to the Survey 1 questions, with each group's five-point Likert scale responses normalized to 100%. 0% indicates no preference. Specifically, question *S1-6* (Figure 12) indicates that users from both human-AI teams, B and C, think that working with the AI makes the task easier than annotating manually. However, in Survey 2 (Figure 14), after users have tried both the AI-assisted and the Manual methods on the same task of similar videos, they express higher preference towards the Manual method regarding multiple aspects. As users took each survey immediately after Part 1 and Part 2 respectively, they might prefer

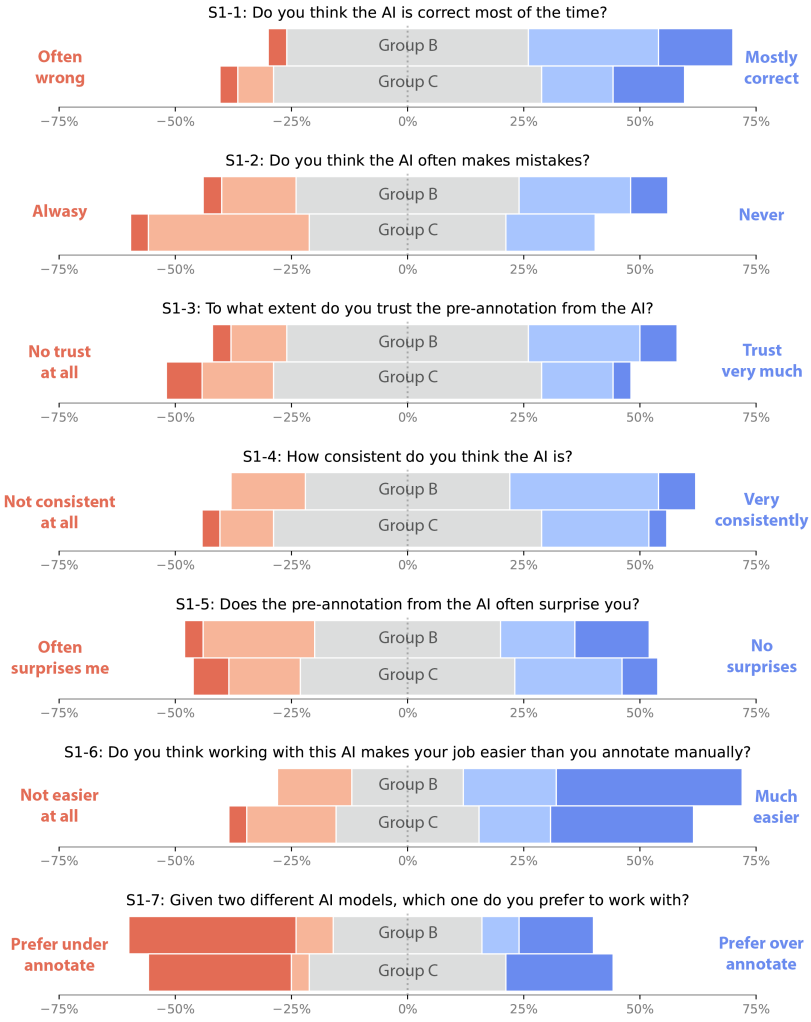


Figure 11: Survey 1 (post-Part 1). We normalize each group's five-point Likert scale responses to 100%. 0% indicates no preference. In Part 1's between-subject study, annotators from Groups B & C only worked with a single AI they were assigned to, so we do not compare the responses between B with C.

Figure 12: Question S1-6 in Survey 1 indicates significant result. The five-point Likert scale responses are converted to [-2, 2] with mean and 95% CI plotted.

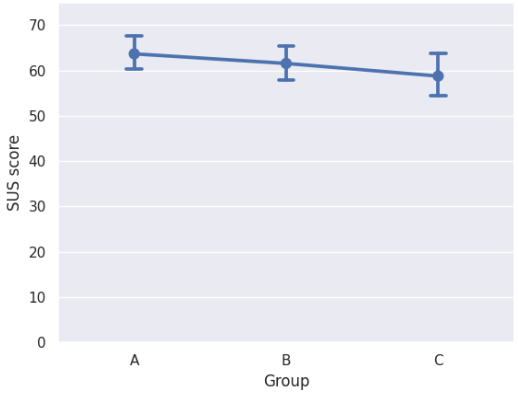


Figure 13: A System Usability Scale (SUS) survey was administered at the conclusion of Part 1 of the study. But we saw no significant difference between the groups. Similar to Survey 1 in Figure 11, participants tend to provide neutral feedback.

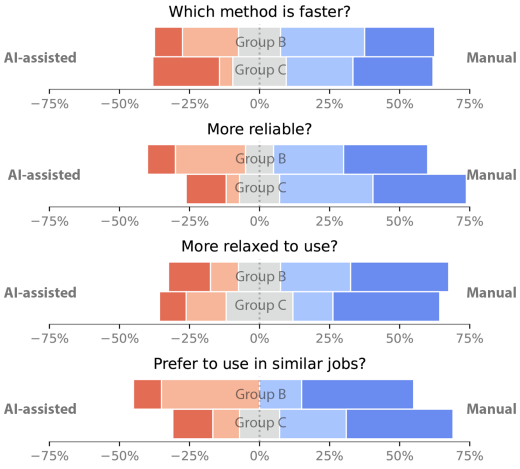


Figure 14: Survey 2. Unlike Survey 1 in which annotators answered questions without comparison, Groups B & C have used both AI-assisted and Manual methods at the end of Part 2. Thus this part of the study is close to a within-subject design where the independent variables are the AI-assisted and Manual method.

the method they just used, but these responses from Groups B & C are in conflict with their continued higher recall in Part 2.

Comparing Figure 8 (left) with Figure 9 (left), we observe that the Group B & C annotators who had shorter tails in recall distribution than Group A in Part 1 ended up with longer tails in Part 2 after they lost the AI's assistance. It shows that a fraction of low-performing users were apparently held at a higher standard by the AI recommendations, and when the AI teammate was gone, they returned to their preferred standard.

This observation might help explain the higher performance with the AI-assisted method but higher user preference for the Manual method. It also reminds us to take users' incentives into account when designing user preference questions in empirical studies – It is well-known that the most favorable method is not necessarily the best performing method. We administered the System Usability Scale (SUS) survey and saw a trend to support this point in Figure 13, but the results are not significant.

6.4 Limitations and Future Work

What are the conditions for which our findings hold? This study investigated a single high-stakes task that met the two aforementioned conditions: 1) either recall or precision is far more important than the other, and 2) the complementary strengths of human and AI can be identified and the precision-recall tradeoff can be exploited to improve the important metric for the given task. We proposed and observed that delegating more recall effort to the zealous AI can significantly improve team performance, which was mainly motivated by our observation that ***reject*** is much easier than ***solve*** for humans in AI-assisted annotation. Will our findings still hold if ***reject*** is easier than ***solve*** in a different task? What about precision-demanding tasks? We would love to see more HCI and AI researchers conduct latitudinal studies in multiple recall- or precision-demanding tasks to test and refine our findings.

Tasks without high-performance models. Face detection is a well-studied problem with high-performance AI models. While we showed in Figure 5 that the AI and human can reach similar performance in this task to achieve complementary team performance, will our findings stand if either the human's performance or the AI's recommendations are much worse than the other? What is the lower bound F1 score limit for either the human or the AI to maintain complementary team performance? What are the F1 or precision/recall conditions for other researchers to reproduce our findings?

Limitation from data and participants. We used a subset of realistic, egocentric video dataset [23] in this study to measure with the skill of locating faces – a human instinct that comes with relatively small inter-personal differences. However, could our findings still play a major role if the task was to identify and track other objects that could have larger inter-personal differences? Furthermore, working with amateurs via crowdsourcing platforms would introduce larger variances between individuals than with the professional workers employed in this study. Researchers would need to put more effort into benchmarking or measuring the human factor in such follow-up studies.

Incentives for users to actively perform better. We discussed in Section 6.3 observations that methods with better performances are not necessarily favored by the users. I.e., the users were involuntarily pushed to have higher performance by their AI teammates. From a system designer's perspective, the AI teammate should help users to voluntarily perform better given the right incentives.

7 CONCLUSION

In this work, we look beyond the accuracy of AI recommendations to explore a new direction to improve human-AI team performance – the tradeoff between precision and recall in model tuning. We propose the concept of restrained and zealous AIs for high-precision and high-recall recommendations and conduct an experiment with 78 professional annotators to compare if and how the different AI recommendations can affect team performance in high-stakes human-AI collaboration. This work serves as a new example of complementary team performance in a large-scale realistic setting.

An in-depth analysis of the task helped us identify an optimization opportunity to harness complementary human and AI strengths utilizing the tradeoff between precision and recall in the AI model tuning – given the importance of recall in face anonymization and the higher cost for humans to improve the recall in video annotation. We showed that the proposed high-recall zealous AI helps annotators achieve significantly better performance than the high-precision restrained AI in the video annotation task. Our follow-up study removed AI assistance and observed potentially negative training effects to the users – if an AI is naively paired with humans without optimizing it for the task at hand or for the human-AI workflow. We feel these findings have important implications for the design of AI assistance in recall-demanding scenarios. We hope this work can also inspire researchers to look for additional directions in model tuning to improve human-AI team performance.

ACKNOWLEDGMENTS

This work was partially done during the first author's research internship at Appen. We thank all anonymous reviewers for their insightful comments and suggestions. We thank Huan Liu for her support and hand-drawn figures, Yue He and Yuedong Wang for their time and discussion, and members of the UCSB Four Eyes and Expressive Computation Laboratories for their helpful feedback. This work was partially supported by ONR awards N00014-19-1-2553 and N00014-23-1-2118.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[2] 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. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 13 (May 2021), 11405–11414. https://ojs.aaai.org/index.php/AAAI/article/view/17359

[3] 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. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 2–11. https://ojs.aaai.org/index.php/HCOMP/article/view/5285

[4] 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. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

[5] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[6] Mustafa Bilgic and Raymond J Mooney. 2005. Explaining recommendations: Satisfaction vs. promotion. In *Beyond personalization workshop, IUI*, Vol. 5. 153.

[7] 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*. 160–169. https://doi.org/10.1109/ICHI.2015.26

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 213–229.

[9] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 1721–1730. https://doi.org/10.1145/2783258.2788613

[10] Tathagata Chakraborti and Subbarao Kambhampati. 2018. Algorithms for the Greater Good! On Mental Modeling and Acceptable Symbiosis in Human-AI Collaboration. arXiv:1801.09854 [cs.AI]

[11] Yimin Chen, Xinjie Zhang, and Junmin Wang. 2021. Robust Vehicle Driver Assistance Control for Handover Scenarios Considering Driving Performances. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 51, 7 (2021), 4160–4170. https://doi.org/10.1109/TSMC.2019.2931484

[12] Peng Chu, Jiang Wang, Quanzeng You, Haibin Ling, and Zicheng Liu. 2021. TransMOT: Spatial-Temporal Graph Transformer for Multiple Object Tracking. arXiv:2104.00194 [cs.CV]

[13] Mark Colley, Benjamin Eder, Jan Ole Rixen, and Enrico Rukzio. 2021. Effects of Semantic Segmentation Visualization on Trust, Situation Awareness, and Cognitive Load in Highly Automated Vehicles. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 155, 11 pages. https://doi.org/10.1145/3411764.3445351

[14] Gabriela Csurka, Diane Larlus, and Florent Perronnin. 2013. What is a good evaluation measure for semantic segmentation?. In *Procedings of the British Machine Vision Conference 2013*. British Machine Vision Association, Bristol, 32.1–32.11. https://doi.org/10.5244/C.27.32

[15] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. 2020. TAO: A Large-Scale Benchmark for Tracking Any Object. In *European Conference on Computer Vision*. https://arxiv.org/abs/2005.10356

[16] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning (ICML '06)*. Association for Computing Machinery, New York, NY, USA, 233–240. https://doi.org/10.1145/1143844.1143874

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Miami, FL, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[18] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019. RetinaFace: Single-stage Dense Face Localisation in the Wild. *CoRR* abs/1905.00641 (2019). arXiv:1905.00641 http://arxiv.org/abs/1905.00641

[19] Abhishek Dutta and Andrew Zisserman. 2019. The VIA Annotation Software for Images, Audio and Video. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 2276–2279. https://doi.org/10.1145/3343031.3350535

[20] Zafer Erenel and Hakan Altınçay. 2013. Improving the precision-recall trade-off in undersampling-based binary text categorization using unanimity rule. *Neural Computing and Applications* 22, 1 (May 2013), 83–100. https://doi.org/10.1007/s00521-012-1056-5

[21] Shi Feng and Jordan Boyd-Graber. 2019. What Can AI Do for Me? Evaluating Machine Learning Interpretations in Cooperative Play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 229–239. https://doi.org/10.1145/3301275.3302265

[22] Deepak Gopinath, Jonathan DeCastro, Guy Rosman, Emily Sumner, Allison Morgan, Shabnam Hakimi, and Simon Stent. 2022. HMIway-Env: A Framework for Simulating Behaviors and Preferences To Support Human-AI Teaming in Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 4342–4350.

[23] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2022. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*.

[24] Joseph Juran, Frederick Taylor, Walter Shewhart, Edward Deming, Philip Crosby, Kaoru Ishikawa, Armand Feigenbaum, Genichi Taguchi, and Elihu Goldratt. 2005. Quality control. *Joseph M. Juran: Critical Evaluations in Business and Management* 1 (2005), 50.

[25] Ece Kamar, Severin Hacker, and Eric Horvitz. 2012. Combining Human and Machine Intelligence in Large-Scale Crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1* (Valencia, Spain) *(AAMAS '12)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 467–474.

[26] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. 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, 1–14. https://doi.org/10.1145/3313831.3376219

[27] Isaak Kavasidis, Simone Palazzo, Roberto Di Salvo, Daniela Giordano, and Concetto Spampinato. 2014. An innovative web-based collaborative platform for video annotation. *Multim. Tools Appl.* 70, 1 (2014), 413–432. https://doi.org/10.1007/s11042-013-1419-7

[28] Matthew Kay, Shwetak N. Patel, and Julie A. Kientz. 2015. How Good is 85%? A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 347–356. https://doi.org/10.1145/2702123.2702603

[29] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300641

[30] 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 *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300717

[31] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 54, 18 pages. https://doi.org/10.1145/3491102.3501999

[32] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology* 49, 4 (2013), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014 (Lecture Notes in Computer Science)*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[34] Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. 2014. Optimal Thresholding of Classifiers to Maximize F1 Measure. In *Machine Learning and Knowledge Discovery in Databases (Lecture Notes in Computer Science)*, Toon Calders, Floriana Esposito, Eyke Hüllermeier, and Rosa Meo (Eds.). Springer, Berlin, Heidelberg, 225–239. https://doi.org/10.1007/978-3-662-44851-9_15

[35] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama,

Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 78, 16 pages. https://doi.org/10.1145/3411764.3445562

[36] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[37] Alan Lundgard, Yiwei Yang, Maya L. Foster, and Walter S. Lasecki. 2018. Bolt: Instantaneous Crowdsourcing via Just-in-Time Training. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3173574.3174041

[38] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. 2022. TrackFormer: Multi-Object Tracking with Transformers. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[39] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. 2016. MOT16: A Benchmark for Multi-Object Tracking. *arXiv:1603.00831 [cs]* (March 2016). http://arxiv.org/abs/1603.00831 arXiv: 1603.00831.

[40] Fred Morstatter, Liang Wu, Tahora H. Nazer, Kathleen M. Carley, and Huan Liu. 2016. A new approach to bot detection: Striking the balance between precision and recall. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 533–540. https://doi.org/10.1109/ASONAM.2016.7752287

[41] Bonnie M. Muir. 1987. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies* 27, 5 (1987), 527–539. https://doi.org/10.1016/S0020-7373(87)80013-5

[42] Yiftach Nagar and Thomas W Malone. 2011. Making business predictions by combining human and machine intelligence in prediction markets. In *International Conference on Information Systems*. Association for Information Systems.

[43] A. Neubeck and L. Van Gool. 2006. Efficient Non-Maximum Suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, Vol. 3. 850–855. https://doi.org/10.1109/ICPR.2006.479

[44] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, et al. 2019. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ digital medicine* 2, 1 (2019), 1–10.

[45] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 83 (apr 2022), 22 pages. https://doi.org/10.1145/3512930

[46] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[47] Yutaka Sasaki et al. 2007. The truth of the F-measure. *Teach tutor mater* 1, 5 (2007), 1–5.

[48] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. 2020. TransTrack: Multiple Object Tracking with Transformer. arXiv:2012.15460 [cs.CV]

[49] Niels Van Berkel, Jeremy Opie, Omer F. Ahmad, Laurence Lovat, Danail Stoyanov, and Ann Blandford. 2022. Initial Responses to False Positives in AI-Supported Continuous Interactions: A Colonoscopy Case Study. *ACM Trans. Interact. Intell. Syst.* 12, 1, Article 2 (mar 2022), 18 pages. https://doi.org/10.1145/3480247

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[51] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. 2019. MOTS: Multi-Object Tracking and Segmentation. *arXiv:1902.03604[cs]* (2019). http://arxiv.org/abs/1902.03604 arXiv: 1902.03604.

[52] Carl Vondrick, Donald J. Patterson, and Deva Ramanan. 2013. Efficiently Scaling up Crowdsourced Video Annotation - A Set of Best Practices for High Quality, Economical Video Labeling. *Int. J. Comput. Vis.* 101, 1 (2013), 184–204. https://doi.org/10.1007/s11263-012-0564-1

[53] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H. Beck. 2016. Deep Learning for Identifying Metastatic Breast Cancer. arXiv:1606.05718 [q-bio.QM]

[54] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[55] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to Complement Humans. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (Jul 2020). https://doi.org/10.24963/ijcai.2020/212

[56] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang 'Anthony' Chen. 2020. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. 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, 1–13. https://doi.org/10.1145/3313831.3376807

[57] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2016. WIDER FACE: A Face Detection Benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[58] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300509

[59] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User Trust Dynamics: An Investigation Driven by Differences in System Performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York, NY, USA, 307–317. https://doi.org/10.1145/3025171.3025219

[60] Jenny Yuen, Bryan Russell, Ce Liu, and Antonio Torralba. 2009. LabelMe video: Building a video database with human annotations. In *2009 IEEE 12th International Conference on Computer Vision*. 1451–1458. https://doi.org/10.1109/ICCV.2009.5459289

[61] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 114, 28 pages. https://doi.org/10.1145/3491102.3517791

[62] Rui Zhang, Nathan J. McNeese, Guo Freeman, and Geoff Musick. 2021. "An Ideal Human": Expectations of AI Teammates in Human-AI Teaming. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 246 (jan 2021), 25 pages. https://doi.org/10.1145/3432945

[63] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. 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* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[64] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2022. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. (2022).

[65] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv preprint arXiv:2010.04159* (2020).