Exploring Collective Theory of Mind on Pedestrian Behavioral Intentions

Md Fazle Elahi
melahi@iupui.edu
Indiana University-Purdue University Indianapolis
Indianapolis, IN, USA

Tianyi Li
li4251@purdue.edu
Purdue University
West Lafayette, IN, USA

Renran Tian
rtian@iupui.edu
Indiana University-Purdue University Indianapolis
Indianapolis, IN, USA

ABSTRACT

While crowdsourcing is commonly used for objective labeling, eliciting subjective annotations, like estimating mental states or perception of other's intention, remains challenging. This study investigates crowdsourcing's potential to predict pedestrian behavioral intentions. We recruited 120 participants to predict pedestrian intentions at different prediction horizons in 24 diverse videos. Our findings revealed that the status-quo bias significantly impacts intention estimation. Specifically, when asked what status the pedestrian will be, predictions inclined towards current state's continuation over transition, with an overall accuracy of 53% at one-second prediction length on a balanced dataset. Rephrasing the annotation question mitigates this bias and improved the estimation accuracy to 79% for one-second ahead predictions, though accuracy drops with longer horizons and is affected by pedestrian actions and contextual information. Overall, this study provides insights into the factors affecting collective estimation of pedestrian intentions and aims to improve crowdsourcing cognitive labels for training better AV-pedestrian interaction algorithms.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **HCI design and evaluation methods**; • **Applied computing** → **Psychology**; • **Computing methodologies** → **Theory of mind**.

KEYWORDS

Human Cognitive State, Crowdsourcing, Behavior Prediction

ACM Reference Format:
Md Fazle Elahi, Tianyi Li, and Renran Tian. 2024. Exploring Collective Theory of Mind on Pedestrian Behavioral Intentions. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3613905.3650930

1 INTRODUCTION

The integration of AI and machine learning in everyday environments such as vehicles and workplaces requires endowing these systems with human-like comprehension of communication, an understanding of interpersonal information conveyance methods with the recognition of the pivotal role played by contextual information [4, 51]. From a computational standpoint, one of the linchpins in this evolving interaction model is the capacity to comprehend and emulate human cognition in intricate naturalistic settings [37].

One prominent example is the navigation of Autonomous Vehicles (AVs) alongside human-controlled vehicles and pedestrians where deciphering human intentions and their actions is challenging and widely discussed [11, 12, 15, 21, 39, 42]. Reports from public-road AV tests shows that 80% to 90% of automation failures occur in city streets, often due to sub-optimal interactions with pedestrians [5]. Furthermore, typical kinematics-based pedestrian-detection algorithms has short response times of 0.5 to 1.5 seconds [44], leaving little room for human drivers to react and thus causing fatalities including pedestrian accidents [6, 35]. Hence, better prediction of pedestrian behavioral intentions is indispensable for safer urban automated driving.

To enhance algorithm performance, many benchmark datasets [17, 24, 40, 41] and prediction algorithms towards pedestrian behaviors, from actions, trajectories, to intentions [7, 18, 53] are being actively developed. The prediction of pedestrian intention primarily adopts two [54] distinct approaches: one focuses on *predicting the future actions to represent the current behavioral intention* [17, 53], while the other employs *crowdsourcing to label pedestrian crossing intentions* and use those labels directly as ground truth to develop sequential prediction models [24, 40]. Both approaches have many models developed, but the reliability of intention labels remains arguable, and the overall pedestrian intention prediction accuracy is not yet satisfactory enough to support driving decisions.

Given the current limitations in algorithmic approaches, it becomes crucial to explore the human ability to predict pedestrian intentions, bridging the gap in decision-making where technology falls short. In the broader domain, while collective intelligence and crowdsourcing excel in various physical world tasks, their efficacy in cognitive tasks like intention estimation in non-verbal human-human interactions varies by tasks [2]. Understanding the role of collective intelligence in intention prediction is critical to improve the crowdsourcing process towards intention labels and corresponding AI algorithm evaluation for practical implementations.

In this work, we investigate the research question: **"How well can human annotators collectively predict behavioral intentions of pedestrians to conduct certain future actions in recorded naturalistic Pedestrian-Vehicle (PV) interaction scenarios?"** Specifically, we assess how the crowd's intention estimation quality are influenced by these four factors: (1) Action continuity, (2) Current action of the pedestrians, (3) Prediction horizon, and (4) Contextual Cues.

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650930

1. **Action continuity** refers to the behavior to continue the pedestrians' current action. If the current action maintains the same state after the moment of prediction, it is called "CONTINUED." If it transitions into a different state, it is called "TRANSITION" cases.
2. **Current action** refers to the pedestrian's present activity before and at the moment of prediction, which can be walking, running, or standing. For simplicity, we only included data that contains "WALKING" and "STANDING" states.
3. **Prediction Horizon** refers to the duration of future to be predicted. In this work, we investigate how well humans can estimate pedestrians' future actions after one, three, and five seconds from the prediction moment.
4. **Contextual Cues** consist of different traffic and environmental factors. We categorize cases into two groups based on the road locations to represent the different levels of contextual cues. "NON-MIDBLOCK" locations have clear right-of-way controls such as traffic signals, signs, crosswalks, and road markings, whereas "MIDBLOCK" locations lack those.

2 RELATED WORK

2.1 Cognitive Models in Driver-Pedestrian Negotiation

In navigating everyday social interactions, humans possess a capacity known as "Theory of Mind" (ToM) [50] to understand other people's beliefs, desires, and intention, recognizing that these can diverge from their own perceptions [1]. According to simulation theory [19], mirror neurons [10] allow people utilise ToM to analyze, judge, and infer others' actions without direct access to others' mind [38] rather by adopting others' perspectives [31].

When people share similar experiences, they develop similar ToM within a specific context since they operate within similar hierarchical cognitive frameworks [46]. Along this line of work, Shteynberg et al. [45] introduced the concept of the *theory of collective mind*, defining it as "the human ability to perceive others' mental states as aligned with one's own." However, in the context of autonomous driving, the mutual awareness between drivers and pedestrians is constrained by spatial separation (being inside the car vs. on the road) and other factors. To address this, we suggest the term "collective theory of mind," which emphasizes the use of crowdsourcing to gather diverse individual perspectives (from different drivers) without shared awareness, each contributing to a more comprehensive understanding of the collective reality (pedestrian's current behavioral intention state).

The automatic and voluntary mentalization process [20] of driver-pedestrian dynamics can be explained through the lens of event segmentation theory [52], where the drivers spontaneously segment the pedestrian's action [28] dictated by pedestrian's current action and final goal. However, few have explored the validity of applying such theory in predicting pedestrian intentions and actions during interactions.

2.2 Crowdsourcing Subjective Tasks

Recently, increasing research interests have been devoted to subjective crowd annotation tasks [2, 3, 9, 14, 23]. It was found that annotators' culture, background, experiences, educational contexts, and their own opinions all interplay to influence the annotators' perceptions and judgments [16, 48]. Prior strategies to elicit more reliable judgement on subjective matters with crowdsourcing include providing additional contexts such as sequential video clips [26], post-hoc filtering or aggregation methods [22, 25], collecting pairwise comparisons rather than item-level annotations [33], and using probabilistic distribution rather than a one-hot encoded label [49]. Yet, these task design strategies cannot be directly applied to pedestrian intention annotation, due to the complexity of the surrounding traffic and social norms. This study draws on prior subjective annotation research and investigates the feasibility of using crowdsourcing to elicit "collective theory of mind" for pedestrian intention prediction.

3 METHOD

3.1 Dataset: Video Trials Sampling

The study uses 24 videos randomly sampled from the TASI-110 driving dataset [47] ensuring each video includes at least one pedestrian encounter in a naturalistic environment. The driving dataset features diverse location, daylight, and weather conditions (Figure 1a, 1b). While sampling, we controlled three factors: the current action (WALKING/STANDING), action continuity (CONTINUED/TRANSITION), and contextual cues (MIDBLOCK/NON-MIDBLOCK). This resulted in three videos for each of the eight combinations of these factors.

The pedestrian-actions in these videos were labeled every 0.1s as "WALKING" or "STANDING" by two team members using the BORIS tool [13]. For each of the videos containing "TRANSITION" cases, a reference timestamp is set at the moment when the current action changes to another state. For the videos with "CONTINUED" actions, the reference timestamp is randomly sampled between 7 seconds and the video's end-time. Then, each of these 24 videos is clipped at 1, 3, and 5 seconds before the reference timestamp, and a total of 72 video-trials are generated. Two video-trials were discarded for being too short, resulting in final 70 video-trials. The duration of these video-trials ranges from three to nine seconds.

3.2 Prediction Task Design: Question Phrasing and Video Trial Distribution

We employed a nested experiment design where three separate experiments were conducted to investigate the impact of action continuity and prediction horizons. Within each experiment, we ensured an equal distribution of tasks representative of the other conditions. The three experiments used different question phrasing that emphasize different aspects of action continuity. The question phrasing used in the experiments are (also shown in Figure 2):

- Experiment without emphasis on action continuity (**Exp-Pre**): The pedestrian is currently [STANDING/WALKING], **what status** do you think the pedestrian is going to be in [1/3/5] second(s)?
- **E**xperiment with emphasis on **T**ransition (**Exp-T**): The pedestrian is currently [STANDING/WALKING], do you think the pedestrian will **start** [WALKING/STANDING] (transition to the other state) after [1/3/5] second(s)?
- **E**xperiment with emphasis on **C**ontinuity (**Exp-C**): The pedestrian is currently [STANDING/WALKING], do you think

(a) A pedestrian-group is trying to cross in rain with walk sign and crosswalk marking.

(b) A pedestrian crossing in front of a car in the dark in a parking lot.

(c) Attention Check 1: Is the scene recorded at night?

(d) Attention Check 2: Is the color of the car in front is yellow?

Figure 1: (a,b): Screenshots of video samples used in the experiment. The videos cover diverse daylight and weather conditions, single or group of pedestrians, locations, traffic control signals, signs, and road markings. (c,d): These two attention-check questions are unambiguous and same for all the participants.

the pedestrian will **continue** [STANDING/WALKING] after [1/3/5] second(s)?

3.3 Participants and Procedure

The study procedure was reviewed and approved by the Institutional Review Board for Human Subject Research at Indiana University. We recruited U.S. participants (≥ 18 years) with over 1000 approved Human Intelligence Tasks (HITs) and a 95% approval rate, who own a car, as human annotators from Amazon Mechanical Turk (MTurk) [1]. We set a 3-hour time limit and paid $3.5 for annotating up to 20 video trials. For each of the three experiments: Exp-Pre, Exp-T, and Exp-C, we collected 10 independent responses from different human annotators for each of the 70 video trials. This resulted in a total of 700 individual responses in each experiment. These 70 video trials were evenly distributed into four batches among 40 annotators. On average, each annotator is assigned 17 to 18 video trials excluding the two attention-check videos.

After accepting the HIT, each participant first completes a questionnaire containing seven demographic questions [36], including age, gender, residence state, and driving habits like current driving status, last month driven, transportation preference, and weekly driving frequency. After that, they sequentially watch a set of 19 or 20 video-trials where a pedestrian either in "WALKING" or "STANDING" state is designated using a red bounding box. When each video-trial finishes playback, the participants can predict the pedestrian's intention by answering the question-phrasing specific to one of the three rounds of experiments (see Figure 2). The participant can watch the same video multiple times if they wish to and refine their answers. Among the assigned 19 or 20 video-trials, two **attention-check** videos were included as the 7th and 14th annotation tasks of each HIT, as shown in Figure 1c and 1d. The sequence of the other 17-18 video trials is randomized for each annotator. Additionally, we ensured that participants did not view multiple cases originating from the same parent video.

To evaluate the crowd's prediction ability, we computed the prediction metrics at two levels: response-level and case-level for each round of experiments. At **response** level, the metrics are calculated on individual annotations made by each crowd worker on each video trial to measure the performance of the crowd annotators across the whole dataset disregarding the differences among video trials. At **case** level, annotations on the same video trial are aggregated with majority voting. A case is considered to be predicted correctly if five or more of the 10 annotations are correct.

A total of 138 workers participated in three rounds of the experiment: 43 in Exp-Pre, 46 in Exp-T, and 49 in Exp-C. We analyzed the data from 120 workers who passed both attention checks and excluded the data from 18 workers who failed either or both of the attention checks (3 from Exp-Pre, 6 from Exp-T, and 9 from Exp-C). Among these 120 participants, 65% are male (Mean Age=36.2 years, SD=8.48) and 35% are female (Mean Age=39.1 years, SD=13.45). 90% of all participants currently drive with an average of 5 days (SD=1.62) per week.

4 RESULTS

4.1 Effect of Action Continuity

4.1.1 Annotation Performance of the Crowd are Influenced by Status Quo Bias. The results indicate a tendency of humans to predict that pedestrians will continue their current action at both the response and case levels (see Figure 3a), despite an equal distribution of "CONTINUED" and "TRANSITION" scenarios in the experiment. Of the 700 responses, 532 (76%) predicted future states identical to the pedestrian's state at the video's pause in Exp-Pre, and a similar percentage in Exp-C predicted "Yes" (pedestrian will continue their current action). This inclination to predict "CONTINUED" actions is also evident at the **case** level, where 61 out of 70 cases (87%) are predicted as "CONTINUED" in Exp-Pre and a similar result was observed in Exp-C. The prevalence of predictions mirroring the pedestrian's current state is confirmed by a strong positive correlation indicated by the Chi-squared test [$\chi^2(1) = 189.66, p < .001$]. We attribute this observation to the status quo bias [43] among human annotators, where human predictions can be biased towards the current states of pedestrians.

4.1.2 Question Rephrasing has Significant Impact on Bias Mitigation. The results from Exp-T provided positive evidence for the framing effect [34] on the status quo bias [43]. Regarding the accuracy metric, crowds in Exp-Pre and Exp-C conditions predicted around 51% of the cases correctly at the case-level (see Table 1). Conversely, in Exp-T, where annotators were specifically instructed to contemplate the possibility of *transitions* in pedestrian states, the accuracy

[1] https://www.mturk.com/

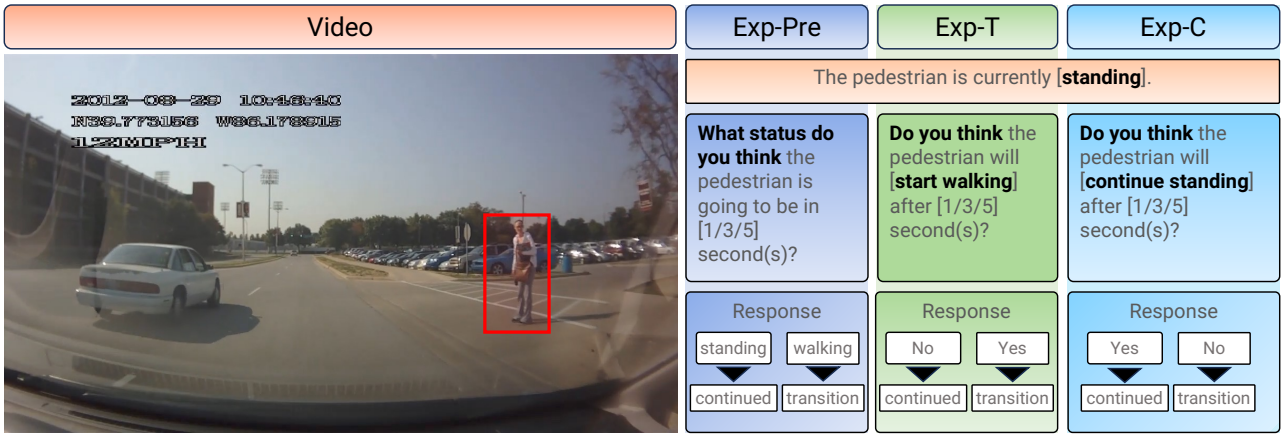
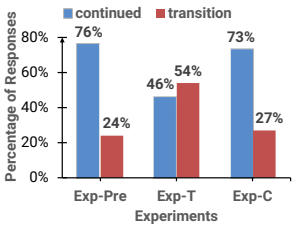
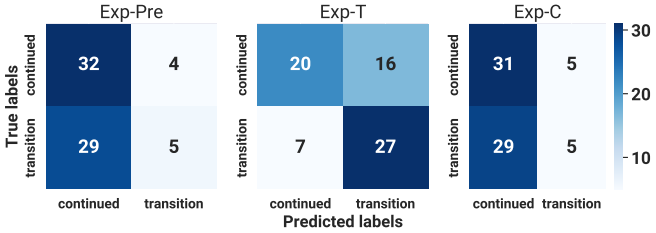


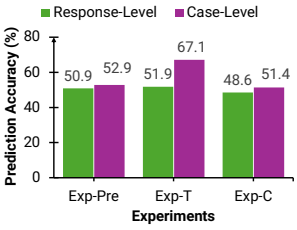
Figure 2: Instructions to predict the behavioral intention of the pedestrian inside the red box in three experiments: Exp-Pre, Exp-T, and Exp-C. In different experiments, participants are asked using three different question phrasings. In Exp-Pre, the question specifically asks to choose the next status from either "STANDING" or "WALKING." Exp-T and Exp-C uses yes/no questions. In Exp-T, the question emphasizes on "opposite action": WALKING. In contrast, in Exp-C, the question emphasizes on the "current action": STANDING.



(a) At response-level, majority predictions are towards CONTINUED cases except for Exp-T.



(b) Confusion matrices at case-level showing biased prediction towards CONTINUED cases, except in Exp-T



(c) Case-level accuracy is higher than that at response-level.

Figure 3: (a) Distributions of predictions at response-level, (b) Confusion Matrices of predictions at case-level. (c) Comparison between accuracy at case-level and at response-level, across three experiments.

became much higher (67%). Similar patterns were observed in Precision and Recall, except the Recall of "CONTINUED" cases. This exception is not surprising, as when crowd workers predict most cases as "CONTINUED", a significant portion of the "CONTINUED" cases — approximately 86.5% (32 out of 37) and 86% (31 out of 36) were predicted correctly. In addition to superior performance, the predictions in Exp-T were also more balanced: of the 47 correctly predicted cases, 43% (20/47) were "CONTINUED" and 57% (27/47) were "TRANSITION"(Figure 3b.) We conducted a Chi-squared test to assess the similarity of responses when different question phrasing were used in the tasks. The responses in Exp-T are significantly different from that of both Exp-Pre [$\chi^2(1) = 129.98, p < .001$] and Exp-C [$\chi^2(1) = 107.25, p < .001$]. Responses in the Exp-C condition are not significantly different from Exp-Pre [$\chi^2(1) = 1.09, p = .29$]. Furthermore, a two-proportion z-test comparing the proportions of correct cases between Exp-T and Exp-C indicated a marginally significant improvement ($z - statistic = 1.89, p = .0585$). This suggests that the changes in question phrasing had a notable but nuanced effect on the accuracy of the responses.

While it is not surprising that the case-level accuracy is higher than response-level accuracy, there was an especially high increase in Exp-T (see Figure 3c.) Aggregating predictions from different people led to an increase in accuracy by 15% in Exp-T and ≈ 2% in both Exp-Pre and Exp-C. This result shows that data aggregation with majority vote is effective for improving the collective accuracy of intention prediction, especially when the instructions emphasize on transitions.

As Exp-T is the least influenced by the status quo bias, we will focus on results from this experiment for the other three factors in the following sections.

4.2 Effect of Current Action

In Exp-T, there are 36 video trials with the current-action labeled as "STANDING" and 34 video trials with the current-action labeled as "WALKING." The influence of the pedestrian's current action, either "WALKING" or "STANDING," on the accuracy of crowd predictions is illustrated in Figure 4a. Notably, when videos depicted pedestrians

Table 1: Prediction Metrics at case level across three experiments when averaged across three prediction horizons. (Highest scores are in bold format. cont: CONTINUED, tran: TRANSITION)

Experiments	Accuracy	Precision (cont)	Precision (tran)	Recall (cont)	Recall (tran)
Exp-Pre	52.86%	52.46%	55.56%	**88.89%**	14.71%
Exp-T	**67.14%**	**74.07%**	**62.80%**	55.56%	**79.41%**
Exp-C	51.43%	51.67%	50.00%	86.11%	14.71%

in the act of walking at the moment of pause, the accuracy of human predictions exhibited modest variability, with performance metrics ranging between 61% and 67% (see Figure 4a). In contrast, when the pedestrians were shown as standing, the accuracy of human predictions increased. Specifically, the predictions were notably more precise in correctly identifying pedestrians who would continue to stand ($Precision_{contd} = 90\%$), and catching most instances where pedestrians would initiate walking ($Recall_{trans} = 93.8\%$.)

To further analyze the impact of a pedestrian's current action on the prediction performance, we applied a Chi-Square test to the data from Exp-T. This test aimed to explore the potential association between the current action of the pedestrian and the correctness of crowd predictions (correct is coded as 1 and incorrect as 0). The results of the test indicated no significant influence of the pedestrian's current action on the predictions $[\chi^2(1) = 0.117, p = .732]$. In other words, the crowd workers did not predict the pedestrian's intentions differently based on their current action.

This discrepancy between observational trends and statistical findings could stem from several factors. It is possible that while the precision and recall metrics show noticeable differences, these differences may not be reflected through the binary correctness coding of the the predictions. Alternatively, the statistical power of the test might not be sufficient to detect the effect, as the sample is now smaller with only results from the Exp-T included.

4.3 Effect of Prediction Horizon

We summarized the influence of prediction horizon on prediction accuracy using the data obtained from Exp-T as shown in Figure 4b. The prediction performance was the highest with the shortest horizon of 1 second. However, an intriguing pattern emerged at longer horizons: predictions made with a 5-second horizon consistently outperformed those with a 3-second horizon across most metrics. This counter-intuitive finding suggests that certain dynamics, perhaps related to human perception or pedestrian movement patterns, become more discernible or predictable over this slightly extended period, thereby enhancing predictive accuracy.

Similar to the "current action" factor, there was no significant differences in correctness across different prediction horizons $[\chi^2(2) = 2.54, p = .281]$. Nonetheless, human prediction shows promising performance in comparison to state-of-the-art algorithms. At prediction horizon of 1 second, the performance of humans and algorithm [30] is comparable with accuracy of $\approx 79\%$. However, while the algorithm's prediction accuracy drops below 59% beyond a 2-second horizon [27], humans can remarkably maintain an accuracy of $\approx 58\%$ even after 3 seconds. This stability demonstrates the potential for training the algorithms beyond 1-2 seconds of prediction horizon.

4.4 Effect of Contextual Cues

Lastly, we analyze the influence of contextual cues on crowd's prediction accuracy at case-level. Each experiment includes 34 video trials at "MIDBLOCK" locations, while 36 trials are at "NON-MIDBLOCK." As opposed to "NON-MIDBLOCK" locations, "MIDBLOCK" locations lack traffic and road factors, generating less contextual cues.

Overall, human predictions had much higher scores across all metrics in scenarios where pedestrians are at NON-MIDBLOCK locations, illustrated in Figure 4c. This suggests that the environmental context, such as traffic control mechanisms and unambiguous traffic norms experienced in the NON-MIDBLOCK settings, plays a crucial role in predicting pedestrian intention.

Similarly, Chi-Square test did not reveal significant difference regarding the correctness of the predictions in these two conditions $[\chi^2(1) = 1.406, p = .236]$.

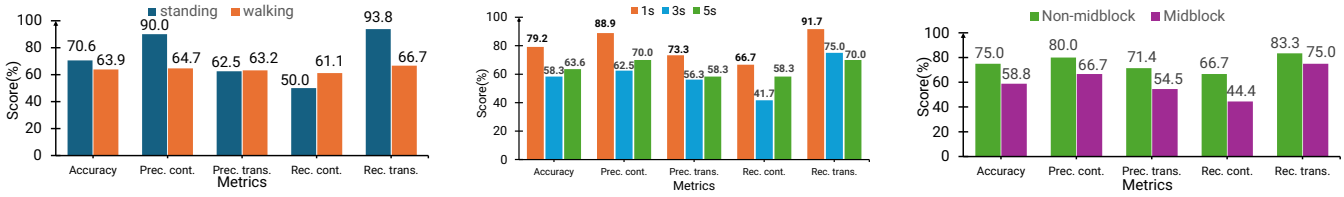
5 DISCUSSION AND TAKEAWAYS

In this section, we discuss the factors affecting crowdsourcing to elicit post-hoc estimations of behavioral intentions based on the current results. For future research, we suggest design strategies for annotation tasks to explore the collective theory of mind, benefiting the HCI community.

Crafting instructions to counteract status quo biases. Our results suggest that post-hoc intention estimations with humans can be challenged by the status quo bias, wherein annotators are inclined to assume subjects will maintain their existing behavior. As a result, the accuracy of these predictions is only marginally better than the chance level, at 52.86%. Moreover, majority of the correct predictions fall into the category of predicting that behaviors will continue as they are. To address this, we implemented instructions that emphasize behavioral transitions. This approach effectively reduced the bias, guiding annotators away from default assumptions of behavioral continuity without unduly influencing them towards contrary predictions. This design change improved the estimation accuracy of the annotators by 14.28% compared to the conditions where the bias is not mitigated.

In addition to modifying instructional phrasing, future work is needed to explore other interventions, particularly in the realm of interface design, to mitigate human cognitive biases in intention estimation. Another avenue worthy of exploration is incorporating interactive and engaging elements, such as gamifying [29, 32] the tasks. For example, asking annotators to role-play as the subjects they are annotating could foster a deeper, more logical understanding of the scenarios, moving beyond instinctive reactions.

Smaller steps for choosing prediction horizons. Our findings indicate that a longer prediction horizon of three to five seconds does not necessarily decrease the accuracy of intention prediction. This suggests a promising future task design consideration where



(a) Performance metrics are higher when current action is "STANDING."

(b) Performance metrics are the highest when the prediction horizon is 1 second.

(c) At NON-MIDBLOCK, all the metrics are higher than those at MIDBLOCK.

Figure 4: Case-level metrics at different levels of (a) Current Action, (b) Prediction Horizon, (c) Contextual Cues in Exp-T.(Prec.: Precision, Rec.: Recall, cont.: CONTINUED, trans.: TRANSITION)

the prediction horizons can range at smaller intervals of 100ms when the time-window is closer to Time-To-Event (TTE). Further, reliable crowd-sourced labels can increase the prediction horizon of existing algorithms from 1-2 seconds [27] to over three seconds. This allows passengers more time to react, thereby enhancing safety and trust. Another potential area for exploration lies in identifying scenarios where human judgment and algorithmic approaches excel or falter in terms of predicting further ahead, with the aim of uncovering opportunities to effectively complement these two methods.

Scaffolding in less favorable cases. Our results suggest that the intent prediction performance can be sensitive to factors such as pedestrian's current action and contextual cues in the videos. For instance, the prediction accuracy is lower when the current action is WALKING as opposed to STANDING. Additionally, the accuracy decreases at NON-MIDBLOCK locations compared to MIDBLOCK locations. In other words, the intention prediction performance may vary dramatically in different cases. Future task design should consider scaffolding the intention predictions in less favorable cases. An interesting approach will be the integration of AI-assisted tools to recognize and emphasize contextual information and behavioral cues, thereby augmenting the human predictions.

Redefine what to annotate. This study explored ability of humans to collectively predict pedestrian behavioral intentions. The results indicated that intention prediction can be influenced by or inferred from various factors, such as contextual cues and observed actions. Therefore, it is valuable to explore alternative human-AI collaboration mechanisms for predicting pedestrian's intentions. For example, rather than directly predicting pedestrian behavioral intention themselves, humans may be better at annotating contextual cues to inform AI models to make the prediction. Collecting contextual information can also help new models to be developed to proactively predict pedestrian intentions. This study considered the contextual cues as a sampling criteria but future research designs can facilitate acquisition of various levels of contextual information to enhance AI's perception.

Continuous labels and measures in intent prediction. The discrepancy between the observational trends and the Chi-Square statistical findings suggests that binary coding (correct/incorrect) might not capture the nuances of prediction performance. Future task designs could benefit from employing more granular or continuous measures of eliciting intention labels, such as soft labels aggregated from the crowd [3, 8], to capture more nuanced understanding of human prediction.

6 CONCLUSION

In this work, we explored the concept of collective theory of mind by investigating four factors affecting the crowd capability in estimating pedestrian behavioral intentions. Key findings include that the status quo bias affects the collective performance in intention prediction tasks, which can be mitigated by simple question rephrasing methods. Prediction horizon affects the intention estimation capability as well; and collectively, human crowd can estimate pedestrian intention more accurately than trained AI algorithms in longer duration. Observed pedestrian actions and contextual cues are not found to have statistically significant effects. However, the observational trends in the data suggests that certain type of pedestrian action and contextual cues affect the crowd's performance, highlighting the need for future research with a larger sample size to explore the potential interactions among these factors. Overall, our findings emphasize on the feasibility and necessity of employing collective theory of mind to help establish the threshold of prediction error which can dictate the degree of accuracy for smooth and proactive human-machine interaction.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No.2145565. The work is also supported by IUPUI Research Support Funds Grant (RSFG).

REFERENCES

[1] Ian A Apperly and Stephen A Butterfill. 2009. Do humans have two systems to track beliefs and belief-like states? *Psychological review* 116, 4 (2009), 953.

[2] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1100–1105. https://doi.org/10.1145/3308560.3317083

[3] Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36, 1 (2015), 15–24.

[4] Jan Auernhammer. 2020. Human-centered AI: The role of Human-centered Design Research in the development of AI. In *Proceedings of DRS (DRS2020)*. Design Research Society, online, 1315–1333. https://doi.org/10.21606/drs.2020.282

[5] Alexandra M Boggs, Behram Wali, and Asad J Khattak. 2020. Exploratory analysis of automated vehicle crashes in California: A text analytics & hierarchical Bayesian heterogeneity-based approach. *Accident Analysis & Prevention* 135 (2020), 105354.

[6] Eamon T Campolettano, John M Scanlon, and Trent Victor. 2023. Representative Pedestrian Collision Injury Risk Distributions for A Dense-Urban US ODD Using

Naturalistic Dash Camera Data. In *27th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration*. NHTSA, Yokohama, Japan, 10.

[7] Tina Chen and Renran Tian. 2021. A Survey on Deep-Learning Methods for Pedestrian Behavior Prediction from the Egocentric View. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, Indianapolis, IN, USA, 1898–1905. https://doi.org/10.1109/ITSC48978.2021.9565041

[8] 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.

[9] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics* 10 (2022), 92–110.

[10] Giuseppe Di Pellegrino, Luciano Fadiga, Leonardo Fogassi, Vittorio Gallese, and Giacomo Rizzolatti. 1992. Understanding motor events: a neurophysiological study. *Experimental brain research* 91 (1992), 176–180.

[11] Joshua E Domeyer, John D Lee, and Heishiro Toyoda. 2020. Vehicle automation–other road user communication and coordination: Theory and mechanisms. *IEEE Access* 8 (2020), 19860–19872.

[12] Frank Ole Flemisch, Klaus Bengler, Heiner Bubb, Hermann Winner, and Ralph Bruder. 2014. Towards cooperative guidance and control of highly automated vehicles: H-Mode and Conduct-by-Wire. *Ergonomics* 57, 3 (2014), 343–360.

[13] Olivier Friard and Marco Gamba. 2016. BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in ecology and evolution* 7, 11 (2016), 1325–1330.

[14] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Jiechao Xiong, Shaogang Gong, Yizhou Wang, and Yuan Yao. 2016. Robust Subjective Visual Property Prediction from Crowdsourced Pairwise Labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 3 (2016), 563–577. https://doi.org/10.1109/TPAMI.2015.2456887

[15] Berthold Färber. 2016. Communication and Communication Problems Between Autonomous Vehicles and Human Drivers. In *Autonomous Driving*, Markus Maurer, J. Christian Gerdes, Barbara Lenz, and Hermann Winner (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 125–144. https://doi.org/10.1007/978-3-662-48847-8_7

[16] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2017. Clarity is a Worthwhile Quality: On the Role of Task Clarity in Microtask Crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media* (Prague, Czech Republic) *(HT '17)*. Association for Computing Machinery, New York, NY, USA, 5–14. https://doi.org/10.1145/3078714.3078715

[17] Harshayu Girase, Haiming Gang, Srikanth Malla, Jiachen Li, Akira Kanehara, Karttikeya Mangalam, and Chiho Choi. 2021. LOKI: Long Term and Key Intentions for Trajectory Prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Montreal, QC, Canada, 9783–9792. https://doi.org/10.1109/ICCV48922.2021.00966

[18] Mahsa Golchoubian, Moojan Ghafurian, Kerstin Dautenhahn, and Nasser Lashgarian Azad. 2023. Pedestrian Trajectory Prediction in Pedestrian-Vehicle Mixed Environments: A Systematic Review. *IEEE Transactions on Intelligent Transportation Systems* 24, 11 (Nov. 2023), 11544–11567. https://doi.org/10.1109/TITS.2023.3291196

[19] Robert M Gordon. 1986. Folk psychology as simulation. *Mind & language* 1, 2 (1986), 158–171.

[20] Giorgio Grasso, Chiara Lucifora, Pietro Perconti, and Alessio Plebe. 2019. Evaluating Mentalization during Driving. In *Proceedings of the 5th International Conference on Vehicle Technology and Intelligent Transport Systems - Volume 1: VEHITS,*. INSTICC, SciTePress, Heraklion, Greece, 536–541. https://doi.org/10.5220/0007756505360541

[21] Surabhi Gupta, Maria Vasardani, and Stephan Winter. 2019. Negotiation Between Vehicles and Pedestrians for the Right of Way at Intersections. *IEEE Transactions on Intelligent Transportation Systems* 20, 3 (March 2019), 888–899. https://doi.org/10.1109/TITS.2018.2836957

[22] Giannis Haralabopoulos, Myron Tsikandilakis, Mercedes Torres Torres, and Derek McAuley. 2020. Objective Assessment of Subjective Tasks in Crowdsourcing Applications. In *Proceedings of the LREC 2020 Workshop on "Citizen Linguistics in Language Resource Development"*, James Fiumara, Christopher Cieri, Mark Liberman, and Chris Callison-Burch (Eds.). European Language Resources Association, Marseille, France, 15–25. https://aclanthology.org/2020.cllrd-1.3

[23] Danula Hettiachchi, Indigo Holcombe-James, Stephanie Livingstone, Anjalee De Silva, Matthew Lease, Flora D. Salim, and Mark Sanderson. 2023. How Crowd Worker Factors Influence Subjective Annotations: A Study of Tagging Misogynistic Hate Speech in Tweets. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 11, 1 (Nov. 2023), 38–50. https://doi.org/10.1609/hcomp.v11i1.27546

[24] Taotao Jing, Haifeng Xia, Renran Tian, Haoran Ding, Xiao Luo, Joshua Domeyer, Rini Sherony, and Zhengming Ding. 2022. InAction: Interpretable Action Decision Making for Autonomous Driving. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Vol. 13698. Springer Nature Switzerland, Cham, 370–387. https://doi.org/10.1007/978-3-031-19839-7_22

[25] Raquel Justo, M. Inés Torres, and José M. Alcaide. 2017. Measuring the Quality of Annotations for a Subjective Crowdsourcing Task. In *Pattern Recognition and Image Analysis*, Luís A. Alexandre, José Salvador Sánchez, and João M. F. Rodrigues (Eds.). Springer International Publishing, Cham, 58–68.

[26] Christina Katsimerou, Joris Albeda, Alina Huldtgren, Ingrid Heynderickx, and Judith A. Redi. 2016. Crowdsourcing Empathetic Intelligence: The Case of the Annotation of EMMA Database for Emotion and Mood Recognition. *ACM Trans. Intell. Syst. Technol.* 7, 4, Article 51 (may 2016), 27 pages. https://doi.org/10.1145/2897369

[27] Iuliia Kotseruba, Amir Rasouli, and John K. Tsotsos. 2021. Benchmark for Evaluating Pedestrian Action Prediction. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Waikoloa, HI, USA, 1257–1267. https://doi.org/10.1109/WACV48630.2021.00130

[28] Christopher A Kurby and Jeffrey M Zacks. 2008. Segmentation in the perception and memory of events. *Trends in cognitive sciences* 12, 2 (2008), 72–79.

[29] Tianyi Li, Kurt Luther, and Chris North. 2018. CrowdIA: Solving Mysteries with Crowdsourced Sensemaking. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 105 (nov 2018), 29 pages. https://doi.org/10.1145/3274374

[30] Bingbin Liu, Ehsan Adeli, Zhangjie Cao, Kuan-Hui Lee, Abhijeet Shenoi, Adrien Gaidon, and Juan Carlos Niebles. 2020. Spatiotemporal Relationship Reasoning for Pedestrian Intent Prediction. *IEEE Robotics and Automation Letters* 5, 2 (2020), 3485–3492. https://doi.org/10.1109/LRA.2020.2976305

[31] Louise McHugh and Ian Stewart (Eds.). 2012. *The self and perspective taking: contributions and applications from modern behavioral science*. New Harbinger Publications, Oakland, CA.

[32] Benedikt Morschheuser, Juho Hamari, and Jonna Koivisto. 2016. Gamification in Crowdsourcing: A Review. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*. IEEE, Koloa, HI, USA, 4375–4384. https://doi.org/10.1109/HICSS.2016.543

[33] Hasti Narimanzadeh, Arash Badie-Modiri, Iuliia Smirnova, and Ted Hsuan Yun Chen. 2023. Crowdsourcing subjective annotations using pairwise comparisons reduces bias and error compared to the majority-vote method. arXiv:2305.20042 [cs.HC]

[34] Thomas E Nelson, Zoe M Oxley, and Rosalee A Clawson. 1997. Toward a psychology of framing effects. *Political behavior* 19 (1997), 221–246.

[35] National Transportation Safety Board 2019. *Collision between vehicle controlled by developmental automated driving system and pedestrian*. National Transportation Safety Board. Retrieved March 16, 2024 from https://www.ntsb.gov/investigations/accidentreports/reports/har1903.pdf

[36] Cynthia Owsley, Beth Stalvey, Jennifer Wells, and Michael E Sloane. 1999. Older drivers and cataract: driving habits and crash risk. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* 54, 4 (1999), M203–M211.

[37] Maja Pantic, Alex Pentland, Anton Nijholt, and Thomas S. Huang. 2007. Human Computing and Machine Understanding of Human Behavior: A Survey. In *Artifical Intelligence for Human Computing*, Thomas S. Huang, Anton Nijholt, Maja Pantic, and Alex Pentland (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 47–71.

[38] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526.

[39] Raul Quintero Minguez, Ignacio Parra Alonso, David Fernandez-Llorca, and Miguel Angel Sotelo. 2019. Pedestrian Path, Pose, and Intention Prediction Through Gaussian Process Dynamical Models and Pedestrian Activity Recognition. *IEEE Transactions on Intelligent Transportation Systems* 20, 5 (May 2019), 1803–1814. https://doi.org/10.1109/TITS.2018.2836305

[40] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John Tsotsos. 2019. PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Seoul, Korea (South), 6261–6270. https://doi.org/10.1109/ICCV.2019.00636

[41] Amir Rasouli, Iuliia Kotseruba, and John K. Tsotsos. 2017. Agreeing to cross: How drivers and pedestrians communicate. In *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, Los Angeles, CA, USA, 264–269. https://doi.org/10.1109/IVS.2017.7995730

[42] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. 2018. Understanding pedestrian behavior in complex traffic scenes. *IEEE Transactions on Intelligent Vehicles* 3, 1 (2018), 61–70.

[43] William Samuelson and Richard Zeckhauser. 1988. Status quo bias in decision making. *Journal of risk and uncertainty* 1 (1988), 7–59.

[44] Friederike Schneemann and Patrick Heinemann. 2016. Context-based detection of pedestrian crossing intention for autonomous driving in urban environments. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Daejeon, South Korea, 2243–2248. https://doi.org/10.1109/IROS.2016.7759351

[45] Garriy Shteynberg, Jacob B. Hirsh, Wouter Wolf, John A. Bargh, Erica J. Boothby, Andrew M. Colman, Gerald Echterhoff, and Maya Rossignac-Milon. 2023. Theory of collective mind. *Trends in Cognitive Sciences* 27, 11 (Nov. 2023), 1019–1031.

https://doi.org/10.1016/j.tics.2023.06.009

[46] Lucille Alice Suchman. 1987. *Plans and situated actions: the problem of human-machine communication.* Cambridge University Press, Cambridge [Cambridgeshire] ; New York.

[47] Renran Tian, Lingxi Li, Kai Yang, Stanley Chien, Yaobin Chen, and Rini Sherony. 2014. Estimation of the vehicle-pedestrian encounter/conflict risk on the road based on TASI 110-car naturalistic driving data collection. In *2014 IEEE Intelligent Vehicles Symposium Proceedings*. IEEE, MI, USA, 623–629. https://doi.org/10.1109/IVS.2014.6856599

[48] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. https://doi.org/10.1126/science.185.4157.1124 arXiv:https://www.science.org/doi/pdf/10.1126/science.185.4157.1124

[49] Peter Washington, Haik Kalantarian, Jack Kent, Arman Husic, Aaron Kline, Emilie Leblanc, Cathy Hou, Cezmi Mutlu, Kaitlyn Dunlap, Yordan Penev, et al. 2021. Training affective computer vision models by crowdsourcing soft-target labels. *Cognitive computation* 13 (2021), 1363–1373.

[50] Henry M. Wellman. 1992. *The child's theory of mind* (1. mit pr. paperback ed ed.). MIT Press, Cambridge, Mass. u.a.

[51] Wei Xu. 2019. Toward human-centered AI: a perspective from human-computer interaction. *interactions* 26, 4 (2019), 42–46.

[52] Jeffrey M Zacks and Khena M Swallow. 2007. Event segmentation. *Current directions in psychological science* 16, 2 (2007), 80–84.

[53] Chi Zhang and Christian Berger. 2023. Pedestrian Behavior Prediction Using Deep Learning Methods for Urban Scenarios: A Review. *IEEE Transactions on Intelligent Transportation Systems* 24, 10 (Oct. 2023), 10279–10301. https://doi.org/10.1109/TITS.2023.3281393

[54] Zhengming Zhang, Renran Tian, and Zhengming Ding. 2023. TrEP: Transformer-Based Evidential Prediction for Pedestrian Intention with Uncertainty. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 3 (June 2023), 3534–3542. https://doi.org/10.1609/aaai.v37i3.25463