Using Models Based on Cognitive Theory to Predict Human Behavior in Traffic: A Case Study

Julian F. Schumann
Cognitive Robotics
TU Delft
Delft, Netherlands
j.f.schumann@tudelft.nl

Aravinda R. Srinivasan
Transport Studies Institute
Leeds University
Leeds, United Kingdom
a.r.srinivasan@leeds.ac.uk

Jens Kober
Cognitive Robotics
TU Delft
Delft, Netherlands
j.kober@tudelft.nl

Gustav Markkula
Transport Studies Institute
Leeds University
Leeds, United Kingdom
g.markkula@leeds.ac.uk

Arkady Zgonnikov
Cognitive Robotics
TU Delft
Delft, Netherlands
a.zgonnikov@tudelft.nl

Abstract—The development of automated vehicles has the potential to revolutionize transportation, but they are currently unable to ensure a safe and time-efficient driving style. Reliable models predicting human behavior are essential for overcoming this issue. While data-driven models are commonly used to this end, they can be vulnerable in safety-critical edge cases. This has led to an interest in models incorporating cognitive theory, but as such models are commonly developed for explanatory purposes, this approach's effectiveness in behavior prediction has remained largely untested so far. In this article, we investigate the usefulness of the *Commotions* model – a novel cognitively plausible model incorporating the latest theories of human perception, decision-making, and motor control – for predicting human behavior in gap acceptance scenarios, which entail many important traffic interactions such as lane changes and intersections. We show that this model can compete with or even outperform well-established data-driven prediction models across several naturalistic datasets. These results demonstrate the promise of incorporating cognitive theory in behavior prediction models for automated vehicles.

Index Terms—autonomous vehicles, gap acceptance, behavior prediction, cognitive theory.

I. Introduction

Automated vehicles have become a major focus of the car industry in recent years due to their potential to revolutionize transportation. The promised benefits of automated vehicles include fewer accidents caused by human errors, increased accessibility of mobility solutions, and more efficient use of time while traveling [1]–[3]. However, despite significant investments [4], there are still only prototypes of automated vehicles on the street, and they are not yet widely available to the public [5], [6]. One major challenge to the widespread adoption of automated vehicles is ensuring that they are both efficient and safe, traveling in a timely and efficient manner while also maintaining a level of safety that is at least equivalent to human driving [5], [7]. However, many automated vehicles currently focus on ensuring safety, avoiding any action that could potentially lead to an accident. While this approach may reduce the risk of traffic participants being harmed, it misses out on travel efficiency and acceptance, requiring further efforts to make automated vehicles truly useful [5],

[6]. One potential solution is to incorporate prediction models to reduce uncertainty about future human behavior and allow for more actions to be classified as safe [8], [9].

Accurate predictions of human behavior are especially critical in scenarios involving gap acceptance [10], which form a significant subset of space-sharing conflicts in traffic, including situations such as crossing an intersection or changing lanes [11]. Many models for predicting human behavior in these scenarios have been developed, including trajectory prediction models [12], [13] and models predicting the binary choice of either accepting or rejecting the gap [14], [15]. However, most of these include few assumptions about human decision-making – using a mainly data driven approach known for being unreliable in safety-critical edge cases [10], [16].

Meanwhile, there is a separate literature of cognitive theory developed to explain human decision-making in traffic [17], [18]. Inclusion of such theory into predictive models might help overcome the unreliability issues of purely data-driven approaches [16]. However, current cognitively plausible models have a number of limitations which hinder their use for behavior prediction. In particular, most such models are limited to a specific scenario [14], [18] and cannot handle complex inputs which prevents their applications to naturalistic datasets. As a result, it is currently unknown if incorporating cognitive theories in behavior prediction models could actually yield any benefits in terms of prediction accuracy and robustness.

This study aims to explore the potential of one possible approach of incorporating cognitive theory into prediction models: the adaption of a specific existing explanatory model [17] to function as a prediction model, using gap acceptance as target scenario type. Adaptation of this model for prediction purposes is non-trivial, and does in itself represent a significant contribution to the field (Section III). Furthermore, we also conduct an ablation study to find the most promising configurations for the model (Section IV). Finally. we compare the performance of the resulting configurations of this model to state-of-the-art data-driven prediction models (Section V).

II. Background

This section provides a description of the general type of gap acceptance scenarios addressed here, a brief overview of
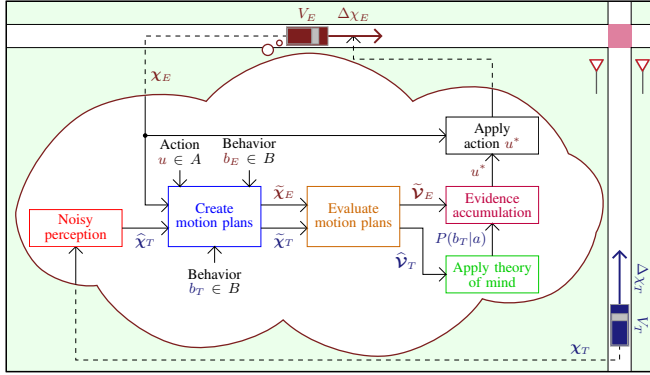
Fig. 1. A depiction of the *Commotions* model [17] and its high-level parts, showing how the position of ego vehicle V_E is updated at one point in time. Simultaneously, the target vehicle V_T also updates its kinematic state by using the same mechanics – only with mirrored inputs.

the tested cognitive model, and an introduction to a framework that facilitates unbiased comparisons of the model's predictive performance against existing benchmarks. We also discuss our changes of the tested model that enable its use as a predictive model.

A. Gap acceptance

Gap acceptance problems are a type of traffic interaction that involves a space-sharing conflict between two agents with intersecting paths, such as intersections, pedestrian crossings, and lane changes on highways [10], [11]. There, these two agents can be differentiated by the possession of the right of way, with the vehicle with priority being referred to as the ego vehicle V_E. In such a situation, the other agent, designated as the target vehicle V_T, must then decide whether to cross V_E's path in front of V_E (i.e., accepting the offered gap) or to wait until V_E has passed, thereby rejecting the gap. For example, if V_T approaches an intersection via a secondary road, it needs to decide whether the gap to the vehicle coming from the perpendicular direction is large enough to cross the intersection without waiting for that car to pass (Fig. 1). Accurately predicting V_T's decision in such scenarios is crucial for V_E, as V_T's future behavior could limit V_E's options, such as V_E being forced to slow down to prevent a collision by V_T accepting the gap.

B. The Commotions model

Markkula et al. [17] proposed a cognitive framework for modeling road user interactions in gap acceptance scenarios. Their framework includes a wide range of cognitive mechanisms, such as decision-making based on evidence accumulation [14], noisy perception [19] and applying a theory of mind [20] (Fig. 1). They implemented this framework in models for interactions between vehicles and/or pedestrians on straight crossing paths, i.e., including gap acceptance scenarios between two vehicles. What we will refer to here as the *Commotions* model (after the name of the project in which the model was developed) is the most successful model variant identified in [17], applied to such scenarios.

As illustrated in Fig. 1, the proposed model postulates that at each time step, both ego vehicle V_E and target vehicle V_T concurrently determine their current control inputs. This decision-making process of each agent is subject to sensory noise and Bayesian filtering during the perception of the position of the other agent. Based on their own short-term control input u (A is a discrete set) and both vehicles' long-term behavior b_E and b_T (i.e., preference for going first or second through the contested space), corresponding pairs of future trajectories – represented by pairs of $\widetilde{\chi}_E$ and $\widetilde{\chi}_T$ – are generated, with the constraint that the resulting interactions are safe. Each pair of trajectories is then evaluated (punishing large control inputs, time delays, and traffic rule violations), resulting in the value $\widetilde{\mathcal{V}}_E$ representing the agent's own opinion and the value $\widehat{\mathcal{V}}_T$, which is the value the agent assumes that the other agent assigns to each trajectory pair for each possible combination of behaviors and control inputs. Each agent then weighs the evaluation $\widetilde{\mathcal{V}}_E$ of their own trajectory based on the probability of the other vehicle behaving accordingly, assuming per the theory of mind that this probability is correlated with the respective value $\widehat{\mathcal{V}}_T$. Evidence accumulation is used to ensure no abrupt and seemingly arbitrary changes in behavior, by combining the weighted values with previous evaluations of a potential action u and only changing the applied control input u^* if this entails a sufficiently substantial improvement in this accumulated value, i.e., control is intermittent. Based on the currently chosen control input u^* each agent's states are then projected forward to the next time step. By repeatedly using this process for both agents, the model can generate a pair of simulated trajectories on the perpendicular intersection. To represent the models randomness, n_p different trajectory pairs are generated in repeated simulations.

C. The framework for benchmarking gap acceptance models

To compare several prediction models in a fair and unbiased manner, we utilize a framework previously developed by Schumann et al. [10]. This framework facilitates the comparison of such models in any gap acceptance scenario according to a wide selection of metrics. Moreover, it grants precise control over the timing of the evaluated predictions and the allocation of individual samples to training and testing sets.

The framework also permits the conversion of different types of predictions, including between binary and trajectory predictions, increasing the number of metrics that can be employed to compare models. For instance, the benchmark enables models that originally predict only gap acceptance probabilities to also generate predictions of full trajectories. Specifically, to transform a predicted probability $a_{\text{pred}} \in [0, 1]$ of accepting the gap into a set of predicted trajectories for a given sample, the framework uses two instances of a state-of-the-art trajectory prediction model [12]. One of these models is trained exclusively on samples with accepted gaps, while the other is trained on samples with rejected gaps. Both models are utilized to predict a set of trajectories based on

the given sample's input, from which the final set is sampled with weights adjusted by a_{pred} [10].

III. COMMOTIONS AS A PREDICTIVE MODEL

Although the *Commotions* model's capability of expressing a number of empirically observed human interaction phenomena was demonstrated successfully in the original paper [17], it was not developed for use as a prediction model. As such, it has many limitations compared to existing models developed for this purpose. For once, the computational efficiency of its existing implementation makes training and testing on most datasets infeasible. In this paper, we address this problem by implementing parallel processing of multiple model predictions on a GPU and using analytical instead of numerical integration inside the model. This achieves a speed increase of roughly four orders of magnitude.

Another problem with the *Commotions* model is that it is constrained to the scenario of perpendicular intersections with straight trajectories seen in Fig. 1, which is incongruent to most real world situations. We utilize an expansion of the benchmarking framework II-C allowing us to project real-world two-dimensional trajectories onto the quasi-one-dimensional-scenario required as the input data. Namely, for each agent, we define a method for determining the most probable path from their current location towards the contested space, where the trajectories of the ego vehicle V_E and the target vehicle V_T intersect. The length of this path is then assumed to be equal to the distances of those agents to the contested space (the purple square in Fig. 1) along the respective perpendicular streets.

While it might be possible to use the same approach to project predicted trajectories from the quasi-one-dimensional scenario to the original two-dimensional space, they could solely be projected onto the aforementioned predefined most probable paths. As this would drastically limit the solution space, we instead use scenario-independent information from the predicted trajectories. For each pair of trajectories from simulation p we can determine if the contested space was reached first by V_T ($a_{\text{pred},p} = 1$ represents an accepted gap) or V_E ($a_{\text{pred},p} = 0$). Simultaneously, the time $t_{A,\text{pred},p}$ of V_T reaching the contested space can be extracted as well. Averaging over all predictions allows us to calculate the probability $a_{\text{pred}} \in [0,1]$ of V_T accepting the gap. Combined with the predicted time of acceptance t_A, which the framework accepts as another type of prediction [10], generating predicted trajectories in the original space then becomes possible (II-C).

Finally, the *Commotions* model is able to process merely the current position and velocity of only the two principal actors in a gap acceptance scenario as *initial model input*, i.e., V_E and V_T, and not any other agents in the scene. However, as this only hinders but does not prevent the model's predictive usage, this issue remains currently unaddressed.

IV. EVALUATING COMMOTIONS' CONFIGURATIONS

In this section, we investigate the predictive performance of several configurations of the *Commotions* model stemming from a number of design decisions that have to be made when using the *commotions* model to predict human behavior. For example, the modeling of the interaction between V_E and V_T can utilize either an *interactive* approach, where both agents utilize all aspects of the *Commotions* model (Fig. 1) to determine their current control inputs u^* (*IM*), or a *non-interactive* approach (*NM*), where only the behavior of V_T is predicted by the model, with V_E set to maintaining its original velocity. Meanwhile, another decision pertains to selecting the form of short-term control inputs u, with the options being the application of either a *constant acceleration* (*AC*) or *constant jerk* (*JC*).

As important parts of the model such as the creation of the trajectories χ_E and χ_T (Fig. 1 and II-B) are non-differentiable, we use Bayesian optimization [21] to fit the *Commotions* model's parameters. However, regarding the optimization procedure, some open questions still remain. First, the user must decide whether to train the model in a *single optimization* round (*1O*) or use a *two-stage optimization* (*2O*), wherein the second stage of optimization is carried out over a reduced parameter search space surrounding the optimized parameters obtained in the first stage. Second, a choice between the two available *loss functions* $\mathcal{L}_1$ and $\mathcal{L}_2$ used to fit the *Commotions* model's parameters must be made. $\mathcal{L}_1$ is adapted directly from the work of Zgonnikov et al. [18] (with t_C being the time when V_E reaches the purple intersection in Fig. 1) and evaluates every prediction p for each sample i, while $\mathcal{L}_2$ expands upon this by enforcing more varied predictions:

$$
\begin{aligned}
\mathcal{L}_1 &= \sum_i \frac{1}{n_p} \sum_{p=1}^{n_p} 4 \left| a_i - a_{\text{pred},i,p} \right| + \left(t_i - t_{A,\text{pred},i,p} \right)^2 \\
\mathcal{L}_2 &= \mathcal{L}_1 + \sum_i 100 V_i - 20\sqrt{V_i} + 1 \\
t_i &= \min \left\{ t_{A,i}, \max \left\{ t_{C,i}, t_{A,\text{pred},i,p} \right\} \right\} \\
V_i &= \min \left\{ \mathbb{V}_p \left(t_{A,\text{pred},i,p} \right), \frac{1}{100} \right\}
\end{aligned}
\tag{1}
$$

A. Setup

1) Datasets: The predictive performance of the different model configurations is compared using three datasets, each focusing on a different scenario.

- *L-GAP* [18], a driving simulator dataset, contains scenarios in which V_T must decide whether to turn left in front of or behind V_E approaching on the opposite lane.
- *rounD* [22], a real-world dataset captured by a drone, covers roundabouts where V_T must decide whether to enter the roundabout in front of or behind V_E which is already inside the roundabout.
- The *UDISS* dataset [23], created in a driving simulator, focuses on a perpendicular intersection where V_T must cross either in front of or behind V_E, which is driving along the other road with the right of way.

While the latter two datasets include other agents besides V_E and V_T, in this paper we ignore those due to the aforementioned limitations of the *Commotions model*, with the

TABLE I
ASSESSING THE INFLUENCE OF THE BINARY CONFIGURATION CHOICES IN THE *Commotions* MODEL (IV) ON THE PREDICTIVE PERFORMANCE.

Binary choice $\mathcal{C}$	$\mathcal{C}_1$	vs	$\mathcal{C}_2$	Cases	$\mathcal{C}_1$ better than $\mathcal{C}_2$	$\mathcal{C}_2$ better than $\mathcal{C}_1$
Ego vehicle's modeling	Interactive (*IM*)	vs	Non-interactive (*NM*)	88	7% (7%)	2% (22%)
Vehicles' control input	Jerk (*JC*)	vs	Acceleration (*AC*)	88	20% (23%)	26% (25%)
Optimization method	Single round (*1O*)	vs	Two rounds (*2O*)	88	2% (15%)	6% (11%)
Optimized loss function	$\mathcal{L}_1$	vs	$\mathcal{L}_2$	24 (*ADE*) 64 (Other)	79% (29%) 0% (11%)	0% (13%) 52% (30%)

The individual results underlying the values shown here can be found in the form of figures and tables in the supplementary materials. In the two right-most columns, statistical significance of the differences in metrics is tested with a paired Student t-test (significance level $\alpha = 0.05$). The first number in each cell represents percentage of cases on the randomly split testing sets, whereas the number in the parentheses corresponds to the critical split (i.e., the testing set including the most unintuitive samples). In the last row, results are split by metric.

resulting datasets being referred to respectively as *roundD*$_{2V}$ and *UDISS*$_{2V}$. We also restrict the provided input trajectories to two input time steps ($n_I = 2$), as this provides sufficient information to extract the two agents' current positions and velocities, which are the only inputs the *Commotions* model is able to process.

2) Train/test splits: On each dataset, we perform eleven training-and-testing cycles for each configuration. In ten of these, the split between training and testing set is random. In the last split however, we place the 20 % of samples that exhibit the most unintuitive human behavior – respectively smallest accepted gaps and largest rejected gaps – into the *critical* testing set. This latter approach allows us to evaluate the robustness of the model's predictive capabilities against the most challenging and safety-critical cases.

3) Metrics: To evaluate the models' predictions made on the testing set, we employ three metrics which have previously been used to assess different aspects of gap acceptance predictions [10], [12], [15]. First, the area under the receiver-operator curve (*AUC*) assesses binary predictions (accept/reject gap) at two different time points: the initial opening of the gaps and the time corresponding to a fixed (dataset-specific) characteristic gap size [10]. Second, the average displacement error (*ADE*) metric evaluates full predicted trajectories at the characteristic gap size. Third, we use the true negative rate under perfect recall [10] (*TNR-PR*), a metric that rates the usefulness of binary predictions made on the smallest possible gaps at the last point in time when they can aid in adjusting V_E's planned path accordingly. However, due to a lack of gaps accepted after this point in time on the *UDISS* dataset, the *TNR-PR* cannot be calculated on that scenario, resulting in eleven viable combinations of metrics and datasets we can use to compare model configurations.

Furthermore, when we transform binary predictions into trajectory predictions, so that for example *ADE* metric can be applied to the *Commotions* model, we use *Trajectron++* [12], a state-of-the-art trajectory prediction model, in accordance to the method laid out in Section II-C.

B. Results

Following the setup described above, we test 16 configurations of the *Commotions* model (resulting from four independent design choices) on the eleven combinations of datasets and metrics, resulting in 88 comparisons for each design choice on both random and critical split test sets. For example, on the *L-GAP* dataset, the *AUC* averaged over the ten random test sets for predictions made at the fixed-size gap (a size of $3.36\,\text{s}$) ranges from 0.936 to the value 0.970 produced by CM_{NA12}, which utilizes the non-interactive modeling approach (*NM*) and acceleration control (*AC*) and was trained in one round of optimizing (*1O*) $\mathcal{L}_2$.

Comparison between the configurations of the Commotions model (Tab. I) indicates that there was no consistently better alternative for any of the four design choices. Still, we are able to make some recommendations. For example, the non-interactive modeling approach (*NM*) appears to be more likely to outperform its opposite on the critical test set, while having the added advantage of faster evaluations by obviating half of the *Commotions* model's calculations updating χ_E (Fig.1). Similarly, using acceleration control (*AC*) produces better predictions slightly more often, possibly by enabling the model to predict faster human reactions. Although the number of optimization rounds appears to be largely irrelevant, using only one round of optimization (*1O*) may make the model even more robust on the critical test sets, with faster training being another benefit. Comparatively, the most significant factor seems to be the choice of the loss function – as long as one differentiates by metric. Specifically, $\mathcal{L}_1$ is a better choice when minimizing *ADE*, while $\mathcal{L}_2$ is superior on the other three metrics. This is expected, as the regularization achieved by $\mathcal{L}_2$ enforcing some variance in the predictions also leads to a larger spread of predicted trajectories, resulting in a larger average displacement error.

When seeking the best configuration of the *Commotions* model, rather than comparing the binary choices, we can compare the 16 configurations among themselves as well, either by the average result over the ten random test sets or the result on the critical test sets. As model performance mainly depends on the chosen metric, here we discuss *ADE* separately from other metrics. Specifically, we found that the CM_{NA11} configuration is best, having a lower *ADE* in 79% of all the 90 possible comparisons – i.e, on two types of results, three datasets, and against 15 other configurations. Using the same approach on the remaining metrics, we find the most promising configuration to be CM_{NA12} with better
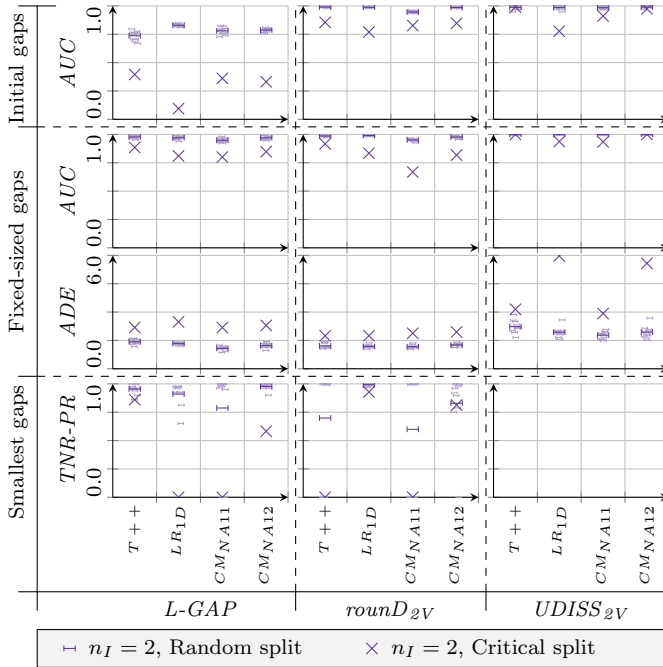
Fig. 2. Behavior prediction performance of the two *Commotions* model (*CM*) configurations compared to *Trajectron++* (*T++*) and logistic regression (*LR*) across three datasets (L-GAP, rounD, and UDISS) according to considered metrics (AUC, ADE, TNR-PR). Large markers indicate the average over repeated simulations.

<div style="text-align:center">

TABLE II

PERCENTAGE OF CASES IN WHICH THE *Commotions* MODEL PERFORMED SIGNIFICANTLY BETTER OR WORSE COMPARED TO THE OTHER TESTED MODELS, BASED ON THE RESULTS SHOWN IN FIG. 2.

</div>

Model M	Cases	CM better than M	CM worse than M
T++	3 (*ADE*)	100% (0%)	0% (0%)
	2 (Other, *UDISS*)	0% (0%)	0% (0%)
	6 (Other, Other)	17% (0%)	33% (33%)
LR_{ID}	3 (*ADE*)	67% (67%)	0% (0%)
	8 (Other)	13% (75%)	25% (0%)

Notation similar to Tab. I. The configuration CM_{NA11} is used for *ADE* and CM_{NA12} for the other metrics. The results are split by metric, and for *T++* partially by dataset.

consistently on the random test sets and, when compared to *LR*, even on the critical test sets. This may be attributed to the model's capacity to forecast both the probability of accepting a gap and the time at which it may be accepted, with the additional information being used to filter out the most aberrant trajectories suggested by the transformation function (II-C).

However, on the other metrics, the *Commotions* model's performance is mostly similar to the other two models (no significant difference on 10/16 random and 8/16 critical splits). Nonetheless, it appears to be more robust than *LR* when predicting unintuitive human behavior, with consistently better outcomes on the critical test. This suggests that constraining a model's predictions using cognitive theory to make it less susceptible to out-of-domain edge cases is a viable way to improve the model's reliability.

The *Commotions* model's worst performance can be observed on the *L-GAP* and *rounD* datasets when compared to *T++* on metrics other than *ADE*. While this might indicate a superiority of the *T++* model, this deviation in performance may be at least partly explained by the aforementioned differences in the inputs provided to the models.

To investigate the extent of potential impact of this difference on our results, we compared the second LR model taking two-dimensional inputs to the original LR model processing the one-dimensional inputs. The results of the comparison (Tab. III) show that, at least for the LR model, processing the two-dimensional original inputs (as *T++* does) appears to simplify the prediction task compared to using the quasi-one-dimensional inputs that the *Commotions* model relies on. This seems plausible, as the projection employed to transform the input data from two-dimensional to one-dimensional likely leads to information loss, leaving fewer cues for the models to make accurate predictions. However, more research is required to accurately assess the impact of input dimensionality on predictions. Thus, a final verdict on the comparative advantage of the *Commotions* model or *T++* is still pending.

V. COMPARING COMMOTIONS TO ESTABLISHED MODELS

In this section, we assess the potential of the *Commotions* model by comparing the predictive performance of two of its configurations (CM_{NA11} for *ADE* and CM_{NA12} for other metrics) against established prediction models. Besides the *Trajectron++* model (*T++*) introduced in Section IV-A3, we also used a logistic regression model (*LR*) as a baseline, with both methods having previously demonstrated good performance on similar gap acceptance problems [10]. While these models have far fewer restrictions on the type of input data they can process, we artificially constrain the used input data to the *Commotions* model's limitations to allow for an equitable comparison. The only exception is the dimensionality of the input for *T++*, as this model can only process the original two-dimensional trajectories, but not the projected quasi-one-dimensional inputs of the *Commotions* model.

Here, the datasets, metrics, as well as testing and training splits are identical to the ablation study (IV-A).

A. Results

Comparison of the models (Fig. 2 and Tab. II) demonstrates that the *Commotions* model can compete with established models, although variations were observed depending on the metric and dataset. Notably, the *Commotions* model routinely outperforms the other models in terms of *ADE*,

metric values in 70% of all cases. These results further support CM_{NA11} and CM_{NA12} (non-interactive modeling, acceleration vehicle input, single-round optimization) as the optimal configurations of the *Commotions* model.

VI. CONCLUSION

This study evaluates the predictive performance of the different configurations of the *Commotions* model, which integrates state-of-the-art theories of human perception, decision-

making, and motor control, in gap acceptance scenarios, comparing the best configurations with other established models. We demonstrate that the *Commotions* model can compete with or even outperform state-of-the-art behavior prediction models, as long as the same input information is provided. Notably, the average displacement error of predicted trajectories is most often significantly lower than the one achieved by other tested models.

We also seek to assess the potential impact of the *Commotions* model's restriction to the quasi-one-dimensional scenario of a perpendicular intersection on its performance. Unable to overcome this restriction, we instead compare two versions of the logistic regression model for this investigation. Our findings suggest that allowing *Commotions* model to instead process two-dimensional trajectories as inputs would be beneficial. As an added benefit, this expansion could also enable the model to function as a dedicated trajectory prediction model. Consequently, such an expansion of the *Commotions* model is likely worlatex set unionthwhile, even if it comes at the cost of more expensive computations. In addition, investigating the impact of other limitations, such as the number of processable input time steps, or the number of interacting road users, should be addressed in future research, as it would provide benefits for model designing even beyond the *Commotions* model.

However, due to its theoretical basis, the *Commotions* model will always be restricted to scenarios such as gap acceptance, where a small number of potential behaviors, like accepting or rejecting a gap, make it feasible to create and evaluate all distinct future trajectories $\widetilde{\chi}$. This limits the model's general applicability compared to models like *Trajectron++*. Additionally, as the model itself is non-differentiable, the resulting need for gradient-free optimization makes the model's training process relatively cumbersome, hampering its feasibility further. Nevertheless, our findings provide encouraging evidence supporting the usefulness of the *Commotions* model, at least for predicting human behavior in gap acceptance scenarios, justifying further research into both this specific model and the general approach of integrating cognitive theory into prediction models. For example, it would be worthwhile to investigate how the cognitive assumptions in the *Commotions* model (or other cognitive models) might be leveraged in model architectures that are specificallylatex set union designed for use in the prediction context.

REFERENCES

[1] J. S. Brar and B. Caulfield, "Impact of autonomous vehicles on pedestrians' safety," in *IEEE Int. Conf. on Intell. Transp. Syst.*

[2] J. Meyer, H. Becker, P. M. Bösch, and K. W. Axhausen, "Autonomous vehicles: The next jump in accessibilities?," *Research Transp. Econ.*, vol. 62, pp. 80–91, June 2017.

[3] J. Pisarov and G. Mester, "The future of autonomous vehicles," *FME Trans.*, vol. 49, no. 1, pp. 29–35, 2021.

[4] D. Holland-Letz, M. Kässer, B. Kloss, and T. Müller, "Mobility's future: An investment reality check | McKinsey," Apr. 2021.

[5] M. Milford, S. Anthony, and W. Scheirer, "Self-Driving Vehicles: Key Technical Challenges and Progress Off the Road," *IEEE Potentials*, vol. 39, pp. 37–45, Jan. 2020.

[6] J. Wang, L. Zhang, Y. Huang, and J. Zhao, "Safety of Autonomous Vehicles," *J. Adv. Transp.*, vol. 2020, p. e8867757, Oct. 2020.

[7] A. Sinha, S. Chand, V. Vu, H. Chen, and V. Dixit, "Crash and disengagement data of autonomous vehicles on public roads in California," *Sci. Data*, vol. 8, p. 298, Dec. 2021.

[8] D. Sadigh, S. Sastry, S. A. Seshia, and A. D. Dragan, "Planning for Autonomous Cars that Leverage Effects on Human Actions," in *Robotics: Sci. Syst. XII*, 2016.

[9] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis, "Deep Learning-Based Vehicle Behavior Prediction for Autonomous Driving Applications: A Review," *IEEE Trans. on Intell. Transp. Syst.*, vol. 23, pp. 33–47, Jan. 2022.

[10] J. F. Schumann, J. Kober, and A. Zgonnikov, "Benchmarking Behavior Prediction Models in Gap Acceptance Scenarios," *IEEE Trans. on Intell. Veh.*, vol. 8, pp. 2580–2591, Mar. 2023.

[11] G. Markkula, R. Madigan, D. Nathanael, E. Portouli, Y. M. Lee, A. Dietrich, J. Billington, A. Schieben, and N. Merat, "Defining interactions: a conceptual framework for understanding interactive behaviour in human and automated road traffic," *Theor. Issues Ergon. Sci.*, vol. 21, Feb. 2020.

[12] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data," in *Comput. Vis. – ECCV 2020* (A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds.), Lecture Notes in Computer Science, pp. 683–700, 2020.

[13] Y. Yuan, X. Weng, Y. Ou, and K. M. Kitani, "AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting," in *IEEE/CVF Int. Conf. on Comput. Vis.*, pp. 9813–9823, 2021.

[14] J. Pekkanen, O. T. Giles, Y. M. Lee, R. Madigan, T. Daimon, N. Merat, and G. Markkula, "Variable-Drift Diffusion Models of Pedestrian Road-Crossing Decisions," *Comput. Brain & Behavior*, vol. 5, pp. 60–80, Mar. 2022.

[15] A. Theofilatos, A. Ziakopoulos, O. Oviedo-Trespalacios, and A. Timmis, "To cross or not to cross? Review and meta-analysis of pedestrian gap acceptance decisions at midblock street crossings," *J. Transport & Health*, vol. 22, p. 101108, Sept. 2021.

[16] G. Marcus, "The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence," Feb. 2020. arXiv: 2002.06177[cs].

[17] G. Markkula, Y.-S. Lin, A. R. Srinivasan, J. Billington, M. Leonetti, A. H. Kalantari, Y. Yang, Y. M. Lee, R. Madigan, and N. Merat, "Explaining human interactions on the road requires large-scale integration of psychological theory," June 2022. PsyArXiv.

[18] A. Zgonnikov, D. Abbink, and G. Markkula, "Should I Stay or Should I Go? Cognitive Modeling of Left-Turn Gap Acceptance Decisions in Human Drivers," *Hum. Factors*, p. 00187208221144561, Dec. 2022. Publisher: SAGE Publications Inc.

[19] O.-S. Kwon, D. Tadin, and D. C. Knill, "Unifying account of visual motion and position perception," *Proc. National Acad. Sci. United States America*, vol. 112, pp. 8142–8147, June 2015.

[20] W. Chen, X. Zhuang, Z. Cui, and G. Ma, "Drivers' recognition of pedestrian road-crossing intentions: Performance and process," *Transp. Research Part F: Traffic Psychol. Behaviour*, vol. 64, pp. 552–564, July 2019.

[21] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient Global Optimization of Expensive Black-Box Functions," *J. Global Optim.*, vol. 13, pp. 455–492, Dec. 1998.

[22] R. Krajewski, T. Moers, J. Bock, L. Vater, and L. Eckstein, "The rounD Dataset: A Drone Dataset of Road User Trajectories at Roundabouts in Germany," in *IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, pp. 1–6, Sept. 2020.

[23] A. R. Srinivasan, J. Schumann, Y. Wang, Y.-S. Lin, M. Daly, A. Solernou, A. Zgonnikov, M. Leonetti, J. Billington, and G. Markkula, "The COMMOTIONS Urban Interactions Driving Simulator Study Dataset," May 2023. arXiv:2305.11909 [cs].

(ITSC), pp. 714–719, Oct. 2017.