Interpretable Trajectory Prediction for Autonomous Vehicles via Counterfactual Responsibility

Kai-Chieh Hsu, Karen Leung, Yuxiao Chen, Jaime F. Fisac, and Marco Pavone

OCTOBER 1 - 5, 2023
IEEE/RSJ International Conference on Intelligent Robots and Systems

Homepage

Trajectory Prediction

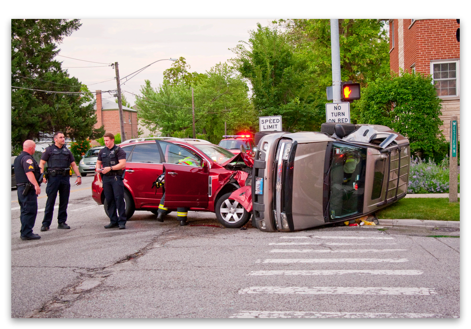
Trajectory prediction is critical for safe and efficient autonomy in human-populated contexts.

Ontological (explicit model)
⚠ poor scalability

Phenomenological (approximation by data)
⚠ not explainable

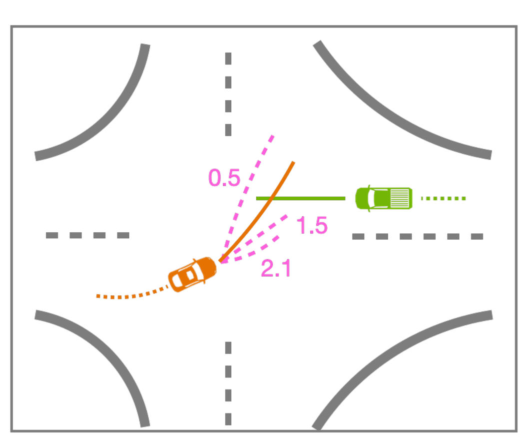
Can we have the best of both worlds by adding an interpretable layer to an underlying predictor?

Counterfactual Responsibility



Responsibility is inherent in human behavior → informative indicator for interpretability

Safety Responsibility β_k^{safety}: ***if*** the query vehicle had acted differently, ***would*** it have been possible to maintain greater separation?

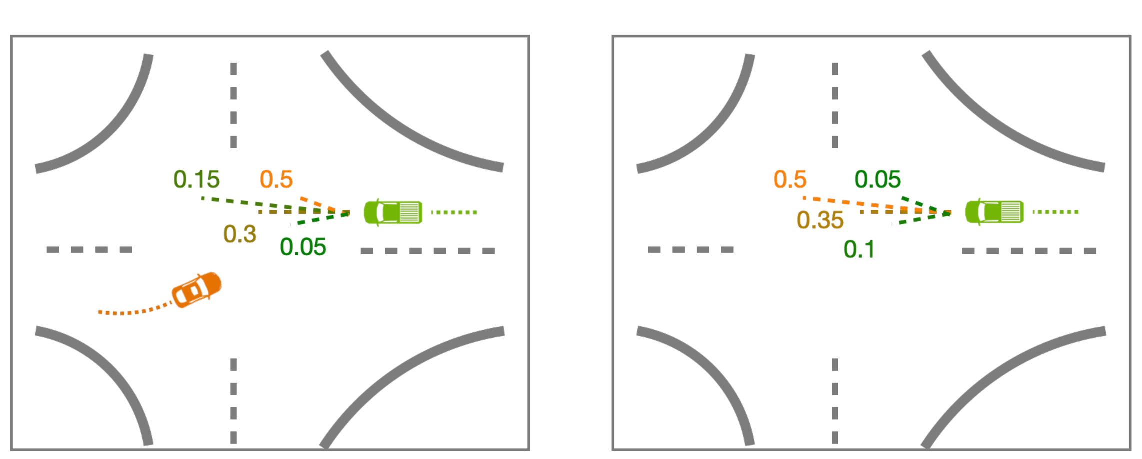


E.g., the closest distance between two trajectories

The queried agent could have maintained larger safety margin (by executing other possible motions) with respect to the target agent.

⚠ irresponsible in the safety metric

Courtesy Responsibility $\beta_k^{\text{courtesy}}$: ***if*** the query vehicle hadn't been there, ***would*** the predicted motions of other agents have changed significantly?



There is a large difference in the distributions of the target agent's motion with/without the queried vehicle's presence.

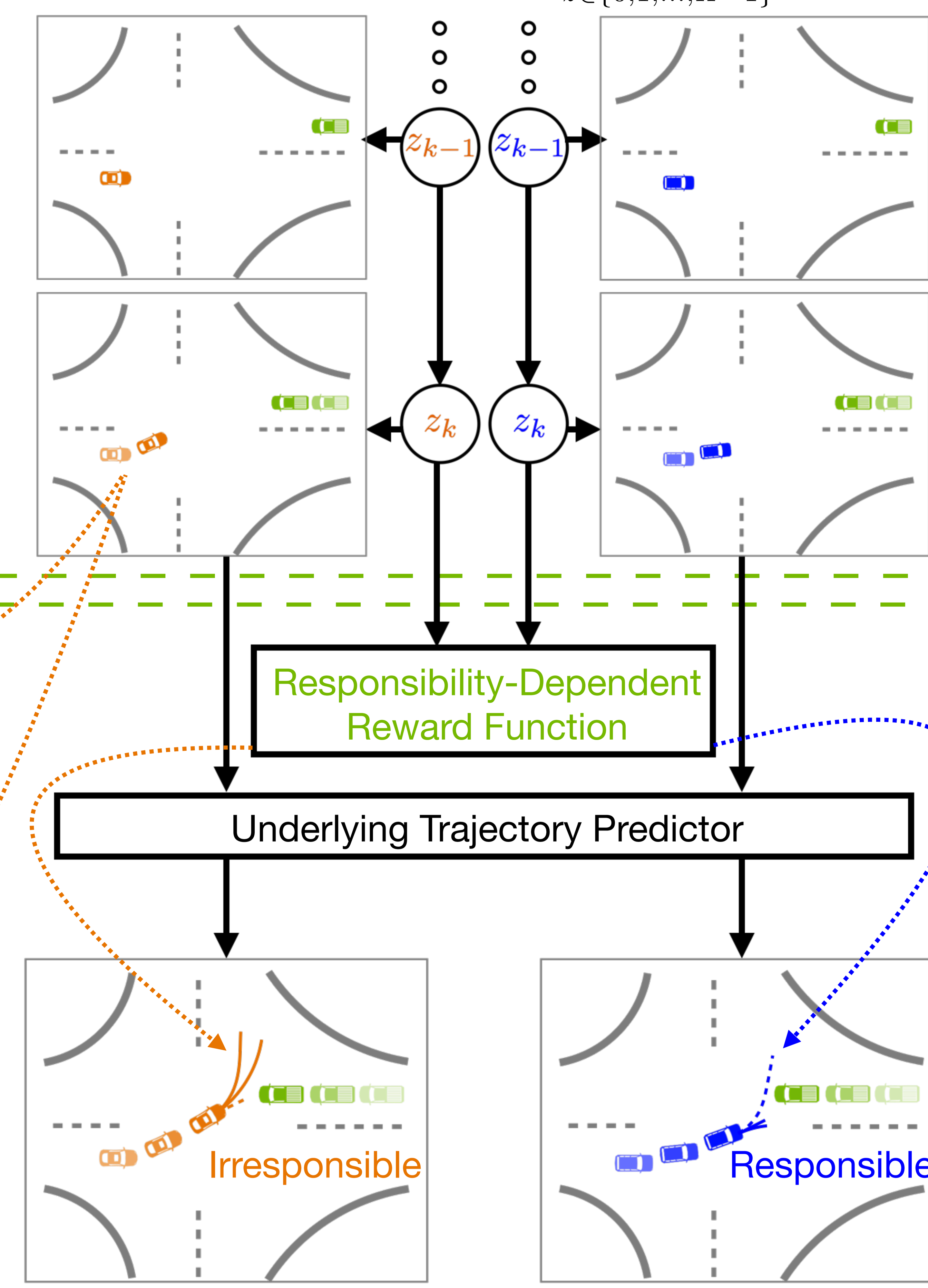
⚠ irresponsible in the courtesy metric

Responsibility-based Interpretability Framework

High-Level: Latent Abstraction Through Responsibility

Simplifies sequences of responsibility metrics into distinct levels by a hidden Markov model to encode long-term dependency.

$$\beta_k := [\beta_k^{\text{safety}}, \beta_k^{\text{courtesy}}] \qquad z_k := \underset{z\in\{0,1,\dots,H-1\}}{\operatorname{argmax}} \; p(z \mid \beta_{1:k})$$



Low-Level: Responsibility-Aware Trajectory Prediction

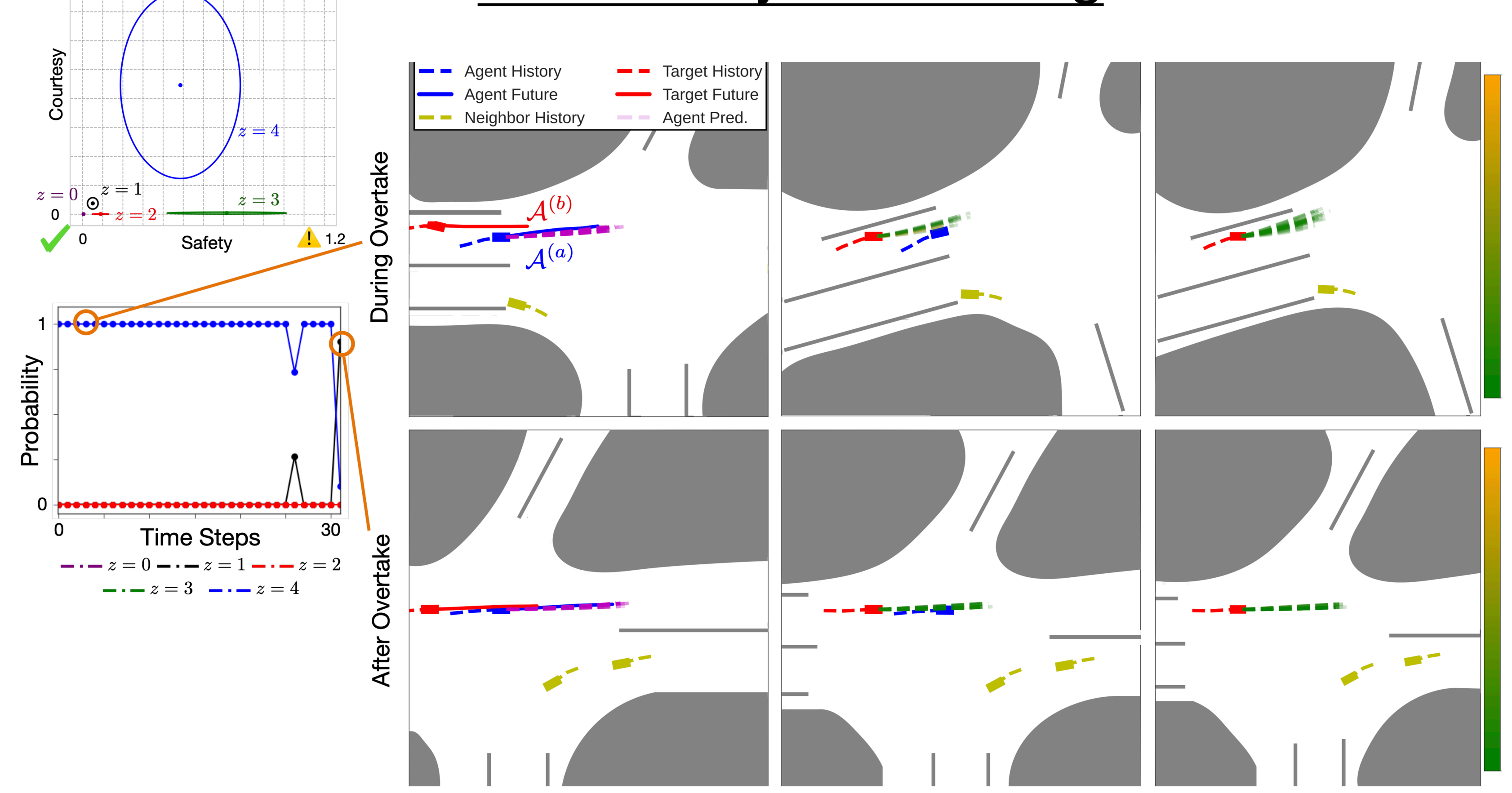
1. Clusters the traffic dataset by responsibility levels z and augments it with artificial trajectories from the underlying trajectory predictor.
2. Learns a reward function R_h for each cluster $\mathcal{D}_h$ to separate the real ($y=1$) and artificial ($y=0$) trajectories by inverse reinforcement learning.

$$\text{Loss}\ (R_h) := -\sum_{(\text{traj},y)\in\mathcal{D}_h} y\log(R_h(\text{traj})) + (1-y)\log(1-R_h(\text{traj}))$$

3. Fine-tunes the underlying trajectory predictor or selects the most likely motion with responsibility-dependent reward functions.

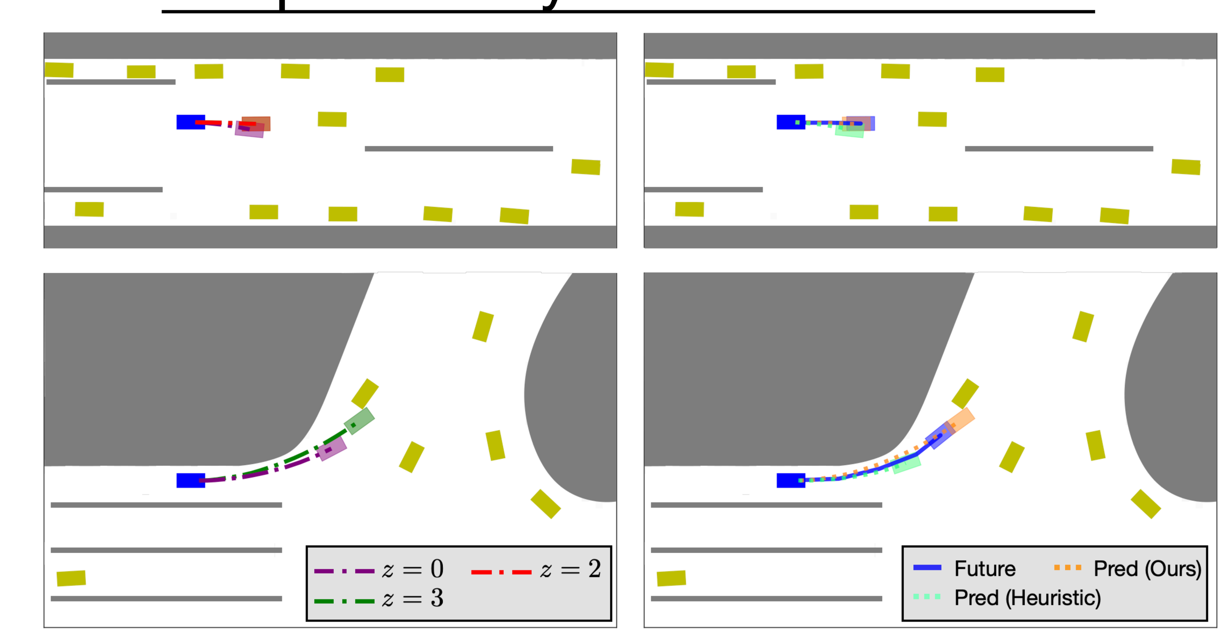
Evaluation with NuScenes Dataset[1]

Case Study: Overtaking



During the overtake, the blue car is aggressive and moves very close to the red car, which leads to safety decrease. Also, the blue car influences the red car's possible motions and is thus irresponsible in courtesy. After overtake, the blue car does not interact with the red car and is classified as responsible.

Responsibility-Aware Prediction



	Train ADE	Train FDE	Val ADE	Val FDE
Responsibility-Aware	0.75 m	0.26 m	1.45 m	0.57 m
Heuristic (BITS[1])	1.02 m	0.39 m	1.64 m	0.65 m

Responsibility-dependent reward functions capture different modes of human behavior. Thus, responsibility aware selection outperforms heuristic reward, which captures human behavior on average only. Additionally, we can generate diverse motions.

Future Work

1. Other types of responsibility: closed-loop, rule-based, information-aware
2. Further integration into AV Software Stack: scenario generation, responsibility-informed planning

References

1. Caesar et al. "nuScenes: A Multimodal Dataset for Autonomous Driving", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
2. Xu et al., "BITS: Bi-level Imitation for Traffic Simulation", IEEE International Conference on Robotics and Automation (ICRA), 2023