Explainable deep learning improves human mental models of self-driving cars

Eoin M. Kenny[1]*, Akshay Dharmavaram[2], Sang Uk Lee[2],
Tung Phan-Minh[2], Shreyas Rajesh[2], Yunqing Hu[2], Laura Major[2],
Momchil S. Tomov[2,3]†, Julie A. Shah[1,4]†

[1]Computer Science & Artificial Intelligence Laboratory (CSAIL),
Massachusetts Institute of Technology, Cambridge, MA, USA.
[2] Motional AD Inc., Boston, MA, USA.
[3]Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA, USA.
[4]Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA, USA.

*Corresponding author(s). E-mail(s): ekenny@mit.edu;
Contributing authors: momchil.tomov@motional.com;
julie_a_shah@csail.mit.edu;
†These authors contributed equally to this work.

Summary

Self-driving cars increasingly rely on deep neural networks to achieve human-like driving [1, 2]. However, the opacity of such black-box motion planners makes it challenging for the human behind the wheel to accurately anticipate when they will fail [3–5], with potentially catastrophic consequences [6–8]. Here, we introduce concept-wrapper network (i.e., CW-Net), a method for explaining the behavior of black-box motion planners by grounding their reasoning in human-interpretable concepts. We deploy CW-Net on a real self-driving car and show that the resulting explanations refine the human driver's mental model of the car, allowing them to better predict its behavior and adjust their own behavior accordingly. Unlike previous work using toy domains or simulations [9–11], our study presents the first real-world demonstration of how to build authentic autonomous vehicles (AVs) that give interpretable, causally faithful explanations for their decisions, without sacrificing performance. We anticipate our method could be applied to other safety-critical systems with a human in the loop, such as autonomous drones and robotic surgeons. Overall, our study suggests a pathway to explainability for autonomous agents as a whole, which can help make them more transparent, their deployment safer, and their usage more ethical.

Keywords: Self-driving cars, Mental model alignment, Interpretable ML, AI safety

1 Introduction

There are hundreds of companies developing autonomous vehicle (AV) technology globally [12], promising to revolutionize transportation for everyone. However, the complexity of fully driverless autonomy has prompted an industry shift towards advanced driver-assistance systems, which require successful communication between the AV and human driver. This is made increasingly difficult by the adoption of deep neural networks in AVs for planning and decision making, the core cognitive functions that determine driving behavior [11]. Deep learning allows motion planners to learn the nuances of human driving behavior from data, but the implicit nature of the learned driving policies makes it challenging to understand the causes of their decisions and to predict their behavior.

A lack of effective communication between the AV and the human driver has contributed to multiple high-profile incidents, some resulting in fatalities [6–8], highlighting the urgent need to make deep motion planners interpretable [11]. Previous studies have sought to address this need using surveys and simulated scenarios [13–20], a human driver emulating the AV [10, 21], or large language models providing post-hoc explanations [2]. However, these studies were theoretical, did not provide faithful explanations of the AV's reasoning process, or were only evaluated in simulation. This leaves open the question of how to provide understandable and useful explanations for the decisions of deep motion planners deployed in real self-driving cars.

To answer this question, we scale up our recent work on interpretable-by-design deep reinforcement learning with prototype-wrapper networks (PW-Nets) [9], using motifs from the literature on concept-bottleneck models [22]. Our key proposal is to ground the reasoning of black-box motion planners in human-interpretable concepts, such as *"Approaching stopped vehicle"* or *"Close to cyclist"*. This method is rooted in case-based reasoning, a classical artificial intelligence (AI) approach [23–25] inspired by cognitive models of human reasoning and memory [26]. It results in causal explanations, such as *"I chose to stop based on recognizing that we are approaching a stopped vehicle."*. In our tests, these explanations help align a safety driver's mental model of the AV with its actual internal decision-making process, increasing transparency and predictability. Importantly, this approach can be applied to arbitrary pre-trained deep neural networks, does not require retraining from scratch, and does not degrade performance of the original black-box planner.

We apply our proposed method, CW-Net (short for concept-wrapper network), to a deep motion planner trained to imitate human driving behavior using inverse reinforcement learning [1]. We replace the final (reward) layer of the pretrained deep neural network with a concept classifier, followed by a new reward layer. We then jointly train the classifier and the new reward layer to predict scenario types and driving decisions, respectively, without modifying the rest of the network. Evaluation on a large-scale benchmark [27] confirms that CW-Net is able to classify concepts without compromising driving behavior. To study the utility of the explanations, we then deploy CW-Net on a real self-driving car [28]. We demonstrate three situations in which the driver has an inaccurate mental model of the motion planner, which is subsequently corrected by the explanations. This ultimately changes the behavior of the driver, for example, by increasing their vigilance in certain situations. Finally, we confirm the statistical significance of these results in an online study (N=120). Overall, our work demonstrates how explainable AI can help users of advanced autonomous systems better understand their behavior in naturalistic settings [29], while also providing insights that can potentially accelerate the development and refinement of such systems.

2 Black-box motion planner

We focus on the motion planning module of the AV stack (Figure 1a), which takes as input a scene context s and outputs a trajectory $\hat{\tau}$. s is a symbolic object-oriented representation of the scene computed by the perception module, while $\hat{\tau}$ is the trajectory that the subsequent controller module should follow. We use a deep neural network architecture consisting of a scene encoder $H(s) \rightarrow \mathbf{h}$ and a trajectory generator $G(s) \rightarrow \{\tau_1 \dots \tau_k\}$, followed by a scene-trajectory encoder $E(\mathbf{h}, \tau_i) \rightarrow \mathbf{z}_i$ and a final reward layer $R(\mathbf{z}_i) \rightarrow r_i$ (Figure 1b). H computes a scene embedding $\mathbf{h}$ and G computes a set of k candidate trajectories $\{\tau_1 \dots \tau_k\}$. Those are combined in E to compute an embedding $\mathbf{z}_i$ for each trajectory τ_i. Finally, R computes an estimated reward r_i for each trajectory τ_i, quantifying how human-like it is. In other words, r_i is higher when τ_i is more similar to how a human would drive in this situation.

During inference, the trajectory with highest reward is selected on each iteration:

$$\hat{\tau} = \tau_{\hat{i}} \quad \text{such that} \quad \hat{i} = \underset{i \in 1,\dots,k}{\arg\max}\, r_i, \quad r_i = R(E(\mathbf{h}, \tau_i))$$

We train the planner on 80 hours of human expert driving using inverse reinforcement learning [1].

3 Planning over human-friendly concepts

To make the planner more interpretable, we replace R with a concept classifier $C(\mathbf{z}_i) \rightarrow \mathbf{c}_i$, followed by a new reward layer $R'(\mathbf{c}_i) \rightarrow r'_i$ (Figure 1c). The concept classifier C computes a logit vector $\mathbf{c}_i$ which is passed through a softmax and/or sigmoid layer which assigns probabilities to different human-interpretable concepts. Since R' computes trajectory rewards from $\mathbf{c}_i$, the final decisions are based solely on these concept assignments and hence they constitute a causally faithful explanation. The rest of the network remains the same.

Similarly to the black-box planner, trajectories are selected according to:

$$\hat{\tau} = \tau_{\hat{i}} \quad \text{such that} \quad \hat{i} = \underset{i \in 1,\dots,k}{\arg\max}\, r'_i, \quad r'_i = R'(C(E(\mathbf{h}, \tau_i)))$$

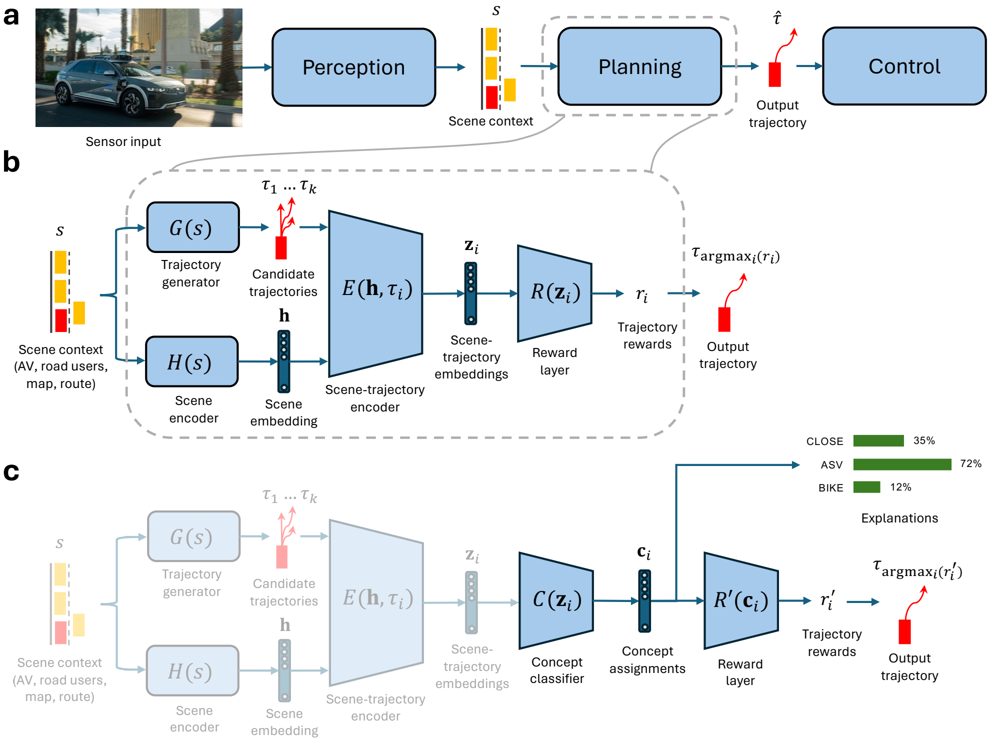


Fig. 1 **Planner architecture. a.** Autonomous vehicle stack. Sensory input is processed by the perception module to generate scene context s. The planning module processes s to compute trajectory $\hat{\tau}$, which is followed by the control module. **b.** Black-box motion planner. s is fed to trajectory generator G, which produces candidate trajectories $\{\tau_1 \dots \tau_k\}$, and scene encoder H, which produces a scene embedding $\mathbf{h}$. These are fed into encoder E, which produces scene-trajectory embeddings $\mathbf{z}_i$ for each τ_i, which are in turn fed into the reward layer R. R computes a reward r_i for each trajectory. The model is trained to output higher rewards for trajectories closer to the the ground-truth human trajectories. **c.** CW-Net. Identical to **b.**, except $\mathbf{z}_i$ is fed to a concept classifier C, which produces concept assignments $\mathbf{c}_i$. These are fed to a new reward layer R' to produce rewards r_i'. In parallel, $\mathbf{c}_i$ is processed to generate the explanations. The reward layer is trained to prefer the same trajectories as the black-box planner, while the concept layer is supervised with ground-truth scenario labels. The weights of the faded components are frozen during training. `CLOSE`, *“Close to another vehicle”*. `ASV`, *“Approaching stopped vehicle”*. `BIKE`, *“Close to cyclist”*.

We train CW-Net to jointly predict concept labels and mimic the driving decisions of the black-box planner. Specifically, $\mathbf{c}_i$ is supervised with multinomial labels corresponding to types of scenarios, such as *“Approaching stopped vehicle”* or *“Close to cyclist”*. This ensures that CW-Net assigns a unique interpretable concept to each unit in $\mathbf{c}_i$. At the same time, r_i' is supervised with trajectories selected by the black-box planner using a cross-entropy loss. During training, the rest of the deep neural network (H, G, and E) is kept frozen (see Section 8.2 for more details).

4 Evaluation in simulation

As a baseline, we first evaluated the black-box planner (without CW-Net) using closed-loop simulations on the nuPlan dataset [27] (Table S1). Overall, the results were competitive with the top submissions to the nuPlan challenge, although performance was slightly lacking when starting from a stop (see Section 8.5). This suggests that there is room for improvement and, importantly, opportunities to study explanations of undesirable behavior.

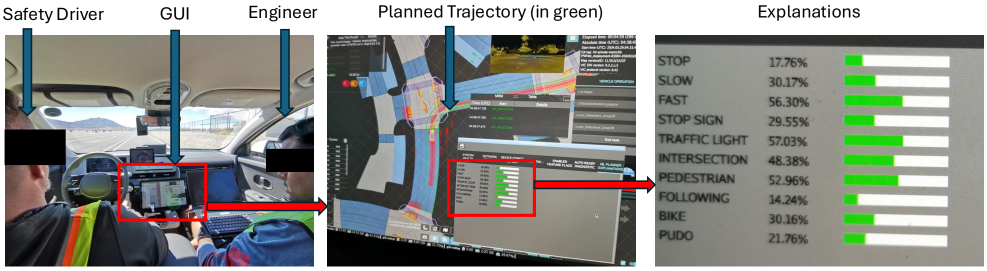
We then evaluated the driving performance of CW-Net wrapped around the black-box planner (Table S1). The results were equivalent, with less than 1% difference across all metrics, confirming our method did not degrade driving performance. We also evaluated concept classification on held-out datasets (Table S2/S3). Mean accuracy was 54%, with 23% precision, 77% recall, and an F1 score of 0.31 (see Section 8.3). Overall, these results indicate that CW-Net can be used to ground the decision making of high-performance deep motion planners in human-interpretable concepts without sacrificing driving performance.

5 Explanations in real-world deployment

To evaluate the usefulness of the explanations in naturalistic settings [29], we deployed CW-Net on a real AV using the Lab2Car wrapper [28]. All tests were performed on a closed course or private lot with an experienced safety driver. Figure 2 shows the experimental setup inside the AV. We next detail three notable situations in which the explanations proved beneficial.

5.1 Unexpected stopping for nearby vehicles

We observed that the AV repeatedly came to a stop shortly before a pedestrian pickup/drop-off zone (Figure 3a). The driver's intuition was that the car stopped because of the pickup/drop-off zone, but the explanations indicated that the planner



STOP	17.76%
SLOW	30.17%
FAST	56.30%
STOP SIGN	29.55%
TRAFFIC LIGHT	57.03%
INTERSECTION	48.38%
PEDESTRIAN	52.96%
FOLLOWING	14.24%
BIKE	30.16%
PUDO	21.76%

Fig. 2 **Deployment setup.** A safety driver, a support engineer, and a researcher were present. The safety driver drove the AV manually between road tests, engaged self-driving mode at the start of each test, monitored AV performance during the test, and took over in case of unsafe driving. The support engineer deployed CW-Net and set scenario destinations. The researcher directed testing. The dashboard included a map with overlaid object detections (s) from the perception module and the output trajectory ($\hat{\tau}$). Explanations $\mathbf{c}_{\hat{i}}$ from CW-Net were shown as percentages for easier interpretation [30].

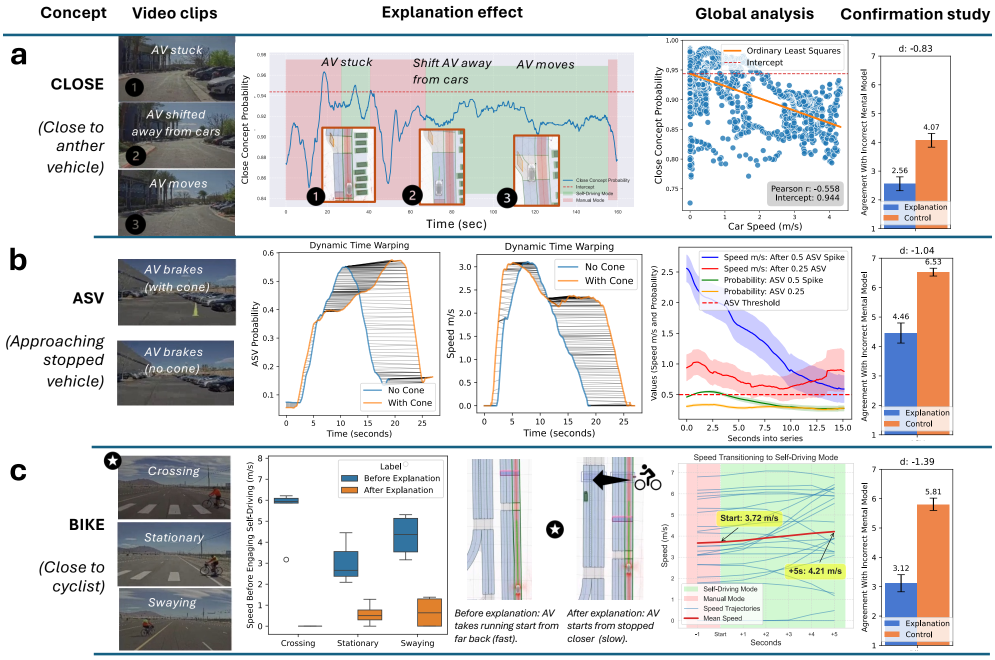


Fig. 3 Results: **a.** The `CLOSE` concept activated when the car got stuck next to parked vehicles. The driver initially thought the pick-up/drop-off area was the cause, but the explanation suggested that it was the nearby vehicles. When the driver engaged self-driving away from the park vehicles, activation of the `CLOSE` concept decreased and the AV started moving again, counter to driver's initial mental model and consistent with the explanations. Across tests, `CLOSE` correlated with speed and the intercept accurately predicted this event. **b.** The `ASV` concept activated when the car stopped next to a traffic cone. The driver initially thought the cone was the cause, but the explanation suggested that the AV was hallucinating a stopped vehicle. When we removed the cone in a counterfactual test, the same phantom braking and concept activation occurred. Across tests, `ASV` correlated with reductions in speed when spiking above 0.5 probability. **c.** The `BIKE` concept failed to activate in our first round of tests with a cyclist, but the car always stopped safely for the cyclist. Responding to the explanation, the driver engaged self-driving from slower speeds in a second round of tests. Follow-up analyses revealed that the AV stopped for the cyclist due to backup safety mechanisms unrelated to CW-Net, which indicates that the driver's increased level of caution was appropriate. These three scenarios validate the driver's updated mental models in response to the explanations. **a-c.** Findings were confirmed in an online study which replicated the events from the road. After seeing the explanations, subjects showed less agreement with the driver's initial mental model, compared to a control group. Standard error of the mean and 3-second rolling averages is shown in relevant plots.

stopped because it detected that it was *"Close to another vehicle"* (the `CLOSE` concept). To test this hypothesis, the driver manually moved the car farther from the parked cars. At this point, the probability of `CLOSE` decreased and the AV began moving again, thus confirming the alternative hypothesis. A full timeline of events is detailed in Figure 3a. We fitted the intercept of the `CLOSE` probability against the speed of the AV globally and found it accurately predicts stopping and starting for this event.[1]

---

[1] This situation used a variant of our architecture detailed in Figure S1.

5.2 Hallucinating a stopped vehicle ahead

At another location, the AV would reliably come to a stop next to a traffic cone (Figure 3b). The driver's initial mental model was that the cone was responsible for the phantom brake. However, the *"Approaching stopped vehicle"* (`ASV`) concept peaked shortly before the car stopped. This suggested an alternative hypothesis, that the planner matched the current situation with training scenarios labeled `ASV`, which in turn promotes stopping behavior associated with such scenarios. As a counterfactual test, the cone was removed. The AV exhibited the same stopping behavior at the same location, along with similar `ASV` probability and speed profiles (L_2 similarities of 1.12 and 1.6 between the respective time-warped profiles, compared to an average $L_2 > 200$ for random events), thus confirming the alternative hypothesis. Note that although there was no vehicle in front of the AV, the explanation is causally faithful to the underlying motion planner and explains why it stopped (namely, because it incorrectly detected a stopped vehicle). Figure 3b illustrates a global analysis of `ASV`, showing it to be a powerful predictor of braking.

5.3 Reacting safely to cyclist

Finally, we tested the ability of the AV to stop safely for cyclists (Figure 3c). For each test, the driver engaged self-driving mode while approaching a cyclist. The driver was instructed to engage self-driving from a speed at which they felt safe, since this determines the subsequent speed of the AV. During the initial tests, the AV reliably stopped for the cyclist. However, the `BIKE` concept maintained a low probability throughout each test ($< 1\%$), indicating that CW-Net was failing to detect the cyclist. Over time, the driver became aware of the concept reading and gradually increased their caution by initiating self-driving from slower speeds. A post-hoc analysis revealed that, although the perception system detected the cyclist, the motion planner failed to take it into account due to a lack of appropriate input features for cyclists. As a result, it chose unsafe trajectories which would have collided with the cyclist. In reality, it was the built-in rule-based systems of the AV that overruled the motion planner and forced the AV to stop. This indicates that the increased caution dictated by the driver's updated mental model was warranted.

6 Confirmation study

We conducted an online study (N=120) to confirm the statistical significance of our results. Following standard procedures in deployment-based research [21, 31], we simulated the sequence of key events from the three situations described previously. We designed a between-subjects study to assess the effect of explanations on the mental model of subjects in the driver's position. For each situation, subjects in the experimental group received the CW-Net explanation, while subjects in the control group received a generic explanation to balance cognitive load (see Figure S4 and Section 8.5 for details). Subjects then rated to what extent they agreed with the driver's incorrect initial mental model. Across all situations, the experimental group gave significantly lower ratings than the control group (Figure 3, right column; mean result: 3.37 ± 1.63 v.

5.46±0.89; t-test $p < 0.001$; Cohen's d=1.58), indicating that the explanations support mental model alignment across the population.

7 Conclusion

Our work shows, for the first time, how explainable deep learning can provide useful explanations for the decisions of self-driving cars in the real world. CW-Net achieves this by grounding the reasoning of a pretrained black-box motion planner in human-interpretable concepts corresponding to types of scenario. By revealing otherwise inaccessible information about the decision-making process of the motion planner in real time, CW-Net helps align the mental model of the human driver with the machine driver. This allows the human driver to better anticipate and account for mistakes of the AV, ultimately resulting in safer driving. Mental model alignment could additionally build trust and understanding with passengers of driverless AVs by helping them anticipate the AV's decisions. This could be particularly beneficial when AV behavior deviates from typical human behavior, such as when driving conservatively or getting stuck. Additionally, the explanations provided by CW-Net can help test engineers provide more precise feedback to the research scientists and engineers working on model improvement. This would be especially relevant for experimental motion planners that are still under development, such as the one used in our study.

CW-Net extends the original PW-Net [9], which reasons over specific scenario *prototypes*, to general scenario *types*. In addition to increasing robustness, reasoning over types has the added benefit of highlighting which parts of the training distribution influence behavior at each point in time. This information can be used by researchers for model improvement. For example, the inability of CW-Net to detect the `BIKE` concept suggests there may not be enough training scenarios with cyclists, or that the focal loss used was ineffective [32], leading to poor performance around cyclists. At the same time, relying on types forgoes some of the benefits of using prototypes. For example, CW-Net can explain that it is stopping because the current scene is similar to `ASV` scenarios in the training data, but it cannot explain *why* it believes so [33, 34]. This suggests a promising avenue of future research: much like humans, who rely on both exemplar-based and rule-based reasoning [35, 36], machines that reason over both prototypes and types could combine the strengths of both approaches, enhancing interpretability while maintaining flexibility in decision making.

Many safety-critical systems involving human-robot interaction require real-time explanations, including AI wingmen, drone navigation systems, and robotic surgeons. Similarly to AVs, many of these applications increasingly rely on deep learning and have a long tail of failure cases, with potentially catastrophic outcomes. This creates an ethical imperative to better understand how they work, so the humans in the loop can intervene when necessary. The success of CW-Net suggests that explainable deep learning may prove essential for meeting the safety and regulatory standards for deploying sophisticated safety-critical agents in the real world.

References

[1] Tung Phan-Minh, Forbes Howington, Ting-Sheng Chu, Momchil S Tomov, Robert E Beaudoin, Sang Uk Lee, Nanxiang Li, Caglayan Dicle, Samuel Findler, Francisco Suarez-Ruiz, et al. Driveirl: Drive in real life with inverse reinforcement learning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1544–1550. IEEE, 2023.

[2] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Video question answering for autonomous driving. *arXiv preprint arXiv:2312.14115*, 2023.

[3] Stefanos Nikolaidis and Julie Shah. Human-robot teaming using shared mental models. *ACM/IEEE HRI*, 2012.

[4] Laura Major and Julie Shah. *What to expect when you're expecting robots: the future of human-robot collaboration.* Hachette UK, 2020.

[5] Rohan Paleja, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, and Matthew Gombolay. The utility of explainable ai in ad hoc human-machine teaming. *Advances in neural information processing systems*, 34:610–623, 2021.

[6] James Titcomb. Uber's safety policies under fire as us watchdog investigates self-driving car death, 11 2019.

[7] Lee Fang. Tesla crash footage shows driver with hands off wheel, raising fresh questions about autopilot safety, 1 2023.

[8] Brad Templeton. Waymo's double crash with pickup trucks and more examined. *Forbes*, 2024. Accessed: 2024-09-22.

[9] Eoin M Kenny, Mycal Tucker, and Julie Shah. Towards interpretable deep reinforcement learning with human-friendly prototypes. In *The Eleventh International Conference on Learning Representations*, 2023.

[10] Gwangbin Kim, Dohyeon Yeo, Taewoo Jo, Daniela Rus, and SeungJun Kim. What and when to explain? on-road evaluation of explanations in highly automated vehicles. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(3):1–26, 2023.

[11] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *IEEE Access*, 2024.

[12] Claudine Badue, Rânik Guidolini, Raphael Vivacqua Carneiro, Pedro Azevedo, Vinicius B Cardoso, Avelino Forechi, Luan Jesus, Rodrigo Berriel, Thiago M Paixao, Filipe Mutz, et al. Self-driving cars: A survey. *Expert systems with*

applications, 165:113816, 2021.

[13] Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. Why did my car just do that? explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 9:269–275, 2015.

[14] Jeamin Koo, Dongjun Shin, Martin Steinert, and Larry Leifer. Understanding driver responses to voice alerts of autonomous car operations. *International journal of vehicle design*, 70(4):377–392, 2016.

[15] Gesa Wiegand, Malin Eiband, Maximilian Haubelt, and Heinrich Hussmann. “i’d like an explanation for that!” exploring reactions to unexpected autonomous driving. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–11, 2020.

[16] Chao Wang, Thomas H Weisswange, Matti Krueger, and Christiane B Wiebel-Herboth. Human-vehicle cooperation on prediction-level: Enhancing automated driving with human foresight. In *2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*, pages 25–30. IEEE, 2021.

[17] Tobias Schneider, Sabiha Ghellal, Steve Love, and Ansgar RS Gerlicher. Increasing the user experience in autonomous driving through different feedback modalities. In *26th International Conference on Intelligent User Interfaces*, pages 7–10, 2021.

[18] Tobias Schneider, Joana Hois, Alischa Rosenstein, Sabiha Ghellal, Dimitra Theofanou-Fülbier, and Ansgar RS Gerlicher. Explain yourself! transparency for positive ux in autonomous driving. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2021.

[19] Daniel Omeiza, Helena Web, Marina Jirotka, and Lars Kunze. Towards accountability: providing intelligible explanations in autonomous driving. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 231–237. IEEE, 2021.

[20] Mehdi Zemni, Mickaël Chen, Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. Octet: Object-aware counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15062–15071, 2023.

[21] Tobias Schneider, Joana Hois, Alischa Rosenstein, Sandra Metzl, Ansgar RS Gerlicher, Sabiha Ghellal, and Steve Love. Don’t fail me! the level 5 autonomous driving information dilemma regarding transparency and user experience. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 540–552, 2023.

[22] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023.

[23] David B Leake. *Case-based reasoning: Experiences, lessons and future directions.* MIT press, 1996.

[24] Mark T Keane and Eoin M Kenny. How case-based reasoning explains neural networks: A theoretical analysis of xai using post-hoc explanation-by-example from a survey of ann-cbr twin-systems. In *Case-Based Reasoning Research and Development: 27th International Conference, ICCBR 2019, Otzenhausen, Germany, September 8–12, 2019, Proceedings 27*, pages 155–171. Springer, 2019.

[25] Frode Sørmo, Jörg Cassens, and Agnar Aamodt. Explanation in case-based reasoning–perspectives and goals. *Artificial Intelligence Review*, 24:109–143, 2005.

[26] Roger C Schank. *Dynamic memory: A theory of reminding and learning in computers and people.* Cambridge University Press, 1983.

[27] Napat Karnchanachari, Dimitris Geromichalos, Kok Seang Tan, Nanxiang Li, Christopher Eriksen, Shakiba Yaghoubi, Noushin Mehdipour, Gianmarco Bernasconi, Whye Kit Fong, Yiluan Guo, and Holger Caesar. Towards learning-based planning: The nuplan benchmark for real-world autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 629–636, 2024.

[28] Marc Heim, Francisco Suarez-Ruiz, Ishraq Bhuiyan, Bruno Brito, and Momchil S Tomov. Lab2car: A versatile wrapper for deploying experimental planners in complex real-world environments. *arXiv preprint arXiv:2409.09523*, 2024.

[29] Mica R Endsley. Autonomous driving systems: A preliminary naturalistic study of the tesla model s. *Journal of Cognitive Engineering and Decision Making*, 11(3):225–238, 2017.

[30] Gerd Gigerenzer and Ulrich Hoffrage. How to improve bayesian reasoning without instruction: frequency formats. *Psychological review*, 102(4):684, 1995.

[31] Federal Aviation Administration. Flight test guide for certification of part 23 airplanes. Advisory Circular AC No. 23-8C, U.S. Department of Transportation, 11 2011. Initiated By: ACE-100.

[32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[33] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*,

1(5):206–215, 2019.

[34] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[35] Edward E Smith and Steven A Sloman. Similarity-versus rule-based categorization. *Memory & cognition*, 22(4):377–386, 1994.

[36] Tyler Davis, Bradley C Love, and Alison R Preston. Learning the exception to the rule: Model-based fmri reveals specialized representations for surprising category members. *Cerebral Cortex*, 22(2):260–273, 2012.

8 Methods

8.1 Architecture

The black-box planner uses a modified version of the DriveIRL architecture (Figure 1b) [1]. For the trajectory generator G, we use a heuristic generator that produces 143 jerk-optimal trajectories to anchor waypoints along the route. For the scene encoder H, we use the hierarchical vector transformer (HiVT) [2] pretrained for multi-agent motion prediction. In addition to the scene embedding $\mathbf{h}$, this produces an additional 3 trajectories for the AV, for a total of $k = 146$ candidate trajectories. In the scene-trajectory encoder E, trajectories are encoded using a recurrent neural network (RNN) and then fed jointly with the scene embedding into a transformer layer which produces the scene-trajectory embeddings $\mathbf{z}_i$. The reward model R is a multilayer perceptron (MLP). In CW-Net (Figure 1c), the classifier C and the new reward model R' are MLPs.

8.2 Training

In all tests, we use one of two datasets, either with 500,000 (and 8 concept labels), or 3 million data (with 10 concept labels), for a full list of the concept labels and their meaning see Section 8.5. Each datum was associated with an additional 141 trajectories, thus giving between 70.5-423 million training data, each with multiple concept labels.

Our algorithm assumes access to the original dataset used to train the black-box planner, along with annotated human-understandable concept labels for each of these data points. The annotations can be multi-label, meaning that one datum can be associated with as many concepts as desired or useful.

During training, the parameters of the trajectory generator G, the scene encoder H, and the scene-trajectory encoder E are frozen, and only the concept classifier C and the new reward model R' are trainable. Two separate losses are trained simultaneously. First, a loss is used to train C to predict the correct concept label(s). In our setting, this loss combines cross-entropy with binary cross-entropy for different concepts, depending on the semantics of the corresponding scenario types. For example, in Dataset 1, we use cross-entropy to model the steering concepts of the car (`LEFT`, `RIGHT`, and `STRAIGHT`), and the speed concepts (`STOPPED`, `SLOW`), while also using binary cross-entropy to predict the presence of other concepts such as `ASV`, `INTERSECTION`, and `CLOSE`. These losses are then averaged into one:

$$\mathcal{L}_{\text{concept}} = \frac{1}{2k} \sum_{i=1}^{k} \left(\frac{1}{M_{\text{CCE}}} \sum_{j=1}^{M_{\text{CCE}}} \mathcal{L}_{\text{CCE}}(c_{i,j}, \hat{c}_{i,j}) + \frac{1}{M_{\text{BCE}}} \sum_{l=1}^{M_{\text{BCE}}} \mathcal{L}_{\text{BCE}}(c_{i,l}, \hat{c}_{i,l}) \right)$$

where M_{CCE} is the number of concepts modeled using categorical cross-entropy (e.g., steering and speed), and M_{BCE} is the number of concepts modeled using binary cross-entropy (e.g., presence of features like `ASV`, `INTERSECTION`, and `CLOSE`). $c_{i,j}$ and $\hat{c}_{i,j}$ represent the true and predicted labels for the j-th concept under CCE for the i-th data point, while $c_{i,l}$ and $\hat{c}_{i,l}$ represent the true and predicted labels for the l-th concept under BCE for the i-th data point. On Dataset 2, we take a different approach and model everything, including the speed concepts (`STOPPED`, `SLOW`, and `FAST`), with binary cross-entropy. These parameters can be tuned to fit the task at hand.

Secondly, a cross-entropy loss is also used to train the network to predict the correct trajectory, which we define as the original trajectory chosen by the black-box planner. Both losses are averaged:

$$\mathcal{L}_{\text{total}} = \frac{1}{2}\left(\mathcal{L}_{\text{concept}} + \mathcal{L}_{\text{trajectory}}\right)$$

A focal loss [3] is applied to counteract data imbalances, just as in the original DriveIRL planner [1]. Computationally, our networks were trained on a large distributed setup using PyTorch Lightning.

8.3 Concept separation

When adding interpretability modules post hoc as we have, there is the possibility that the network will not have learned to separate the concepts of interest, and thus fail to be able to predict them accurately [4]. In fact, we observed this in Motional's experimental prototype which we tested (see Table S3), when certain concepts such as `CLOSE` and `PEDESTRIAN` had poor precision and high recall, relatively speaking. There are two important points to note here. First, the better trained and more sophisticated an architecture is, the more it naturally learns to separate an impressive number of concepts in an unsupervised manner [5, 6] (often in the millions), so this is unlikely to be an issue for most companies with the flagship models in the future. Secondly, even if the car has not learned to separate the concepts of e.g. red traffic lights compared to green ones, this would likely highlight the reason why the car would fail to stop (or go) in such a situation, so from an explainability point of view, it would never be an issue, in fact it is potentially very useful information, which we showed in the main paper.

8.4 Alternative architecture

Alongside our primary causal architecture illustrated in Figure 1, we also developed an alternative which gave post-hoc justifications for the car's actions (Figure S1). Specifically, we simply left the black-box planner to drive the car as normal. However, we trained a concept classifier to work in parallel it to the black-box planner, which classified the scene-trajectory embeddings $\mathbf{z}_i$, and displayed these predictions while the car drove, similarly to the causal architecture. This approach is beneficial because of its relative simplicity and accessibility, although the drawback is that it may be less faithful to the model's reasoning process, as the concept classifications are not directly used by the model to rank state-trajectory pairs. However, there is ample evidence that such explanations are often faithful and capable [7, 8], so we include both as an option and demonstrate the utility of both.

8.5 Concept details

Dataset 1 concepts were as follows:

- *Steering*: A classification of either left/right/straight concepts, trained with cross entropy loss. The concept of e.g. left represents training data where the car was turning left.
- *Speed*: A classification of either slow/stopped concepts, trained with cross entropy loss. The concept of e.g. stopped represents training data where the car was stopped.
- *ASV (Approaching stopped vehicle)*: Trained with binary cross entropy. The concept represents data in which the car was approaching stopped vehicles.
- *Intersection*: Trained with binary cross entropy. The concept represents data in which the car was at an intersection.
- *Close*: Trained with binary cross entropy. The concept represents data in which the car was within 3 Meters of another vehicle.

Dataset 2 had the following concepts:

- *Slow*: Trained with binary cross entropy. The concept represents data in which the car was driving 1-2 meters per second.
- *Stopped*: Trained with binary cross entropy. The concept represents data in which the car was stationary.
- *Fast*: Trained with binary cross entropy. The concept represents data in which the car was driving faster than 2 meters per second.
- *Stop Sign*: Trained with binary cross entropy. The concept represents data in which the car was close to a stop sign.
- *Traffic Light*: Trained with binary cross entropy. The concept represents data in which the car was close to a traffic light.
- *Intersection*: Trained with binary cross entropy. The concept represents data in which the car was at an intersection.
- *Pedestrian*: Trained with binary cross entropy. The concept represents data in which the car was close to a pedestrian.
- *Following*: Trained with binary cross entropy. The concept represents data in which the car was following another vehicle.
- *Bike*: Trained with binary cross entropy. The concept represents data in which the car was close to a cyclist.
- *PUDO (Pedestrian Pickup-Drop-off)*: Trained with binary cross entropy. The concept represents data in which the car was in a pedestrian pickup drop-off zone.

Evaluation

In this section we give much greater detail about various aspects related to the evaluation in the main paper. In our tests we deployed a highly experimental AV from Motional, partly because these datasets had the necessary annotations, but also to maximize the number of potentially surprising events which would require explanation during the deployment. Our evaluation encompassed (1) a simulation phase with concept accuracy verification, (2) deployment of the AV itself, and (3) a final confirmation study. All experiments involving users obtained IRB approval.

Simulation Results

We tested our CW-Net model across the entire nuPlan validation dataset to see how its performance compared to the original black-box algorithm it was trained from. The dataset represents the world's first large-scale planning benchmark for autonomous driving, and measures how close a trained AV is to a human expert in L_2 distance. In the black-box model, when following the lane or decelerating from high speed, the planner was able to make progress along the route ($> 93\%$ of human driving distance), while avoiding collisions ($> 90\%$ collision-free) and staying close to the ground-truth human expert trajectory (< 1 m displacement at 5 s). Performance was worse when starting from a stop, with less progress (74% of human driving distance), more collisions (81% collision-free), and greater deviation from the human expert (1.2 m displacement at 5 s). Overall, the results showed our variation of the AV architecture had less than 0.01 L_2 difference to the original black-box agent on average across all measurements, and not meaningfully different, showing that it is possible to train our more interpretable model in Figure 1 without sacrificing performance. The full results are in Table S1. For the concept accuracy verification, we used 5% holdout data from our training datasets, the results are given in Table S2 and Table S3. Across both datasets, the mean accuracy was 0.54, precision 0.23, recall 0.77, and F1 Score 0.31. Overall, the results suggested that the prototype AV did not separate all concepts equally well, which suits our purposes as the explanations will highlight when and how this happens, and how it relates to driving performance, thus helping with mental model refinement (see Section 8.3). Notable results include an F1 score of 0.82 for detecting the `SLOW` concept, and < 0.00 for detecting the `BIKE` concept, showing the latter is perhaps not well encoded or understood by the car.

Distribution Comparison

In this section we demonstrate how the distribution of concept activations differs based on the deployment environment. The data here focuses on two deployments of the same model in (1) a large carpark with many tight lanes and obstacles, and (2) a large open court test track with the opportunity to drive long, straight distances at a higher speed. This serves somewhat as validation for the concept accuracy in a deployment setting. The data shown in Figure S2 is the full concept activation explanations across the entire deployments when the AV was in self-driving mode only (i.e., all data when the safety driver was in control in manual mode was deleted for this analysis). The difference between deployments is perhaps best highlighted with how the concepts for `RIGHT` and `LEFT` have generally higher activations in the parking lot compared to the large track, which involved less turning in general. Other notable differences can be seen in `SLOW` and the AV's speed, in which the parking lot had generally lower values in both. Moreover, the `STRAIGHT` concept has a higher mass on the large track, again reflecting the actual environment around it. Lastly, ASV was also higher on the large track, which was caused by issues with the trajectory generator (see Section 8.5). Overall, relative to each other, the classification of concepts reasonably represent the environment around them and give evidence our system performs fittingly in various environments.

An important note is how the AV had a large bias towards predicting the `CLOSE` and `STOP` concepts, which indicates it often conflates the environment around it with training data in which it was close to other vehicles and had to stop. Having said this, the AV also had a poor ability to predict `SLOW` correctly, but recall there is no "fast" concept here, so this concept simply refers to the AV moving. We believe this demonstrates the debugging (i.e., model improvement) power of our network as it likely accounts for the AV's general tendency to drive slowly in our tests, but this would require a long validation process to authenticate and is separate to the scope of this paper which is concerned with mental model alignment.

User Study

This section serves to give much greater detail about our user study in the main paper. We crowd sourced responses (N=120) simulating the events in the car to see if they correlated with the driver's mental model during the events, which would allow us to further extrapolate the results. Note, this is the same principle used in the U.S. Air Force called "spot checking", and similar research in academia [9, 10]. The point is to acquire additional evidence that results would generalize to a larger population, without the expense of repeating our tests in an expensive deployment environment (which was infeasible). Hence, we designed a between subjects study (N=120) to test for the effect of concept-based explanations on the accuracy of a human's mental model of the car. Both groups were shown the real videos of the three events, and asked to rate on a Likert Scale (1-7) how much they agreed with the driver's initial mental model of the event. Ideally, after viewing the explanation, they should begin to disagree with the initial mental model (which was proven wrong in our deployment study) and move towards the more correct one based on our causal architecture. We avoided asking how much they agreed with the driver's new mental model, because (1) it is best practice to minimize the number of metrics in a user study to avoid p-hacking, and (2) the explanations which essentially state this new mental model may lead users to simply agree with such a metric (e.g., one question states that the AV detects a stopped vehicle ahead, so asking people how much they agreed that the AV stopped due it detecting a stopped vehicle ahead was judged to be too leading).

Initially, subjects were given a disclaimer, introduction to the task, and a simple practice question before the main study. As an attention check we presented a video of the car driving straight, and asked the question *"I think the car drove straight because the detected the `LEFT` concept"*. As a second attention check we also measured how long users spent on each question, if they took less than 10 seconds, they would be excluded. A final survey was also given to subjects which used questions extrapolated from our post-hoc interviews with 4 safety-drivers and 1 engineer, but they are not relevant to this paper and not reported.

Materials

In total there were five videos shown to users in the main materials, three situations of the car acting in unusual ways (see Figure S4), and one attention check (a final question was removed post hoc, see discussion later). One group was given the car's parsed explanation as outlined in the main paper, and the control group a replacement explanation stating *"the car learned to drive from human expert demonstrations"*. This was the only modification in the user study between groups, thus serving to isolate the concept-based explanation as the confounding factor of interest. After seeing the explanation, subjects were presented with a statement of the driver's initial (wrong) mental model, and asked how much they agreed with it on a 7-point Likert scale.

Subjects

The users were recruited via Prolific.com, and purposefully selected to be 18+, residents of the U.S., native English speakers, and a 50/50 splits of men and women. U.S. citizens were purposefully chosen due to the car commonly being shown to drive on the right side of the road. All subjects were paid a rate of 12 USD per hour. The study obtained IRB approval from MIT.

Metrics

The measure of interest is Mann–Whitney U test between groups on all three questions. As another metric, we also averaged each respondent's scores across all questions and performed a t-test between groups as other work has done [11]. Both were two-sided tests. These two approaches allow us to analyze these data on a per-question basis, but also from a global perspective.

Results

Only one user failed the attention checks and was excluded. Figure S3 (right) displays the results of individual questions, and Figure S3 (left) the results with each user's questions averaged. Overall, the experimental group's mean was significantly lower than the control (3.37±1.63 v. 5.46±0.89; t-test $p < 0.001$; t(df)=-8.5; Cohen's d=1.58), lending strong evidence our results in deployment would generalize at scale. When considering each individual question, a large effect of explanation is observed for the scenario with `ASV` and the traffic cone (Mann-Whitney U test: $p < 0.001$; Cohen's d: 1.04), the `CLOSE` concept (Mann-Whitney U test: $p < 0.001$; Cohen's d: 0.83), and the `BIKE` (Mann-Whitney U test: $p < 0.001$; Cohen's d: 1.39).

Lastly, note that there was also an additional 4th situation involving phantom braking and the `ASV` concept, the data from this was not reported as further analysis showed the car broke not because of the `ASV` concept being activated, but rather because of a failure in the trajectory generator itself. It could be argued that the explanation pointed us towards this discovery (which it ultimately did), but we nevertheless opted to omit it. As with the other three questions in the study, this showed statistical significance in favor of the explanation group.

8.6 Data Availability

The data used for plotting in the paper's figures is available at https://drive.google.com/drive/folders/1Lz6OGGi2gFeBOnC3ddyzFJMztqUTC_Am?usp=sharing. The user study data are also available at the same address. The concept classification and ranker classification data is not available, along with the model weights, videos, and nuPlan results, due to data privacy and intellectual property issues. However, any person may reproduce similar results by training their own model on the nuPlan dataset available online, and following the instructions in the paper, although they will need to label the data with concepts of interest. Motional and MIT are happy to assist any research effort to do this.

8.7 Code Availability

CW-Net code for the models used is available at https://drive.google.com/drive/folders/1Lz6OGGi2gFeBOnC3ddyzFJMztqUTC_Am?usp=sharing. Due to Motional intellectual property issues, the code for training the AV used in the paper cannot be made available. However, code to implement and train CW-Net architectures will be available at https://github.com/EoinKenny/CW_Net, which can be used to train an interpretable agent in any domain as long as concept labels are present, this will help reproduce similar results in any domain. The full user survey will also be available to reproduce the user study in the same repo, but without the videos.

Acknowledgements

We thank Tom Bewley of JPMC AI research, Mycal Tucker of Anthropic, and Winthrop Gillis of Harvard for reviewing an early version of the manuscript.

Author contributions

E.M.K contributed to conceptualization of the research, model training, experimental evaluation, writing. A.D., S.U.L., T.P.M., S.R., Y.H., and M.S.T. all contributed to the technical implementation of the algorithm in the self-driving car. L.M. contributed to project organization, and writing. M.S.T. and J.A.S. contributed to conceptualization of the research, project organization, and writing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information available at https://drive.google.com/drive/folders/1Lz6OGGi2gFeBOnC3ddyzFJMztqUTC_Am?usp=sharing and https://github.com/EoinKenny/CW_Net. Correspondence and requests for materials should be addressed to Eoin M. Kenny.

References

[1] Tung Phan-Minh, Forbes Howington, Ting-Sheng Chu, Momchil S Tomov, Robert E Beaudoin, Sang Uk Lee, Nanxiang Li, Caglayan Dicle, Samuel Findler, Francisco Suarez-Ruiz, et al. Driveirl: Drive in real life with inverse reinforcement learning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1544–1550. IEEE, 2023.

[2] Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8823–8833, 2022.

[3] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[4] Simon Schrodi, Julian Schur, Max Argus, and Thomas Brox. Concept bottleneck models without predefined concepts. *arXiv preprint arXiv:2407.03921*, 2024.

[5] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

[6] Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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*, pages 1135–1144, 2016.

[8] M Scott, Lee Su-In, et al. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30:4765–4774, 2017.

[9] Federal Aviation Administration. Flight test guide for certification of part 23 airplanes. Advisory Circular AC No. 23-8C, U.S. Department of Transportation, 11 2011. Initiated By: ACE-100.

[10] Tobias Schneider, Joana Hois, Alischa Rosenstein, Sandra Metzl, Ansgar RS Gerlicher, Sabiha Ghellal, and Steve Love. Don't fail me! the level 5 autonomous driving information dilemma regarding transparency and user experience. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 540–552, 2023.

[11] Eoin M Kenny, Courtney Ford, Molly Quinn, and Mark T Keane. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in xai user studies. *Artificial Intelligence*, 294:103459, 2021.

Supplementary Material

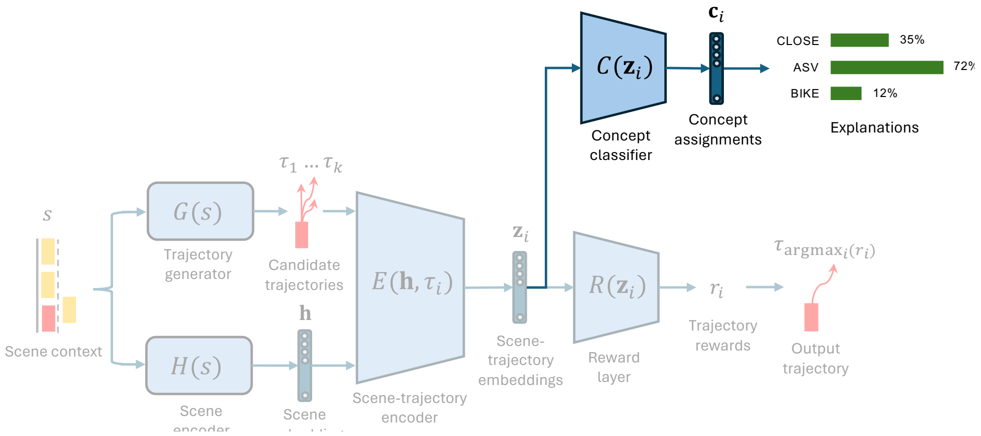


Fig. S1 **Parallel architecture:** This is identical to the black-box planner (Figure 1b), except the scene-trajectory embeddings are fed to the concept classifier C in parallel to the (original) reward model R. These concept classifications are then converted to probabilities (x100 to convert to percentages) and presented to the user.

Scenario	Metric	Our Model	Original
nominal lane follow (968)	Average L2 error	4.202058	4.200189
	Average L2 error 3s	0.401251	0.400440
	Average L2 error 5s	0.780286	0.778495
	Average L2 error 10s	2.141165	2.136004
	Progress along expert route	0.939595	0.939382
	No at fault collisions	0.909091	0.909091
decelerating from high speed scenarios (1099)	Average L2 error	3.139028	3.144618
	Average L2 error 3s	0.533925	0.533905
	Average L2 error 5s	0.930384	0.929786
	Average L2 error 10s	2.056039	2.053457
	Progress along expert route	0.990187	0.990213
	No at fault collisions	0.945405	0.942675
	Deceleration time difference	0.443201	0.439381
start accelerating from stationary scenarios (919)	Average L2 error	8.919248	9.019697
	Average L2 error 3s	0.426693	0.425365
	Average L2 error 5s	1.244357	1.249598
	Average L2 error 10s	4.611888	4.673006
	Progress along expert route	0.740881	0.736437
	No at fault collisions	0.807399	0.803047
	Start from stationary max speed difference	0.239367	0.241552
	Start from stationary time delay	3.477263	3.494675

Table S1 Full nuPlan results which compares the performance of our main model used in testing against the original black-box IRL agent.

	Accuracy	Precision	Recall	F1 Score
Steering	0.62	(cross entropy)		
Speed	0.83	(cross entropy)		
Approaching stopped vechicle	0.55	0.2	0.89	0.33
Intersection	0.51	0.42	0.61	0.49
CLOSE to another vechicle	0.45	0.1	0.81	0.17
Ranker Accuracy	0.94			

Table S2 Concept performance for Dataset 1 for CW-Net

	Accuracy	Precision	Recall	F1 Score
Slow	0.83	0.73	0.93	0.82
Stopped	0.48	0.16	0.81	0.27
Fast	0.68	0.49	0.86	0.63
Stop sign	0.43	0.03	0.84	0.05
Traffic light	0.39	0.16	0.62	0.25
Intersection	0.42	0.20	0.65	0.31
Pedestrian	0.44	0.03	0.84	0.07
Following	0.43	0.02	0.84	0.04
BIKE	0.21	0.00	0.42	0.00
PUDO	0.58	0.30	0.86	0.44
Ranker Accuracy	0.95			

Table S3 Concept performance for dataset 1 with CW-Net.

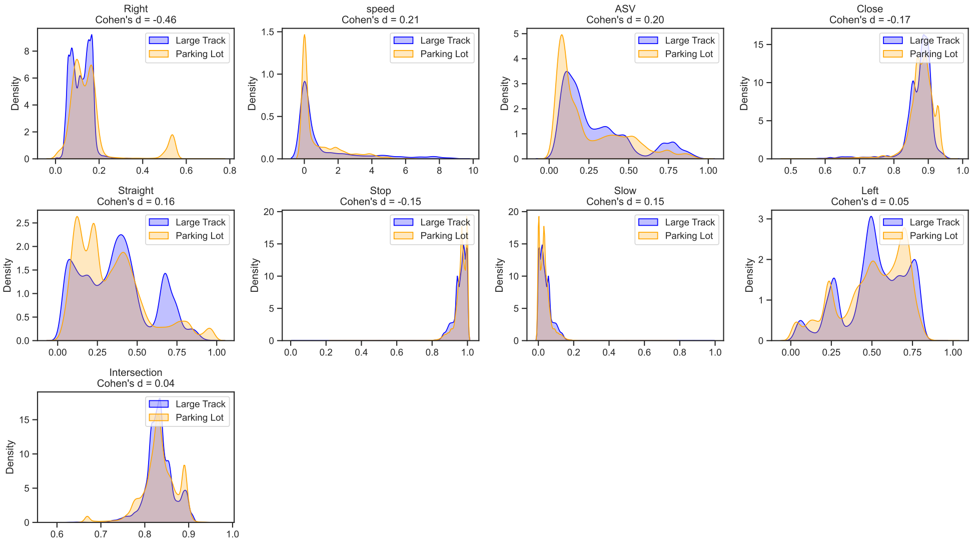


Fig. S2 A comparison of the concept predictions for the same model deployed in a large test track with wide open, long straight roads, and a smaller parking lot with many turns.

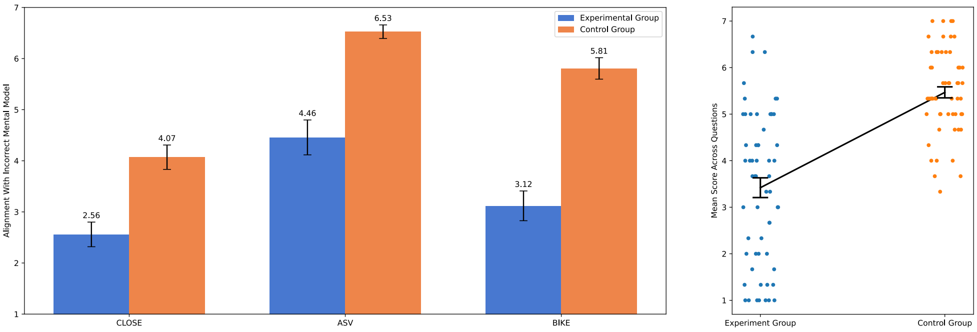
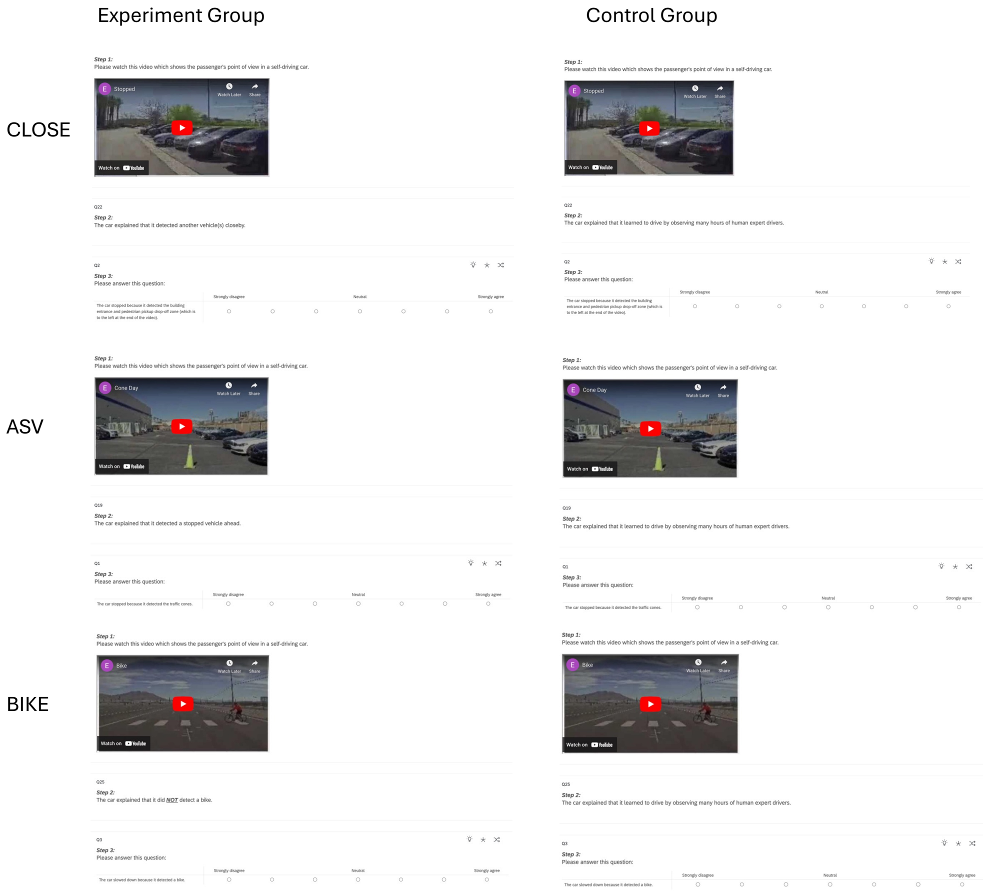


Fig. S3 Statistical results for user study: (left) Each individual question.(right). Averaged across all questions



Experiment Group

CLOSE

Step 1:
Please watch this video which shows the passenger's point of view in a self-driving car.

Q22
Step 2:
The car explained that it detected another vehicle(s) closeby.

Q2
Step 3:
Please answer this question:

	Strongly disagree			Neutral			Strongly agree
The car stopped because it detected the building entrance and pedestrian pickup drop-off zone (which is to the left at the end of the video).	○	○	○	○	○	○	○

ASV

Step 1:
Please watch this video which shows the passenger's point of view in a self-driving car.

Q19
Step 2:
The car explained that it detected a stopped vehicle ahead.

Q1
Step 3:
Please answer this question:

	Strongly disagree			Neutral			Strongly agree
The car stopped because it detected the traffic cones.	○	○	○	○	○	○	○

BIKE

Step 1:
Please watch this video which shows the passenger's point of view in a self-driving car.

Q25
Step 2:
The car explained that it did ***NOT*** detect a bike.

Q3
Step 3:
Please answer this question:

	Strongly disagree			Neutral			Strongly agree
The car slowed down because it detected a bike.	○	○	○	○	○	○	○

Control Group

CLOSE

Step 1:
Please watch this video which shows the passenger's point of view in a self-driving car.

Q22
Step 2:
The car explained that it learned to drive by observing many hours of human expert drivers.

Q2
Step 3:
Please answer this question:

	Strongly disagree			Neutral			Strongly agree
The car stopped because it detected the building entrance and pedestrian pickup drop-off zone (which is to the left at the end of the video).	○	○	○	○	○	○	○

ASV

Step 1:
Please watch this video which shows the passenger's point of view in a self-driving car.

Q19
Step 2:
The car explained that it learned to drive by observing many hours of human expert drivers.

Q1
Step 3:
Please answer this question:

	Strongly disagree			Neutral			Strongly agree
The car stopped because it detected the traffic cones.	○	○	○	○	○	○	○

BIKE

Step 1:
Please watch this video which shows the passenger's point of view in a self-driving car.

Q25
Step 2:
The car explained that it learned to drive by observing many hours of human expert drivers.

Q3
Step 3:
Please answer this question:

	Strongly disagree			Neutral			Strongly agree
The car slowed down because it detected a bike.	○	○	○	○	○	○	○

Fig. S4 Survey questions for the experimental and control group.