Proceedings of Machine Learning Research vol 242:1–14, 2024

A framework for evaluating human driver models using neuroimaging

Christopher A. Strong CHRISTOPHER_STRONG@BERKELEY.EDU
University of California, Berkeley, Department of Electrical Engineering and Computer Sciences

Kaylene C. Stocking KAYLENE@BERKELEY.EDU
University of California, Berkeley, Department of Electrical Engineering and Computer Sciences

Jingqi Li JINGQILI@BERKELEY.EDU
University of California, Berkeley, Department of Electrical Engineering and Computer Sciences

Tianjiao Zhang T.ZHANG@BERKELEY.EDU
University of California, Berkeley, Helen Wills Neuroscience Institute

Jack L. Gallant GALLANT@BERKELEY.EDU
University of California, Berkeley, Department of Psychology

Claire J. Tomlin TOMLIN@EECS.BERKELEY.EDU
University of California, Berkeley, Department of Electrical Engineering and Computer Sciences

Abstract

Driving is a complex task which requires synthesizing multiple senses, adapting to a constantly changing environment, and safely reasoning about the behavior of others. Failures of human driver models can become failures of vehicle safety technology or autonomous driving systems that rely on their predictions. Although there has been a variety of work to model human drivers, it can be challenging to determine to what extent they truly resemble the humans they attempt to mimic. The development of improved human driver models can serve as a step towards better vehicle safety. In order to better compare and develop driver models, we propose going beyond driving behavior to examine how well these models reflect the *cognitive* activity of human drivers. In particular, we compare features extracted from human driver models with brain activity as measured by functional magnetic resonance imaging. We demonstrate this approach on three human driver models with brain activity data from two human subjects. We find that model predictive control is a better fit for driver brain activity than classic non-predictive models, which is in good agreement with previous works that obtain better predictions of human driving behavior using model predictive control.

Keywords: Human Driver Modeling, Neuroscience, fMRI, Model Predictive Control

1. Introduction

Because autonomous vehicles often operate in dynamic environments with humans, it is important for them to be able to understand and predict human motion. These predictions are safety-critical, as their outputs inform potentially life-threatening decisions made by autonomous vehicles. As a result, there have been substantial efforts to develop accurate human driver models (Brown et al., 2020; Mozaffari et al., 2020; Leon and Gavrilescu, 2021). Additionally, while there exist increasingly large driving datasets used to fit human driver models (Kang et al., 2019; Ettinger et al., 2021; Houston et al., 2021), these datasets only include driving behavior, treating the driver as a black box and providing no insight on the cognitive algorithms that output these behaviors. While models fitted to such datasets may replicate human behavior to a certain extent, there is no guarantee they use the same algorithms and as a result may not generalize to novel situations.

In this work, we aim to open up this black box by examining the brain activity underlying driving behavior. We propose a novel framework for evaluating human driver models not only by their ability to match the behavior of human drivers, but also their ability to match the algorithms used by human drivers to produce those behaviors. Concretely, we use functional magnetic resonance imaging (fMRI) to record brain activity during driving, and apply a data-driven regression-based approach to evaluate how well the computations underlying human driver models matches those performed by the brain. We provide a preliminary demonstration of this framework with a case study that evaluates three human driver models.

Key Contribution. This paper proposes a novel framework for comparing algorithms used by human driver models with those used by humans while driving. It validates the framework by applying it to three human driver models with brain activity from two human subjects.

2. Background and related work

Current models of human driving behavior generally fall into one of three categories: dynamics-based, data-driven, and optimization-based (Rudenko et al., 2020). Dynamics-based models predict driver behavior reactively based on current conditions, e.g. the distance to the car in front (Treiber et al., 2000; Bando et al., 1995). Data-driven models base their predictions on what drivers have been observed to do in similar situations in the past (Salzmann et al., 2020; Chen and Krähenbühl, 2022). Finally, optimization-based models predict that the driver will choose control actions that optimize a reward function (Prokop, 2001; Guo et al., 2013; Sundaram et al., 2022). Models from all classes have free parameters that can be tuned to better match the behavior in datasets of human driving (Burnham et al., 1974; Bekey et al., 1977; Lee et al., 2022; Xu et al., 2022). However, to date this fitting process have relied solely on drivers' behavior, and it is unclear to what extent these models accurately reflect the internal processes humans use to produce that behavior. This is important because if the inferred features or reward functions are incorrect, the models will likely fail to generalize appropriately to novel situations (Fu et al., 2017).

Fortunately, human drivers are not a "black box," and there has been an increasing interest in the neural basis of human driving behavior over the past two decades. Noninvasive brain recording techniques such as electroencephalography, functional near-infrared spectroscopy, and magnetoencephalography have all been investigated (Haghani et al., 2021), but functional magnetic resonance imaging (fMRI) provides the highest spatial resolution and enables the localization of brain activity (Gross, 2019). Both Spiers and Maguire (2007) and Schweizer et al. (2013) have used fMRI to examine brain activity from subjects performing simulated driving tasks. Navarro et al. (2018) pooled these and seven other fMRI studies of driving to find characteristic patterns of brain activity for low-level tasks such as turning, medium-level tasks such as planning to overtake, and high-level tasks such as route planning, and related these patterns to conceptual (non-predictive) models of human driving. However, the analysis is hampered by contrast-based fMRI analysis methods that only produce activation maps and provide little insight into the computations performed by the brain.

3. Approach

We propose a framework for evaluating human driver models that incorporates the brain. In this framework, we first record (i) brain activity of drivers, (ii) the visual and auditory stimuli seen

by drivers and the task performed by drivers, and (iii) the resulting behavior of drivers. Next, we tune any free parameters in the human driver models to match the observed behavior. Finally, we examine whether features derived from the human driver models are represented in the brain activity of the drivers. To implement this framework, we used fMRI to collect brain activity from two drivers performing a taxi-driver task in a virtual world. We then implemented three car-following human driver models: the optimal velocity model (OVM), the intelligent driver model (IDM), and model predictive control (MPC) (Bando et al., 1995; Treiber et al., 2000; Guo et al., 2013). We fit these models to the behavior of the drivers. Finally, we used voxelwise modeling (Naselaris et al., 2011) to investigate which human driver models use representations and algorithms most similarly to the human brain. This framework enables us to select for human driver models that not only match human behavior, but also use algorithms similar to those used by the brain.

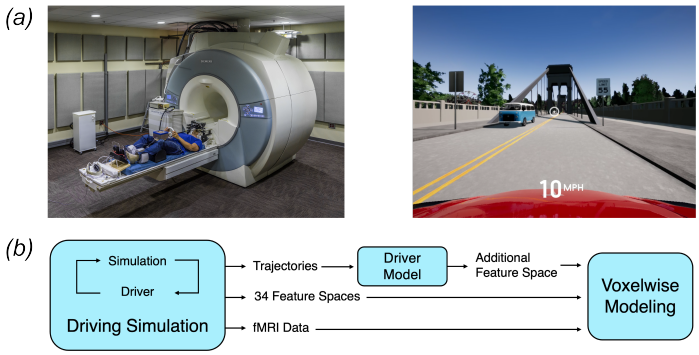


Figure 1: **An overview of the framework to evaluate human driver models with brain activity**. (a) Subjects used a custom MR-safe steering wheel and pedal set (left) to drive in a naturalistic virtual world (right) while brain activity was recorded with fMRI. (b) A block diagram of the proposed framework. First, subjects drive in a driving simulator while inside the MRI scanner. Trajectories of driving behavior, stimulus and task data, and brain activity are simultaneously recorded. 34 feature spaces are computed from the stimulus and task data. A human driver model to be evaluated is tuned to match the observed behavior, then a feature space consisting of the driver model's inputs, intermediate computations, and outputs at each moment in time is gathered. Finally, voxelwise modeling is used to evaluate the human driver model: features produced by the model, along with the stimulus and task feature spaces, are regressed onto brain activity to produce encoding models. The ability of the encoding models from different driver models to predict brain activity provides a metric of the similarity of the human driver models to algorithms implemented by the brain.

3.1. Behavioral and brain data

In this work we used fMRI to record blood-oxygenation-level dependent (BOLD) (Ogawa et al., 1990) activity from the brain while subjects drove in a large realistic virtual city. We used Unreal Engine 4 and Carla (Dosovitskiy et al., 2017) to build a 2×3 km virtual city populated by dynamic pedestrians and vehicles where participants drove a virtual car using an MR-compatible steering wheel and pedals (figure 1). Prior to scanning, subjects learned the layout of the world. In the scanner, subjects performed a taxi-driver task in which they were cued to navigate to particular locations. Data were collected in 11-minute runs (180 minutes for subject 1 and 110 minutes for

subject 2). The experimental procedures were approved by the Institutional Review Board at the University of California, Berkeley, and written informed consent was obtained from all participants.

3.2. Human driver models

We implemented three driver models: OVM, IDM, and MPC. Each model has several **free parameters** that can be tuned to affect their predictions. At each time point we can extract a given model's input, intermediate computations, and output to obtain a set of **features** that can be compared with brain activity. The output of all three models is the predicted acceleration along the road. We denote the longitudinal position of the subject's vehicle as y_{ego}, the position of the nearest vehicle in front of the subject in their lane as y_{other}, and the gap between them as $\Delta y = y_{\text{other}} - y_{\text{ego}}$. The longitudinal velocity of the subject's vehicle is v_{ego}, the other vehicle v_{other}, the difference between these is $\Delta v = v_{\text{other}} - v_{\text{ego}}$. The vehicle length is l. The modeling setup is visualized in figure 2. When rolling out a model to evaluate its behavior, the predicted acceleration is used to update the velocity at the next time step in a closed loop. In comparison, when gathering features from the model to compare to the brain activity the ground truth state at every timepoint serves as the model input. If there is no vehicle in front, the features that rely on the other vehicle are set to 0.

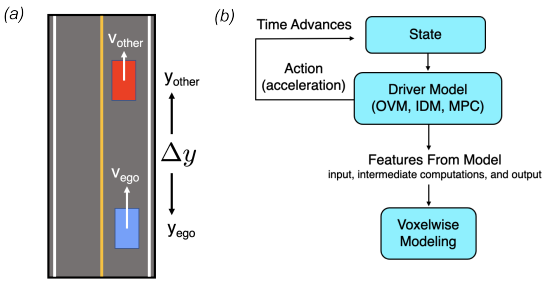


Figure 2: **Visualization of the simplified human driver modeling setup**. (a) The subject's vehicle (in blue) following another vehicle (in red). (b) A block diagram showing how each driver model takes in the current state and outputs the acceleration and a set of features consisting of the model's input, intermediate computations, and output. This corresponds to the driver modeling block in figure 1 (c).

Optimal velocity model. OVM models the acceleration along the road by setting a desired velocity as a function of the distance to the next vehicle $V(\Delta y)$ and accelerating proportionally to the velocity error (Bando et al., 1995, 1998). In discrete time, this can be written as

$$v_{\text{ego}}[k+1] = v_{\text{ego}}[k] + a\left[V(\Delta y[k]) - v_{\text{ego}}[k]\right], \quad V(\Delta y) = \text{scale} \cdot (\tanh(d(\Delta y) - e)) + \text{shift}$$

with 5 free parameters a, scale, d, e, and shift. The free parameter a scales how aggressively you accelerate while the remaining parameters stretch and shift the desired velocity function. The features representing this model, consisting of its inputs, intermediate computations, and outputs, are

$$\left\{\mathbb{1}_{\Delta y<\infty}, \Delta y, v_{\text{ego}}, V(\Delta y), a[V(\Delta y) - v_{\text{ego}}]\right\}$$

where $\mathbb{1}_A$ is 1 when A is true and 0 otherwise.

Intelligent driver model. IDM models the acceleration along the road by trying to maintain a desired velocity v_0 and minimum gap $s^*(v_{\text{ego}}, \Delta v)$ with the form

$$s^*(v_{\text{ego}}, \Delta v) = s_0 + T v_{\text{ego}} + \frac{v_{\text{ego}} \Delta v}{2\sqrt{ab}}$$

for maximum acceleration a, comfortable braking deceleration b, and desired minimum spacing s_0. The gap between vehicles is $s = \Delta y - l$. In discrete time the model is

$$v_{\text{ego}}[k+1] = v_{\text{ego}}[k] + \Delta T \cdot \tilde{a}[k], \quad \tilde{a}[k] = a\left[1 - \left(\frac{v_{\text{ego}}[k]}{v_0}\right)^{\delta} - \left(\frac{s^*(v_{\text{ego}}[k], \Delta v[k])}{s[k]}\right)^2\right]$$

with 7 free parameters $v_0, s_0, \delta, T, a, b$ and l. The free parameter δ affects how deviations from the desired velocity are responded to, and T is the desired time headway. To improve stability, the acceleration is clipped to between -7m/s^2 and a. The features are

$$\left\{\mathbb{1}_{\Delta y<\infty}, \Delta y, v_{\text{ego}}, v_{\text{other}}, \Delta v_{\text{ego}}, (\frac{v_{\text{ego}}}{v_0})^{\delta}, \frac{v_{\text{ego}}\Delta v}{2\sqrt{ab}}, s^*(v_{\text{ego}}, \Delta v), \frac{s^*}{s}, \left(\frac{s^*}{s}\right)^2, \tilde{a}\right\}$$

Model predictive control. MPC models behavior by (i) finding an optimal sequence of actions for the near future with respect to a cost function and (ii) applying the first action, then repeating this process. The longitudinal positions y_{ego} and y_{other} are modeled as double integrators and the acceleration of the modeled vehicle is optimized at each moment while the acceleration of the other vehicle is assumed to stay constant. The cost function penalizes being too close to the vehicle in front, deviations from a desired velocity, and control effort. The joint state vector is defined as $\boldsymbol{x}[k] := \left[y_{\text{ego}}[k], v_{\text{ego}}[k], y_{\text{other}}[k], v_{\text{other}}[k]\right]^\top$. Let $u_{\text{ego}}[k]$ be the subject vehicle's acceleration and $u_{\text{other}}[k]$ be the other vehicle's acceleration. This results in linear longitudinal dynamics:

$$\boldsymbol{x}[k+1] = A \cdot \boldsymbol{x}[k] + B_{\text{ego}} \cdot u_{\text{ego}}[k] + B_{\text{other}} \cdot u_{\text{other}}[k]$$

where A, B_{ego}, and B_{other} are discrete-time approximations of the dynamics. We let our cost function be

$$c(\boldsymbol{x}[k], u[k]) := (\text{ReLU}(d_{\text{safe}} - \Delta y[k]))^2 + \theta_1(v_{\text{ego}}[k] - v^*)^2 + \theta_2(u_{\text{ego}}[k])^2$$

where ReLU represents the rectified linear unit activation function, θ_1 scales the velocity error cost, θ_2 scales the control cost, and v^* is the desired velocity. Our optimization problem is then

$$\begin{aligned}
\underset{\{\boldsymbol{x}[k], u_{\text{ego}}[k]\}_{k=0}^{T}}{\text{minimize}} \quad & \sum_{k=0}^{T} c(\boldsymbol{x}[k], u_{\text{ego}}[k]) \\
\text{subject to} \quad & \boldsymbol{x}[0] = \boldsymbol{x}_0 \\
& \boldsymbol{x}[k+1] = A \cdot \boldsymbol{x}[k] + B_{\text{ego}} \cdot u_{\text{ego}}[k] + B_{\text{other}} \cdot u_{\text{other}}[k], \quad k = 0, \ldots, T-1 \\
& u_{\text{ego}}[k] \in [-1, 1], \; u_{\text{other}}[k] = u_{\text{front}}, \quad k = 0, \ldots, T \\
& d_{\text{safe}}[k] = d_{\text{safe}_0} + v_{\text{ego}}[k] \cdot d_{\text{safe}_1}, \quad k = 0, \ldots, T
\end{aligned}$$

with 6 free parameters $\theta_1, \theta_2, d_{\text{safe}_0}, d_{\text{safe}_1}, v^*$, and T . The free parameter d_{safe_0} is the comfortable distance when not moving, d_{safe_1} is the rate of growth of the comfortable distance as v_{ego} grows, and T is the time horizon. The features are

$$\{v_{\text{ego}}[k], v_{\text{other}}[k], \Delta y[k], d_{\text{safe}}[k] - \Delta y[k]\}_{k=0}^{T} \cup \{u_{\text{ego}}[k]\}_{k=0}^{T-2} \cup \{c(\boldsymbol{x}[k], u_{\text{ego}}[k])\}_{k=0}^{T} \cup \{u_{\text{front}}\} \quad (1)$$

as well as several induced terms $\{(v_{\text{other}}[k] - v^*)^2, \mathbb{1}_{d_{\text{safe}} - \Delta y[k] \le 0}, (d_{\text{safe}} - \Delta y[k])^2, \text{ReLU}(d_{\text{safe}} - \Delta y[k]), \text{ReLU}(d_{\text{safe}} - \Delta y[k])^2\}_{k=0}^{T} \cup \{u_{\text{ego}}[k]^2\}_{k=0}^{T-2} \cup \{\sum_{k=0}^{T} c(\boldsymbol{x}[k], u_{\text{ego}}[k])\} \cup \{\mathbb{1}_{\Delta y \le \infty}\}$. If there

is no vehicle in front, then (3.2) is solved by dropping terms related to the front car.

Optimizing the behavior of human driver models. We focus on segments where the subject is not turning to match the car-following scenario for which OVM and IDM are designed. The free parameters for the models were tuned with three losses: a "full rollout" loss, where the model was rolled out for a full straight segment and the loss is defined as the average squared position error; a "subtrajectory loss" where the predictions were rolled out for 4 seconds starting at points spaced evenly along the segment at 2 second intervals and the loss is defined as the average squared position error across all rollouts; and an "acceleration loss," where the loss equals the average squared error between the predicted and actual accelerations, and where the prediction at each moment across the segment is evaluated independently (the model is not rolled out in a closed loop). For each loss, we tuned the human driver models via a grid search over the free model parameters.

3.3. Testing algorithmic similarity between human driver models and the brain

We applied the voxelwise modeling (VM; Naselaris et al. (2011); Nunez-Elizalde et al. (2019); Dupré la Tour et al. (2022)) framework to the features generated from OVM, IDM, and MPC to test whether these three human driver models used algorithms similar to those implemented by the human brain. In VM, the stimulus and task are first nonlinearly transformed into feature spaces that are hypothesized to capture some particular aspects of the stimulus and task. Then, the time series of activity in each brain voxel is modeled as a linear combination of the feature time series. Models are evaluated by predicting brain activity on a held-out dataset. High prediction performance suggests that the brain represents information as parameterized by that feature space (Naselaris et al., 2011). Intuitively, each feature space can be considered as a hypothesis of how information is represented in the brain; the model's prediction performance then tests the hypothesis and provides an estimate of its effect size. VM has enabled new insights about brain activity while subjects listen to narrative stories (Huth et al., 2016) and watch movies (Nishimoto et al., 2011; Huth et al., 2012).

We fit voxelwise encoding models separately for each of the three human driver models to every voxel in each subject. In addition to the human driver models, 34 other feature spaces were included to capture other aspects of the taxi-driver task such as the visual motion-energy content of the scene or progression along a planned route (Zhang, 2021). The inclusion of these additional feature spaces enables the regression procedure to accurately attribute variance to the representation of human driver model features, minimizing confounding effects due to correlation with other stimulus or task variables. The hemodynamic response function for each voxel was modeled by a finite impulse response (FIR) filter with five delays, where the shape of the filter varies across features. Banded ridge regression (Nunez-Elizalde et al., 2019; Dupré la Tour et al., 2022) was used to impose a different regularization parameter on each feature space for each voxel. The optimal feature weights, regularization parameters, and FIR filter shapes were empirically selected by cross-validating over 9000 random samples. The overall model performance was quantified by computing the R^2 score (explained variance) between predicted and actual activity in each voxel on a held-out test set. Individual model performances for each of the 35 feature spaces were determined by partitioning the R^2 between feature spaces (Pratt, 1987; Dupré la Tour et al., 2022). Since this task is interactive and open-ended, no experimental conditions could be perfectly repeated twice, precluding the calculation of a noise ceiling for encoding model performances.

Table 1: **The MPC human driver model performs comparably to IDM and better than OVM at matching behavior**. To test how well IDM, OVM, and MPC predict behavior, each model was evaluated on a held-out test set. On the subtrajectory and acceleration losses, MPC was comparable to IDM, while on the full rollout loss MPC performs inconsistently across subjects. With respect to their behavioral outputs MPC and IDM are comparable and both are better than OVM.

Model	Full Rollout Loss (m^2) S1 | S2	Subtrajectory Loss ($(m/s^2)^2$) S1 | S2	Acceleration Loss (m^2) S1 | S2
OVM	502 | 1335	18.8 | 45.8	3.8 | 7.1
IDM	**440** | 1179	13.9 | **43.3**	**3.2** | 6.8
MPC	621 | **1164**	**13.2** | 43.8	3.4 | **6.3**

4. Results

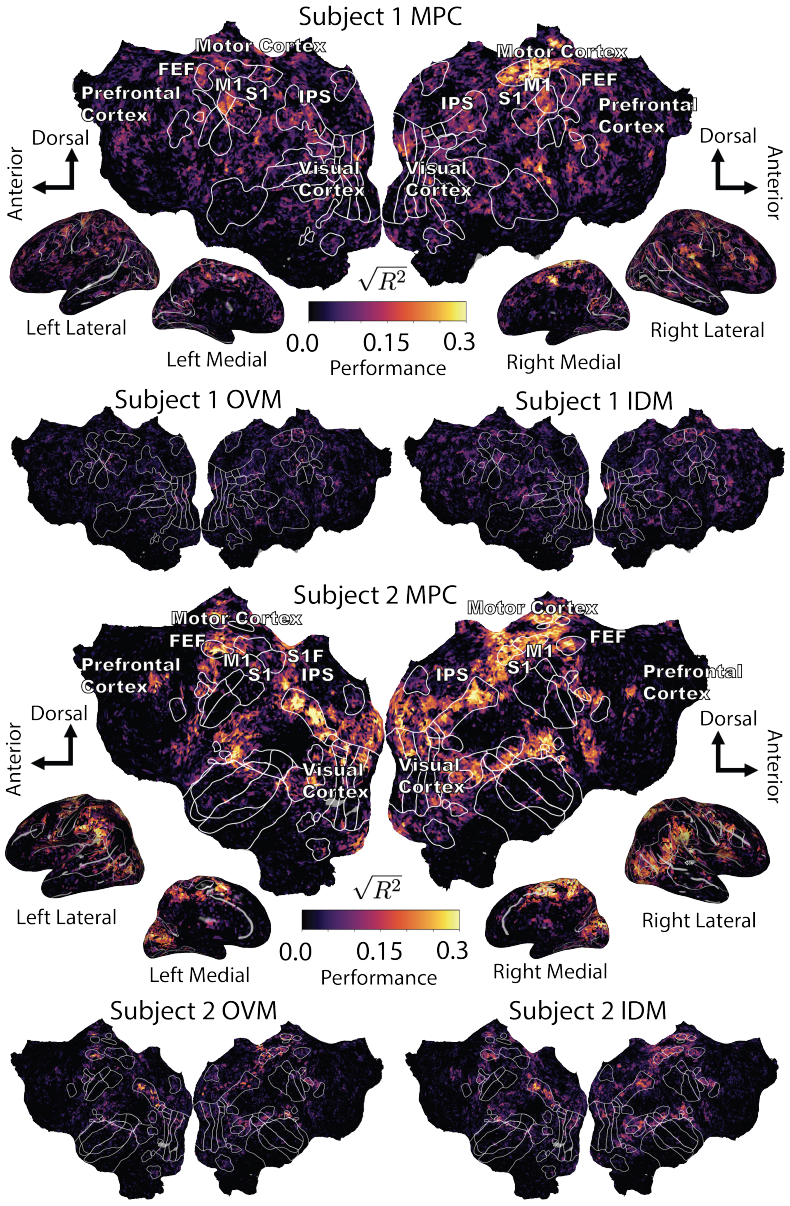
In order to evaluate how closely human driver models match human drivers both behaviorally and algorithmically, the proposed framework consists of three steps. The first is to acquire paired brain activity and driving data. The second is to tune the human driver models and enumerate a set of features for each. The third is to compare these features to the brain activity. Here, we demonstrate this framework with fMRI data from two subjects driving in a realistic simulation. We chose the OVM, IDM, and MPC human driver models and used the voxelwise modeling framework to test the algorithmic similarity of these models to human drivers. The free parameters for each model were tuned to match the behavior of the two participants using the subtrajectory, acceleration, and full rollout losses described in section 3.2. Table 1 shows the performance of each model on a held-out test set for each subject. MPC and IDM perform comparably, and both outperform OVM.

While both MPC and IDM match driver behavior to a similar degree, they use different algorithms to do so. Figure 3 shows voxelwise modeling results from human driver models fit to the subtrajectory loss. Models fit using the subtrajectory loss were chosen as this loss represents a balance between immediate action and longer horizon planning. The MPC model explains more variance in brain activity than the OVM and IDM models, particularly in the intraparietal sulcus (IPS), frontal eye field (FEF), and the supplementary motor area (SMA). Histograms of the variance explained (R^2) scores are shown in figure 3 (b). These results suggest that while MPC and IDM can behaviorally match human drivers equally well, the representations and algorithms used by MPC are more similar to those implemented by the human brain. As a result, of the three human driver models tested, MPC best captures human drivers' cognitive activity.

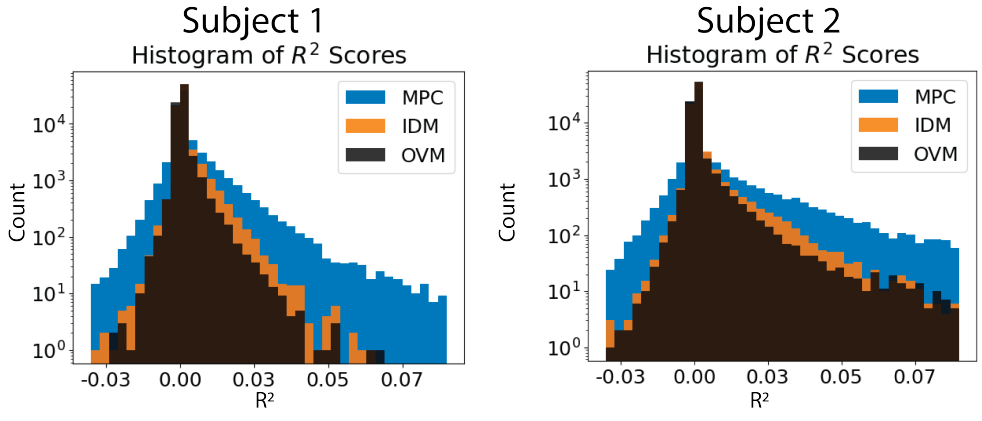
5. Discussion

Behaviorally, MPC and IDM captured human drivers comparably, and both outperformed OVM. However, the MPC encoding model explains more variance in the brain than the IDM or OVM models, suggesting that the computations of the MPC model better reflect those used by the human brain. While previous work has shown advantages of MPC over IDM in empirically matching observed behavior (Guo and Jia, 2019), here we provide fundamentally novel evidence for the advantage of MPC over IDM: by including brain activity, we support MPC being more closely aligned at an algorithmic level with how humans drive. Including brain activity enabled us to distinguish models with similar predictive performances of behavior.

Figure 3


Subject 1 MPC
Motor Cortex
FEF
Prefrontal Cortex
M1
S1
IPS
Visual Cortex
Dorsal
Anterior
Left Lateral
Left Medial
Right Medial
Right Lateral
$\sqrt{R^2}$
0.0
0.15
0.3
Performance
Subject 1 OVM
Subject 1 IDM
Subject 2 MPC
S1F
Subject 2 OVM
Subject 2 IDM

(*a*)


Subject 1
Histogram of R^2 Scores
Subject 2
MPC
IDM
OVM
Count
10^0
10^1
10^2
10^3
10^4
-0.03
0.00
0.03
0.05
0.07
R²

(*b*)

Figure 3: **The MPC human driver model explains much more variance in brain activity than IDM or OVM**. To test whether the OVM, IDM, and MPC human driver models use algorithms similar to that used by the human brain, we used the inputs, internal representations, and outputs of each of the three human driver models to create voxelwise encoding models of brain activity. We then observed which models better predict brain activity. In (a) we show the performance of voxelwise models for subtrajectory loss optimized OVM, IDM, and MPC on the flattened cortical surfaces of both subjects (inflated views are shown for reference). The first row shows the MPC model performance in subject 1, and the second row shows the OVM and IDM model performances in subject 1. The third and fourth rows follow the same format for subject 2. Brighter voxels are better predicted. In (b) we show the performance of these same models through histograms of the R^2 scores for each voxel. We find that the MPC model explains much more variance in brain activity than the IDM or OVM models, and this is consistent across subjects. Well-predicted regions include the IPS, FEF, and SMA, all of which have previously been implicated in visuomotor control; the model additionally predicts brain activity in the primary foot motor area, which controls the action of the subject. The ability of the MPC encoding model to explain brain activity in known visuomotor control regions suggests that MPC implements a more human-like algorithm than IDM or OVM.

The MPC encoding model was able to explain brain activity in multiple functional regions, including visual attention regions such as the intraparietal sulcus (IPS) and the frontal eye field (FEF), and motor planning regions such as the supplementary motor area (SMA). The IPS is associated with visuomotor tasks and coordinate transforms (Grefkes et al., 2004; Grefkes and Fink, 2005), the FEF is associated with eye movement and attention (Crowne, 1983; Paus, 1996), and the SMA is associated with motor planning (Tanji and Shima, 1994; Gerloff et al., 1997; Hoshi and Tanji, 2004; Nachev et al., 2008). Both the SMA and IPS have previously been shown to have significant activity during driving (Navarro et al., 2018), but their representational properties for driving have not been well-characterized. Additionally, the MPC encoding model explains brain activity in many areas outside known functional regions, particularly in the prefrontal cortex. The MPC encoding model does not predict as well in visual regions, such as the human middle temporal cortex (hMT+), that had previously been implicated in driving (Navarro et al., 2018). However, this difference in fact provides further evidence that MPC is a good human driver model. While vision is necessary for driving, the visual system likely does not perform the underlying computations for control. Previous analyses may have been confounded by the co-activation of vision and control during driving. Here, the specificity of MPC only to the control algorithms of the brain likely enabled its encoding models to be decorrelated from vision.

Thus, the brain data provides two targets against which to optimize human driver models: maximize encoding model performance in known motor control regions, such as SMA, IPS, and the motor cortex, and minimize spurious predictions in co-activated regions, such as the visual cortex. These results provide instructive examples of how the voxelwise modeling framework allows us to examine human driver models with finer granularity than possible when using behavior alone.

One potential concern is that this work includes data from only two subjects, while previous fMRI studies of human driving have upwards of 20 (Spiers and Maguire, 2007). However, we note that while previous studies had a much larger number of subjects, they collected very few data in each individual subject (20-30 minutes each). Because fMRI is a noisy modality, this small amount of data is insufficient to provide detailed descriptions of brain activity in individual subjects. To

ameliorate this, previous studies have projected individual subject data into a common space, and averaged across subjects. However, because there are significant individual differences in both structural and functional neuroanatomy, this averaging loses significant amounts of information about individual subjects. On the other hand, here we have collected a large amount of data for each subject, roughly 2 hours driving and 4-6 hours of localizers and anatomical data. This enables us to fit accurate encoding models that have detailed descriptions of brain function in individual subjects. Another concern could be that although all driver models had a similar number of free parameters for tuning their behavior (5, 7, and 6), MPC generates about an order of magnitude more features than IDM and OVM. This results in more weights being fit for the MPC feature space in the encoding model as well. This difference in the number of features is part of the different hypotheses that these models provide about visuomotor control during driving. We plan to further explore the role of the number of features in the future.

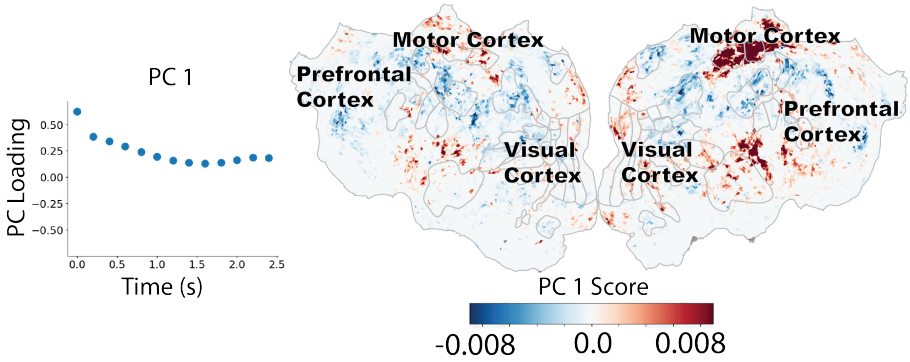


Figure 4: **The importance of the cost feature over time displays regional patterns**. We analyze the weights of the MPC cost feature ($c(x[t], u[t])$ in equation (1)) over the MPC prediction horizon and across the brain. The principal components (PCs) of the weight matrix represent common patterns in these weights while the component scores represent how present that pattern is in each voxel. The left image displays the top PC for the cost feature, which explains 86.5% of the variance and the right image shows its scores. We observe contiguous regions of blue and red, suggesting different regions have different relationships to immediate versus distant predictions.

6. Conclusion

In this work we provided a framework for comparing human driver models to the algorithms used by human drivers. By integrating this framework into the human driver model design process, we hope to facilitate the search for better models, which could in turn lead to safer behavior in autonomous driving systems. Additionally, we hope this framework can provide insights to those aiming to understand how the brain engages with complex tasks. There are several directions for future work. First, we will apply our framework to additional subjects and driver models, especially data-driven and strategic models. Second, we will continue exploring new ways to analyze how the brain can aid in understanding driver models and vice versa. For example, figure 4 demonstrates the potential for analysis of the regression weights to examine how the brain may represent predictions. Lastly, we hypothesize that human driver models that represent information more similarly to the brain will have better generalization performance. Future work will be necessary to validate this hypothesis.

Acknowledgments

We'd like to thank the friends and colleagues who supported our exploration of this work and provided fruitful discussion and feedback. This includes but is not limited to Anand Siththaranjan, Francisco Utrera, and Serena Tang.

References

Masako Bando, Katsuya Hasebe, Akihiro Nakayama, Akihiro Shibata, and Yuki Sugiyama. Dynamical model of traffic congestion and numerical simulation. *Physical review E*, 51(2):1035, 1995.

Masako Bando, Katsuya Hasebe, Ken Nakanishi, and Akihiro Nakayama. Analysis of optimal velocity model with explicit delay. *Physical Review E*, 58(5):5429, 1998.

George A Bekey, Gerald O Burnham, and Jinbom Seo. Control theoretic models of human drivers in car following. *Human Factors*, 19(4):399–413, 1977.

Kyle Brown, Katherine Driggs-Campbell, and Mykel J Kochenderfer. A taxonomy and review of algorithms for modeling and predicting human driver behavior. *arXiv preprint arXiv:2006.08832*, 2020.

G Burnham, Jinbom Seo, and G Bekey. Identification of human driver models in car following. *IEEE transactions on Automatic Control*, 19(6):911–915, 1974.

Dian Chen and Philipp Krähenbühl. Learning from all vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17222–17231, 2022.

Douglas P Crowne. The frontal eye field and attention. *Psychological Bulletin*, 93(2):232, 1983.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

Tom Dupré la Tour, Michael Eickenberg, Anwar O Nunez-Elizalde, and Jack L Gallant. Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728, 2022.

Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

Christian Gerloff, Brian Corwell, Robert Chen, Mark Hallett, and Leonardo G Cohen. Stimulation over the human supplementary motor area interferes with the organization of future elements in complex motor sequences. *Brain: a journal of neurology*, 120(9):1587–1602, 1997.

Christian Grefkes and Gereon R Fink. The functional organization of the intraparietal sulcus in humans and monkeys. *Journal of anatomy*, 207(1):3–17, 2005.

Christian Grefkes, Afra Ritzl, Karl Zilles, and Gereon R Fink. Human medial intraparietal cortex subserves visuomotor coordinate transformation. *Neuroimage*, 23(4):1494–1506, 2004.

Joachim Gross. Magnetoencephalography in cognitive neuroscience: a primer. *Neuron*, 104(2): 189–204, 2019.

HY Guo, Yan Ji, Ting Qu, and Hong Chen. Understanding and modeling the human driver behavior based on mpc. *IFAC Proceedings Volumes*, 46(21):133–138, 2013.

Longxiang Guo and Yunyi Jia. Modeling, learning and prediction of longitudinal behaviors of human-driven vehicles by incorporating internal human decisionmaking process using inverse model predictive control. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5278–5283. IEEE, 2019.

Milad Haghani, Michiel CJ Bliemer, Bilal Farooq, Inhi Kim, Zhibin Li, Cheol Oh, Zahra Shahhoseini, and Hamish MacDougall. Applications of brain imaging methods in driving behaviour research. *Accident Analysis & Prevention*, 154:106093, 2021.

Eiji Hoshi and Jun Tanji. Differential roles of neuronal activity in the supplementary and presupplementary motor areas: from information retrieval to motor planning and execution. *Journal of neurophysiology*, 92(6):3482–3499, 2004.

John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. In *Conference on Robot Learning*, pages 409–418. PMLR, 2021.

Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532 (7600):453–458, 2016.

Yue Kang, Hang Yin, and Christian Berger. Test your self-driving algorithm: An overview of publicly available driving datasets and virtual testing environments. *IEEE Transactions on Intelligent Vehicles*, 4(2):171–185, 2019.

Keuntaek Lee, David Isele, Evangelos A Theodorou, and Sangjae Bae. Spatiotemporal costmap inference for mpc via deep inverse reinforcement learning. *IEEE Robotics and Automation Letters*, 7(2):3194–3201, 2022.

Florin Leon and Marius Gavrilescu. A review of tracking and trajectory prediction methods for autonomous driving. *Mathematics*, 9(6):660, 2021.

Sajjad Mozaffari, Omar Y Al-Jarrah, Mehrdad Dianati, Paul Jennings, and Alexandros Mouzakitis. Deep learning-based vehicle behavior prediction for autonomous driving applications: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):33–47, 2020.

Parashkev Nachev, Christopher Kennard, and Masud Husain. Functional role of the supplementary and pre-supplementary motor areas. *Nature Reviews Neuroscience*, 9(11):856–869, 2008.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Jordan Navarro, Emanuelle Reynaud, and François Osiurak. Neuroergonomics of car driving: A critical meta-analysis of neuroimaging data on the human brain behind the wheel. *Neuroscience & Biobehavioral Reviews*, 95:464–479, 2018.

Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011.

Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage*, 197:482–492, 2019.

Seiji Ogawa, Tso-Ming Lee, Alan R Kay, and David W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences*, 87(24):9868–9872, 1990.

Tomáš Paus. Location and function of the human frontal eye-field: a selective review. *Neuropsychologia*, 34(6):475–483, 1996.

John W Pratt. Dividing the indivisible: Using simple symmetry to partition variance explained. In *Proceedings of the second international Tampere conference in statistics, 1987*, pages 245–260. Department of Mathematical Sciences, University of Tampere, 1987.

Günther Prokop. Modeling human vehicle driving by model predictive online optimization. *Vehicle system dynamics*, 35(1):19–53, 2001.

Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8):895–935, 2020.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 683–700. Springer, 2020.

Tom A Schweizer, Karen Kan, Yuwen Hung, Fred Tam, Gary Naglie, and Simon J Graham. Brain activity during driving with distraction: an immersive fmri study. *Frontiers in human neuroscience*, 7:53, 2013.

Hugo J Spiers and Eleanor A Maguire. Neural substrates of driving behaviour. *Neuroimage*, 36(1): 245–255, 2007.

Ganesh Sundaram, Sai Krishna Chada, and Daniel Görges. A novel approach to classify and replicate human drivers using model predictive control. In *International Commercial Vehicle Technology Symposium*, pages 36–51. Springer, 2022.

Jun Tanji and Keisetsu Shima. Role for supplementary motor area cells in planning several movements ahead. *Nature*, 371(6496):413–416, 1994.

Martin Treiber, Ansgar Hennecke, and Dirk Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical review E*, 62(2):1805, 2000.

Jiawei Xu, Sicheng Pan, Poly ZH Sun, Seop Hyeong Park, and Kun Guo. Human-factors-in-driving-loop: Driver identification and verification via a deep learning approach using psychological behavioral data. *IEEE Transactions on Intelligent Transportation Systems*, 24(3):3383–3394, 2022.

Tianjiao Zhang. *Modelling Navigation Representations During Naturalistic Driving*. PhD thesis, University of California, Berkeley, 2021.