MIRROR: Differentiable Deep Social Projection for Assistive Human-Robot Communication (Supplementary Material)

VII. Appendix

This section contains supplementary material, specifically details on the experiment domains, the simulated human agents, and experimental results relating to the learnt implants. We also provide additional details on the human-subject experiment.

A. Domains for Simulated Human Experiment

Our experiments made use of three Gridworld domains, each with two communication/observation modalities (see Fig. 5 in the main paper and Table I below).

Gridworld Driving In this task, a human drives a car at a constant speed along a road with two lanes. The driver has two possible actions, to stay in his lane or switch lanes. There are four cars around the driver that may speed up or slow down. These cars will not avoid the human's car. The position of the cars are randomly initialized at the beginning of each episode. We examined 8 different scenarios in total with varying car movements. While simple, the state space of this domain is larger than 10^5.

In addition to observing their lane, the human driver can receive two types (modalities) of discrete observations (i) the relative position of the other cars and (ii) specific "speech" symbols that describe the position of the other cars as well as the speed of the other cars. When driving in clear weather, all cars are visible. However, in dense fog, the driver can only see two units ahead and cannot see cars in the rear.

Search & Rescue In this task, a firefighter has to rescue a victim from a burning house; the firefighter has to find the victim's position and bring them back to the entrance. The victim may appear at one of the three potential positions. There are two corridors to the victim's potential positions, but one of them may be blocked by an obstacle. At the outset, the firefighter is unaware of the positions of victim and the obstacle. We examined 9 different scenarios in total with different situations of corridors and the victim's position.

The firefighter can receive two kinds of observations in addition to his own position: (i) a raw image of the map and (ii) specific speech utterances that describe the corridors (blocked/unblocked) and victim's potential position (victim is/isn't there). In the original training (clear) setting, all information are visible. In the transfer setting (dense smoke), the firefighter can only see 1.5 units around himself and does not observe the speech utterances.

Bomb Defusal In this task, a human tele-operates a robot arm to defuse a bomb. To defuse the bomb, the robot needs to press 3 buttons (a button at each stage). Which button to press at each stage depends on the bomb type and the "terminals" alongside the bomb.

The human can receive two kinds of observations: (i) raw images of specific terminals and (ii) specific speech utterances that indicate the type of the bomb. The robot knowledge of the rules is out-of-date, i.e., its policy is wrong. As such, the robot cannot defuse the bomb by itself. The human knows the updated rules but is slower than the robot at identifying the correct terminals and is unable to identify the bomb type. To help human quickly defuse the bomb, the robot advises the human which button to press and provides explanations (images of specific terminals and the bomb type). The human then chooses a button to press. Note that the human is unable to complete the task on their own since they cannot perceive the bomb type.

TABLE I: Communication Modalities and Costs.

Domain	Visual	Cost	Verbal	Cost
Driving	Relative positions of other cars in front of the agent.	$0.01 \cdot (\text{Num. cars shown})^2$	Symbols representing position and speed of other vehicles.	$0.03 \cdot (\text{Num. utterances})^2$
Search-&-Rescue	Raw image pixels revealing a location in the map. Each map is divided into 9 locations.	$0.02 \cdot (\text{Num. locations shown})^2$	GPT-2 embedding of speech utterances indicating the position of the victim and obstacles, e.g., "*victim in top-right*".	$0.1 \cdot (\text{Num. utterances})^2$
Bomb-Defusal	Raw image pixels of specific terminal.	$0.3 \cdot (\text{Num. terminals shown})^2$	GPT-2 embedding of speech utterances indicating the bomb type, e.g., "*Type A Bomb*".	$0.5 \cdot (\text{Num. utterances})^2$

B. Simulated Humans Agents.

In the following, we describe the simulated agents that were created for our experiments. For each domain, we collected data from 10 participants and trained a simulated agent for each participant. The simulated agents are able to perceive state information, up to their perceptual limitations. Qualitatively, we find the behavior of the simulated agents to be very similar to their human counterparts.

Gridworld Driving Each participant played 24 rounds in each setting (clear weather and dense fog). We trained a reward function on all the collected human data using Maximum Entropy Inverse Reinforcement Learning [44]. Similar to previous work [18], we use as features (i) the distance to the center of the road, (ii) distances to the other four cars and (iii) driver's action. Given the learnt reward function, the simulated human agent plans actions at each time-step using CE. Each simulated agent plays 40 rounds (clear weather and dense fog). The last 20 rounds of each simulated agent's data were used as validation set (10 rounds) and testing set (6 rounds). We assumed that the simulated human agent can digest all communicated information and never forgets.

Search & Rescue Each participant played 36 rounds in each setting (clear and smoke). The simulated agent was a planning agent that computes and executes the shortest path between subgoal positions. The subgoals were manually specified, and the transitions probabilities between subgoals were learnt using the collected data. In the smoke setting, if the simulated agent does not know the position of victim, he will search all victim's potential positions until the victim is found. Similar to Gridworld Driving, each simulated agent plays 40 rounds (clear and smoke). The last 20 rounds of each simulated agent's data were used as validation set (10 rounds) and testing set (6 rounds). As before, we assumed that the simulated human agent can digest all communicated information and never forgets.

Bomb Defusal Each participant played 36 rounds. For the human participants, we informed them of the type of bomb to ease the number of samples that needed to be collected. To determine which button to press, the simulated human has to guess the type of bomb and identify one correct terminal out of the six displayed. The type of the bomb is not visible to the simulated human and hence, the chance of guessing the correct type is 0.5. To model the length of time the human takes to find the correct terminal, we use a geometric distribution with success probability p learnt from human data. Once the simulated human finds the relevant terminal, it will always press the correct button. Otherwise, it refrains from pressing any button and continue searching for the correct terminal. Same as before, each simulated agent plays 40 rounds. The last 20 rounds of each simulated agent's data were used as validation set (10 rounds) and testing set (6 rounds). We also assumed that the simulated human agent can digest all communicated information and never forgets.

C. MIRROR *Implants*

The implants used in our experiments are similar to those described in III-B.

Perceptual Implants. In gridworld driving, the perceptual implant is a threshold filter governed by 4 parameters. Each parameter specifies the distance a human can see along a specific direction (front/back and left/right lane) on the road. If a car on a lane is within the specified threshold distance, the model can observe the position of the car. For Search & Rescue, the perceptual implant is a single parameter threshold filter; we split the image map into 9 portions and if the center of a portion is within this range, it is observable. In bomb defusal, the perceptual implant are 6 parameters — each models the probability that a human perceives 1 of 6 terminals within 1 second (1 time step). Given these 6 probabilities, we sample a 6 dimensional vector with the value of each dimension to be 0 or 1. If the value of a dimension is 1, we show the corresponding terminal in the image, if not, we mask it out.

Policy Implants. For all three domains, the policy distributions are modeled as categorical distributions $\mathrm{Cat}(K, \mathbf{p})$. The policy implant was a small neural network that takes in the latent state z_t^{H} as input and outputs a residual term $\delta(z_t^{\mathrm{H}})$ that was added to $\mathbf{p}$ of original policy distribution. The resultant action distribution is then parameterized by $\hat{\mathbf{p}} = \sigma(\mathbf{p} + \delta(z_t^{\mathrm{H}}))$.

Learnt perceptual implants. Samples of the learnt perceptual implants by MIRROR are shown in Fig. 11, Fig. 12 and Table. II). Both in Gridworld Driving and Search & Rescue, the learnt perceptual implants indicate that the human was able to see in the original setting and was only able to see a small area nearby in transfer setting (e.g., fog). For the Bomb Defusal task, we see the average success rate was approx 0.05.

TABLE II: 5 samples of learnt perceptual implants in the bomb defusal game.

Terminals	1	2	3	4	5	6
Sample 1	0.054	0.051	0.045	0.059	0.185	0.05
Sample 2	0.054	0.071	0.044	0.068	0.085	0.068
Sample 3	0.045	0.033	0.035	0.045	0.206	0.043
Sample 4	0.037	0.034	0.026	0.047	0.012	0.029
Sample 5	0.031	0.018	0.026	0.029	0.019	0.037

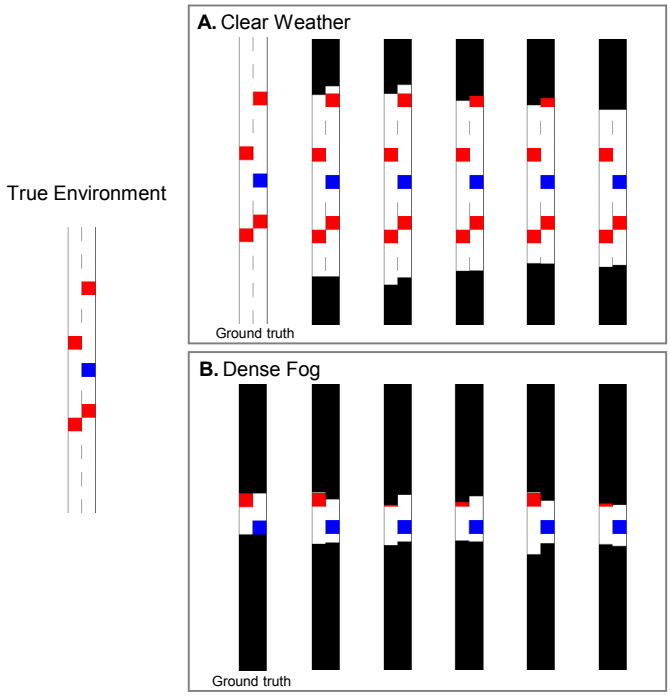

Fig. 11: Samples of learnt perceptual implants for Gridworld Driving. The blue block represents the ego car and red blocks represent other cars. The black areas represent the region that the human cannot see. Ground truth images represent what human actually sees in the original (clear weather) and transfer (dense fog) settings. (**A**) The learnt implants indicate that the human was able to see the most of road in clear weather. (**B**) The implants show the human was only able to see a small nearby area in dense fog.

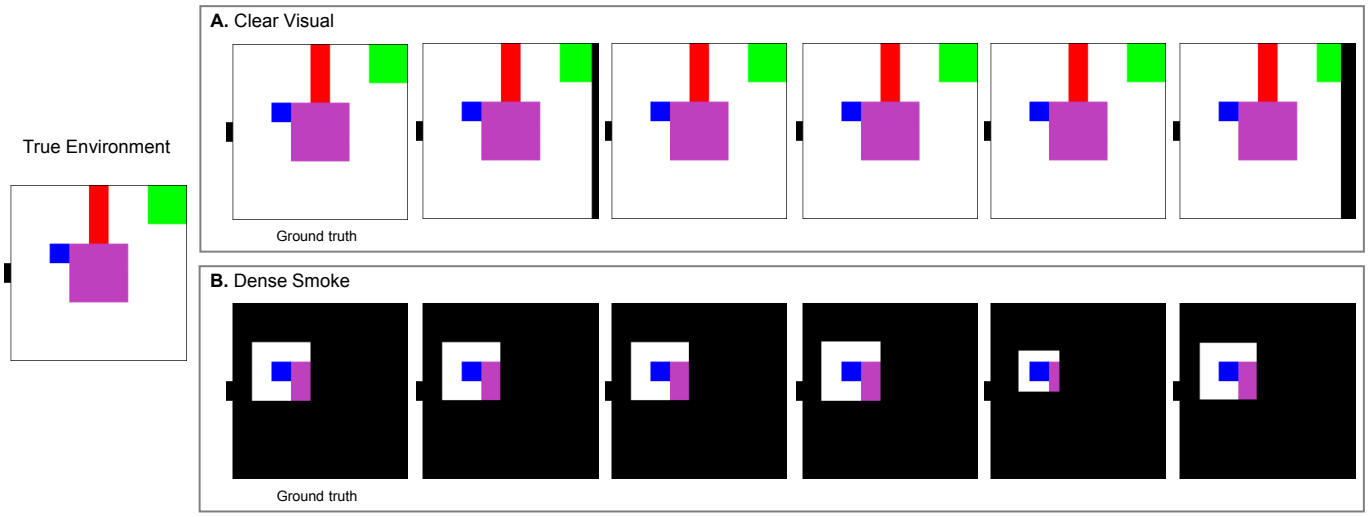

Fig. 12: Samples of learnt perceptual implants in the Search & Rescue task. The red and purple blocks represent obstacles and the green box represents the victim. The black areas represent the region that the human can not see. The implants are similar to the ground truth images that represent what human actually sees in the original (clear weather) and transfer (dense fog) settings. (**A**) For the clear setting, the learnt implants indicate the human was able to see the most of map. (**B**) In the dense smoke setting, the implants show the human was only able to see a small surrounding region. The implants are similar to the ground truth images that represent what human actually sees in the original (clear weather) and transfer (dense fog) settings.

D. Model Architectures and Training

In all three domains, we approximate the posterior $p(z_{1:T}|x_{1:T}^{1:M}, a_{1:T-1}) = \prod_{t=1}^{T} q_\phi(z_t|z_{t-1}, x_t^{1:M}, a_{t-1})p(z_0)$. Below, we give an overview of our models; source code implementing these models is available at http://blinded-for-review.

Gridworld Driving. Recall that in this domain, state (e.g., position of cars) information is available to the agent. The MIRROR model consists of the following main components:

- The transition distribution $p_\theta(z_t|z_{t-1}, a_{t-1})$ and variational distribution $q_\phi(z_t|z_{t-1}, x_t^{\text{visual,verbal}}, a_{t-1})$ are both 48-dimensional Gaussian distribution with diagonal covariance. We set the transition network to be 3 FC layers deep with hidden size 32. The network parameterizing $q_\phi(z_t|z_{t-1}, x_t^{\text{visual,verbal}}, a_{t-1})$ was a MLP with 3 fully-connected (FC) layers of hidden size 64.
- The observation (decoder) distributions $p_\theta(x_t^{\text{visual}}|z_t)$, $p_\theta(x_t^{\text{verbal}}|z_t)$ and reward $p_\theta(r_t|z_t)$ are Gaussian distributions with a diagonal covariance. Each decoder is 3 FC layers deep with hidden size 32.
- The Q function is modeled by a Q-network, which takes z_t, a_t as input and outputs the Q value for the particular a_t. The Q-networks consist of 3 FC layers with hidden size 128.

The learning rate was set to 0.0003. Before each experience collection phase in Deep Q-Learning, we probabilistically determine whether to use the expert or the model policy via the hyperparameter β. In our experiments, β was initialized to 0.5 with a decay factor of 0.8. Furthermore, whenever the policy is chosen to perform rollout, we perform ϵ-greedy action selection feature to allow for further randomization; ϵ was initialized to 0.5 with a decay of 0.75 after every rollout phase.

For BC, the policy distribution is a categorical distribution $\text{Cat}(K, \mathbf{p})$. The policy network is modeled as a 48-dimensional GRU followed by one FC layer of size 32; At each time step, the GRU first takes x_t^{visual}, x_t^{verbal}, a_t and hidden state h_{t-1} from the previous time step as input and outputs a new hidden state h_t, which is feed into a fully-connected layer to produce the parameter $\mathbf{p}$ of the categorical policy distribution $\text{Cat}(K, \mathbf{p})$. For SQIL, the Q function is modeled as a combination of a 48-dimensional GRU followed by a 3 FC layers deep with hidden size 32; the GRU first takes x_t^{visual}, x_t^{verbal}, a_t and hidden state h_{t-1} from the previous time step as input and outputs a new hidden state h_t, which is fed into 3 FC layers to produce the Q value.

Search and Rescue The distributions and neural networks in the MIRROR model are similar to the Gridworld driving environment. Accommodating differences in the input ("raw" image and text) led to differences in:

- The encoder $q_\phi(z_t|z_{t-1}, x_t^{\text{visual,verbal}}, a_{t-1})$, which uses convolutional layers for the raw image input, and feedforward layers for the symbolic speech observations. Due to computational limits, we pre-trained an autoencoder to reduce the dimensionality of the speech observations (768 dimensional GPT-2 word vector). We set the encoder to have 1 convolutional layer with both kernel size and stride size 3 and outputs 4 channels followed by 3 FC layers with 48 dimensions, while the feedforward portion for the symbolic speech input contains just 1 layer. Each network (Conv for image and FC for speech) produces a 16-dimensional vector, which are concatenated and passed through 3 FC layers (hidden size 128) to derive the state parameters for z_t.
- The image decoder, which consists of a deconvolution layer with the same kernel size and stride as the convolution layer in the encoder. The symbolic speech decoder has 3 FC layers with hidden size 64.

The Q-network was a MLP with three FC layers (hidden size 200). During the training process, β was initialized to 1.0 with a decay factor of 0.98 while ϵ was initialized to 0.5 with a decay factor of 0.9.

The BC and SQIL models are also similar to the Gridworld driving domain, except for the observation input networks. We used the same networks as the MIRROR model but the concatenated features are fed into a GRU instead of FC layers.

Bomb Defusal The distribution and model setups were very similar to the above domains; there were only minor differences in the convolutional layers (stride size 2) and FC layers (64 dimensions). The Q network was a larger MLP with 3 FC layers (2048 neurons each). During the training process, β was initialized to 1.0 with a decay factor of 0.98 while ϵ was initialized to 0.5 with a decay factor of 0.9. Likewise, the BC and SQIL models were similar to other domains above, but the networks were larger (three layers as before, but with 2048 neurons in each layer).

CARLA The MIRROR model used higher capacity representations/networks; both $q_\phi(z_t|z_{t-1}, x_t^{lidar,verbal}, a_{t-1})$ and $p_\theta(z_t|z_{t-1}, a_{t-1})$ were 128-dimensional Gaussian distributions with diagonal covariance. Similarly, the observation distributions ($p_\theta(x_t^{\text{lidar}}|z_t)$, $p_\theta(x_t^{\text{verbal}}|z_t)$), reward $p_\theta(r_t|z_t)$ and policy distributions were Gaussian distributions with diagonal covariance. The above distributions and the Q-function were modeled using neural networks with 3 fully-connected layers of 128 neurons each. During the training process, β was initialized to 1.0 with a decay factor of 0.995 while ϵ was initialized to 0.5 with a decay factor of 0.9.

For BC, the policy distribution is a Gaussian distribution with a diagonal covariate matrix. The policy network is modeled as a GRU with hidden size 128 followed by a 3 FC layers with hidden size 128; The GRU first takes $x_t^{\text{lidar,verbal}}$, a_{t-1} and the hidden state h_{t-1} from the previous time step as input and outputs a hidden state h_t, which is feed into the 3 FC layers to produce the mean and variance of the Gaussian policy distribution.

E. Task Reward

In the following, we describe the task reward functions used to train our agents and to model the simulated humans.

Gridworld Driving If the distance of a car and the ego car is within 2 units, the agent will receive a -2 penalty. If the ego car collides with other cars or goes off the road, the agent will receive a -5 penalty. If the ego car changes the lane, the agent will receive a -1 penalty.

Search and Rescue During each episode, if the human agent finds the victim, they will receive a $+1$ reward. After the agent finds the victim, if they return to the entrance, they will receive another $+15$ reward. Colliding with obstacles incurs a -10 penalty. At each time step, the agent will receive a 0.1 time penalty.

Bomb Defusal During each stage, if the human agent press a correct button, they will receive a $+5$ reward. If the agent press a wrong button, they will receive a -5 penalty. Each time step incurs a -1 cost.

CARLA The reward function comprises several components:

- Speed reward r^{speed}, which encourages the agent to drive as fast as possible without exceeding $v_{\max}$ (set to 40 km/h in the experiments).

$$r^{\text{speed}} = \begin{cases} \dfrac{v_{\text{current}}}{v_{\max}}, & v_{\text{current}} \leq v_{\max} \\ -(v_{\text{current}} - v_{\max}), & \text{otherwise} \end{cases} \tag{7}$$

- Braking and steering penalties r^{brake}, r^{steer} to promote smooth driving.
- Lane change reward r^{change}, which provides a negative reward of -1 whenever the agent changes a lane, and lane center r^{center} reward which penalizes off-center driving using the normalized value of $-\dfrac{\text{distance to the lane center}}{0.5 \times \text{ lane width}}$.
- Proximity reward $r^{\text{proximity}}$, which is separated into front and back proximity (penalty of -2 whenever a car is detected within 20 meters by the front and back facing LIDAR beams), and the immediate surrounding proximity (penalty of -4 whenever a car is detected within 1.6 meters by any LIDAR beams)
- Road shoulder penalty r^{road} of -4 whenever the ego car goes into the road shoulder.

F. Communication during Simulated Human Experiment

This section provides quantitative results that MIRROR reveals information more selectively, often focusing on the relevant information (e.g. the location of the goal item and obstacle in the Search & Rescue task). We report the percentage of times that relevant information was communicated by the agents (Fig.13) (relative to the total number of times communication occurred). We define relevant information as:

- **Gridworld Driving.** information about cars that were in a collision course with the ego car.
- **Search & Rescue.** The location of the goal item and the obstacle.
- **Bomb Defusal.** The correct terminals and the bomb type.

Fig.13 shows that MIRROR communicate higher percent of relevant information compared to BC and SQIL. As the proportion of training data increases, the percentage of relevant communication also increases.

G. Human-Subject Experiments

To supplement the main results in the paper, this section provides the specific survey questions (Table III) and an analysis of the cognitive workload experienced by the participants.

We focus on the responses collected using NASA TLX [51]. Fig. 14.B. shows the user ratings for the individual subscales, which supports the notion that MIRROR lessened cognitive load compared to BC. Differences between the communication agents were statistically significant at the $\alpha = 0.05$ level across the subscales except for Physical Demand and Temporal Demand; this was likely because the task (driving along a relatively straight highway) was not physically and temporally demanding in nature. Pairwise t-tests (adjusted-$\alpha = 0.0167$) indicate differences between MIRROR v.s. BC and MIRROR-KL v.s. BC to be statistically significant except for Performance (MIRROR-KL vs BC: $p = 0.017$).

The overall Raw-TLX scores (Fig. 14.C.) show that the participants felt less mentally burdened when they interacted with MIRROR ($F_{3,60} = 10.071$, $p < 0.001$; MIRROR vs BC: $t(9.5) = 6.658$, $p < 0.001$; MIRROR-KL vs BC: $t(9.5) = 5.580$, $p < 0.001$). NC was always done first (to collect data for training the models), so we cannot completely ignore ordering effects. That said, participants only started the task after sufficient practice. The differences in scores between the NC and MIRROR conditions are large, which suggests that MIRROR is effective at reducing cognitive load via assistive communication.

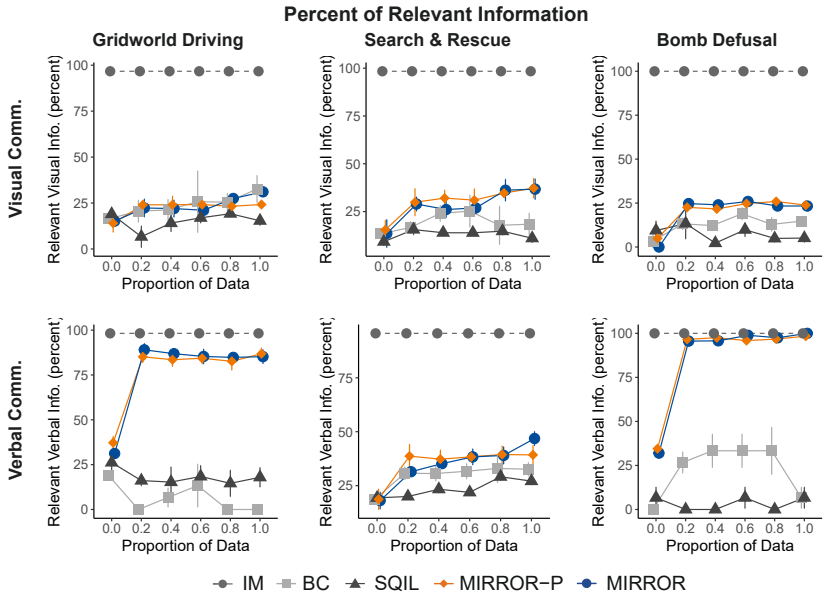

Fig. 13: The percent of relevant information during communication in the Transfer/Test Environment. NC indicates agents that received no communication. In the Gridworld Driving tasks, the BC agents can achieve similar percent of relevant information compared to MIRROR, but BC gives much less relevant information in the verbal communication channel. In the Search & Rescue and Bomb Defusal tasks, the simulated agents that were paired with the assistive MIRROR revealed information more selectively compared to BC and SQIL.

TABLE III: Subjective measures in the human experiment.

	Subjective Measures
After each Method (Including NC)	**Cognitive Load** *(7-point Likert Scale)* - NASA Task Load Index (NASA-TLX) [51]
After each Method (Excluding NC)	**What, When, How of Human-Robot Communication** *(7-point Likert Scale)* - The assistive driving agent's communication was helpful in accomplishing the task. - The assistive driving agent's communication was redundant. - The assistive driving agent's communication was timely. - I feel comfortable with the mode of communication (visual and/or speech) selected by the assistive driving agent at different scenarios. **Human-Robot Trust** *(7-point Likert Scale)* - I trust the assistive driving agent to provide useful communication on this task.
After Interaction	**Perceived Relative Method Preferences** *(Short Distance vs Long Distance)* - Which agent were you the most comfortable with? - Imagine that you have to drive for long distances of at least one hour under the same weather conditions and traffic, which agent would you be the most comfortable with?

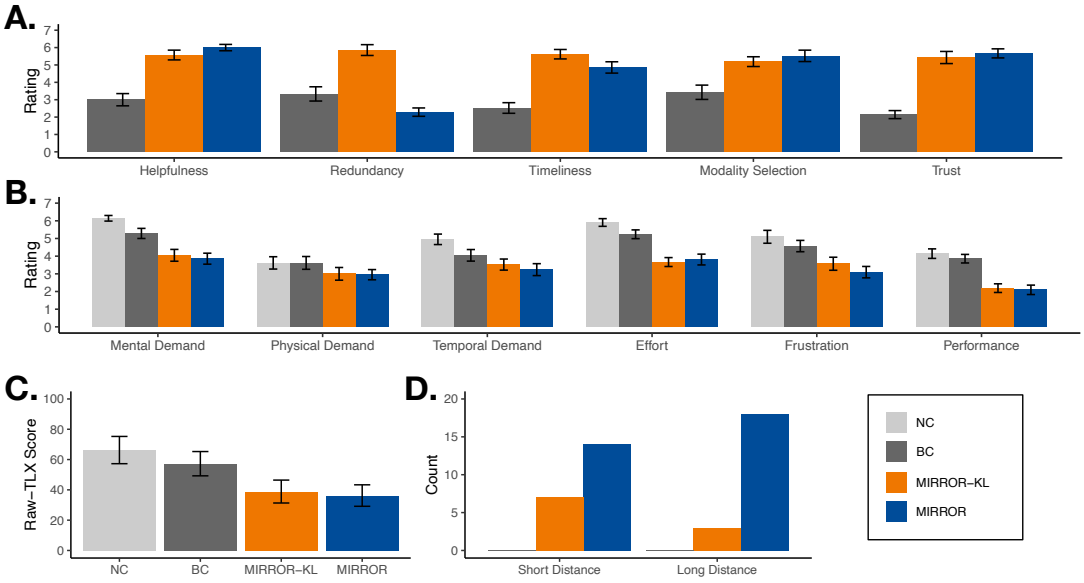

Fig. 14: Subjective Measures in the CARLA Driving Experiments across conditions. Error bars indicate one standard error. (**A**) Human-Robot Communication and Trust; (**B**) Individual NASA-TLX ratings; (**C**) Raw-TLX score; (**D**) Counts of which agent was preferred for short and long distance driving. Overall, MIRROR was perceived more positively than BC. Please see the main text for additional details.