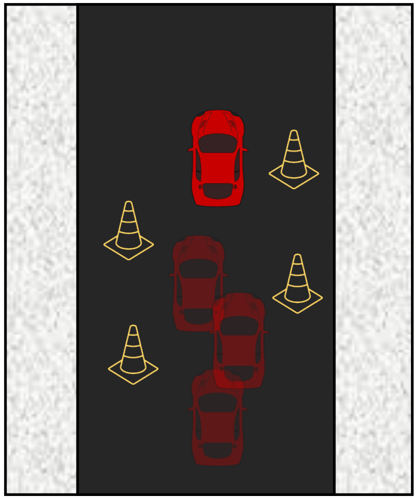
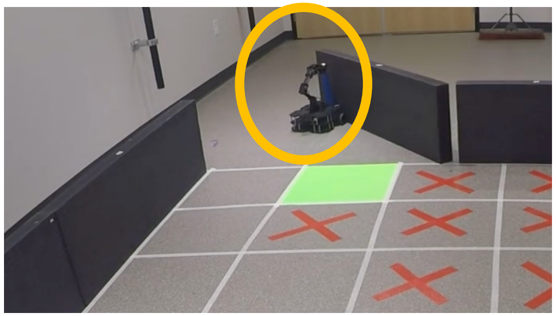
No Bad Surprises: Aligning Agent Behavior with Human Norms via Specification Refinements

PIs: Sandhya Saisubramanian (Oregon State University), Houssam Abbas (Oregon State University), Joydeep Biswas (UT Austin), Shlomo Zilberstein (UMass Amherst)

Bad Surprises

- Autonomous robots can fail in surprising ways, like a self-driving car swerving dangerously to avoid traffic cones.
- The interaction between perception and action is particularly prone to such bad surprises.





Project Objectives

A formal design pipeline for autonomous agents, aimed at reducing surprises that result from the misalignment between the agent's (social and ethical) norms, and those that the human designers expect it to have.

1. Develop a way for RL agents to learn an incentive structure subject to *normative (deontic) constraints*.
2. Uncover unknown unknowns by generating norms that would surprise the engineer
3. Enable the agent to reason for itself about what it should *know* given what it should *do*, and to explain its actions in high-level logic.
4. Develop runtime monitors to detect and predict safety violations, and a "meta-reasoner" to reason about multiple objectives to restore safety
5. Industrial collaboration with Agility Robotics and Toyota for better technology transfer
6. Undergraduate student mentoring and curriculum development

Team: Alena Makarova, Houssam Abbas

Research Progress #1

- Bad surprises often arise from the violation of implicit norms that govern the operating context.
- In year 1, we aim at explicitly embedding such norms in the training stages of the agent, to reduce the likelihood of bad surprises at runtime.

Current Progress:

- The norms are expressed in variants of probabilistic deontic logic and *constrain* the utility-maximizing objectives of the agent.
- Expected Results: A process to design an RL agent guaranteed to meet explicitly given norms while accomplishing an independently specified mission

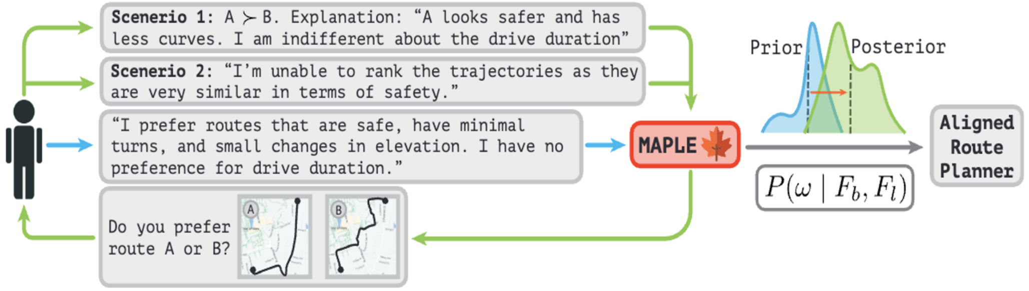
Team: Saaduddin Mahmud, Mason Nakamura, Shlomo Zilberstein

Research Progress #2 (AAAI 2025)

- Aligning systems to specific user preferences at test-time requires high sample efficiency to minimize user interaction and a natural way for communicating preferences to the system.

Active learning + LLMs = MAPLE

- Incorporates linguistic feedback as explanations in addition to binary preferences in a Bayesian preference learning framework
- Uses Oracle-Guided Active Query Selection to select easy-to-answer queries for the user while minimizing uncertainty
- Introduces abstract linguistic concepts to facilitate generalizability to new domains and interpretable preference mechanisms



Scenerio 1: A ≻ B. Explanation: "A looks safer and has less curves. I am indifferent about the drive duration"
Scenerio 2: "I'm unable to rank the trajectories as they are very similar in terms of safety."
"I prefer routes that are safe, have minimal turns, and small changes in elevation. I have no preference for drive duration."
MAPLE
Prior
Posterior
$P(\omega \mid F_b, F_l)$
Aligned Route Planner
Do you prefer route A or B?
A
B

Figure 1: Application of MAPLE to the Natural Language Vehicle Routing Task.

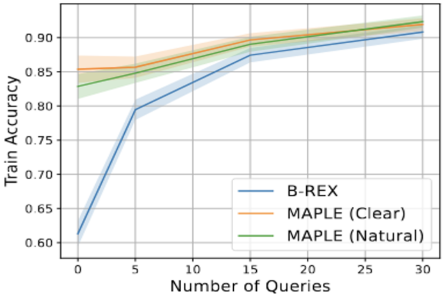


Figure 4: (Test accuracy (OSM Routing)

Improves sample efficiency; higher similarity to ground truth preferences

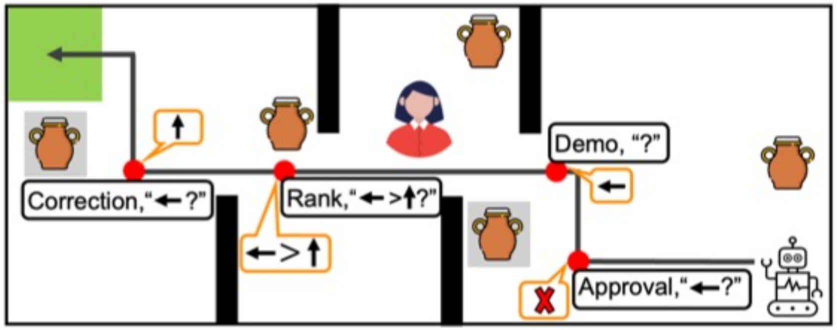
Team: Yashwanthi Anand, Sandhya Saisubramanian

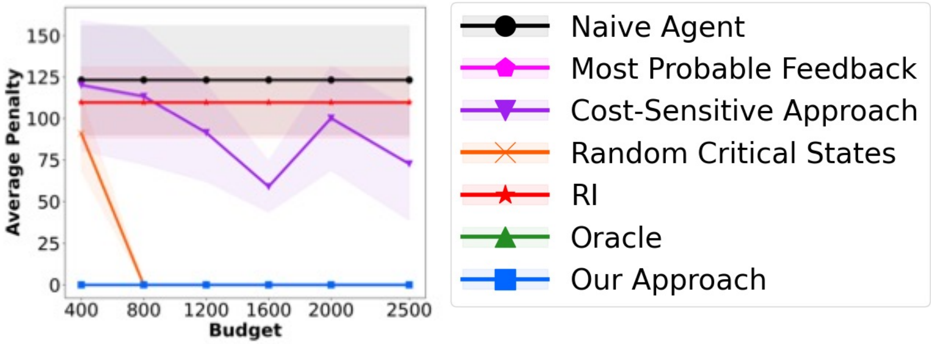
Research Progress #3 (Under review)

How to balance the trade-off between querying a human to accelerate learning and reducing human effort involved?

Information-theoretic approach to decide when to query and in what format, accounting for user preferences

- Identify critical states where gathering data maximizes information gain for the agent
- Identify the most informative feedback in critical states, based on past interactions and using human preference model
- Update learned model of desirable behavior based on gathered data





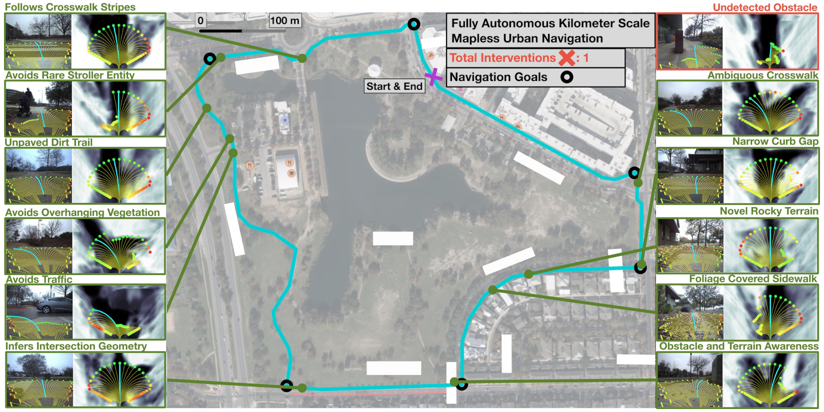
Improves sample efficiency; fewer “bad surprises”

Team: Arthur Zhang, Harshit Sikchi, Amy Zhang, Joydeep Biswas

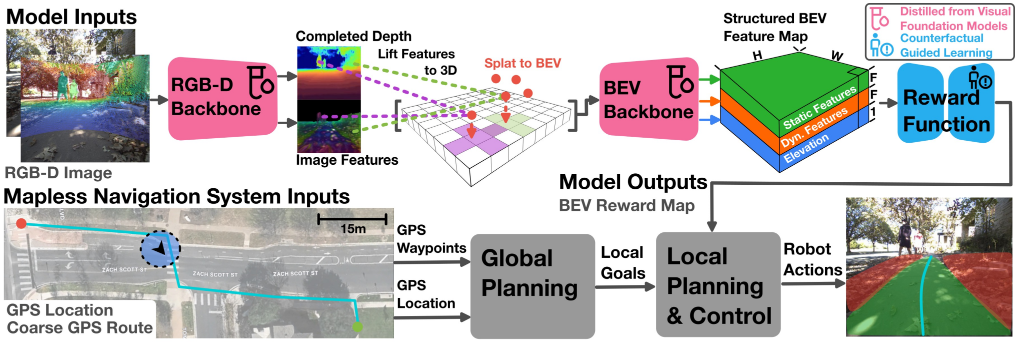
Research Progress #4

Learn a generalizable representation that addresses the full mapless navigation problem in urban environments, robust to perceptual aliasing and aligned with human preferences.

- A new architecture and learning framework for distilling navigation factors from visual foundation models trained on internet-scale data.
- A principled inverse reinforcement learning (IRL) framework for aligning navigation behavior with human preferences using counterfactual demonstrations.


Follows Crosswalk Stripes
Avoids Rare Stroller Entity
Unpaved Dirt Trail
Avoids Overhanging Vegetation
Avoids Traffic
Infers Intersection Geometry
0
100 m
Start & End
Fully Autonomous Kilometer Scale Mapless Urban Navigation
Total Interventions ✗: 1
Navigation Goals ○
Undetected Obstacle
Ambiguous Crosswalk
Narrow Curb Gap
Novel Rocky Terrain
Foliage Covered Sidewalk
Obstacle and Terrain Awareness

Satellite image of 2-km long horizon autonomous navigation experiment.



CREStE Mapless Navigation System. Using a RGB-D camera and GPS, our solution predicts bird's eye view (BEV) reward maps that enable robots to navigation safely to user specified goals in urban environments.