Towards Trustworthy Autonomy:

Human-Centered Approaches for Modeling, Decision-Making, and Control

Katherine Driggs-Campbell

It is an exciting and pivotal moment in the history of robotics and transportation, as autonomous vehicles become tangible technologies that will have a huge impact on the human experience. However, the desirable impacts of autonomy are only achievable if the underlying algorithms can handle unique challenges humans in the real world present. Human-dominated fields are prone to peculiar problems, as people tend to defy expected behaviors and do not conform to many of the standard assumptions made in robotics. Moreover, these human agents appear in many different forms, making perception, planning, and control extremely difficult to execute with formal guarantees. To unlock the potential of autonomy and design autonomous systems that we can trust to safely operate among us, we must transform how we think about how intelligent systems interact with, influence, and predict human agents.

We aim to develop rigorous human models that make minimal assumptions, design control schemes that interact and integrate with human agents effectively, and employ high-fidelity simulators and methodologies to ensure realistic interaction in conjunction with large, realistic datasets. In essence, this research agenda combines ideas from robotics and transportation, focusing on nonparametric approximations of human behaviors that account for the combinatorial aspects of decision making and control. The key impacts and contributions of this work are:

1. Developing robust models of human-in-the-loop systems that capture nonparametric distributions over trajectory sets, which can be readily integrated into semi- and fully autonomous control schemes;
2. Considering the impact of autonomy by formalizing control policies for cooperative maneuvers and developing prescriptive tools for optimal communication between humans and autonomy; and
3. Validating models and control schemes via immersive human-in-the-loop testbeds and developing tools to ensure the overall safety of the autonomous system can be guaranteed.

Robust, Informative Models of Highly Uncertain Agents for Control

Assuming that the transition to ubiquitous autonomy will not be instantaneous, we must rigorously model human drivers in a manner that is easily integrated into control. To predict driver trajectories, we take a control theoretic approach and rephrase reachability as an optimization problem to estimate an empirical reachable set for human-in-the-loop systems, as a data-driven alternative to reachable sets. We find the representative subset of likely human actions using a branch-and-bound optimization paradigm, reducing the conservativeness of traditional safety approaches while maintaining accuracy up to a chance constraint [1,2]. This can be integrated into control schemes for semi- and fully autonomous vehicles.

- Semiautonomous Application: Taking results from driver monitoring to detect the driver state, this set prediction method can effectively predict driver errors. This was integrated as a constraint in a model predictive control framework for an optimal shared control scheme that kept the driver safe in a minimally invasive fashion [3].
- Interactive Autonomous Framework: This model was extended to approximate the distribution over possible reactions during cooperative maneuvers, which can be entered as a soft constraint trajectory planning for autonomous vehicles. Resulting trajectories capture the nuanced interactions and improves understanding and collaboration [4].

Optimizing Interaction via Information Constraints

As the role of the driver transitions from controlling the vehicle to monitoring the autonomy's operation, possible unsafe behaviors due to coupling human controlled and autonomous systems arise that must be

mitigated. In the case of transfer of control, we demonstrate a tradeoff between information flow and driver performance. This tradeoff can be used to optimize interaction between humans and autonomy by constraining the information flow through interface design. Using information theory to model this as a noisy channel, we can compute the total entropy of displayed signals to quantify and optimize the informativeness and the conciseness of a user interface, which has a significant impact on developing systems that must smoothly interact with humans [5].

Validation of Human-in-the-Loop Systems

Experimental design and validation is a key component in the development of safe autonomous systems, particularly for human-robot systems where human actions have significant impact on the outcomes. An immersive human-in-the-loop testbed was setup to collect driving datasets, which has been made available for public use in research [6], and validated our human-in-the-loop control frameworks in a realistic fashion [7]. Further, to properly verify safety guarantees of autonomy, we must develop adaptive collision avoidance strategies for rare events that might not be accounted for in the original control design. By extending results from adaptive stress testing, we aim to selectively validate on scenarios that are likely to lead to collision [currently on-going].

Future and On-Going Work

The work presented here provides a brief overview of our efforts in developing safe, interactive autonomy that we can trust to operate in human-dominated fields. There are many extensions to the work, including expanding into other broader applications and implementation on a test vehicle [currently on-going]. Current research thrusts include:

- Assuming that actions are determined by a mode of operation or intent, effective prediction becomes a problem of estimating the underlying control objective. By computing the associated control laws, we treat the resulting trajectories as expert forecasts. Merging results from statistical learning and game theory, this expert advice can dynamically predict driver behavior in complex situations.
- In robotics, the sensing capabilities of any single agent may be limited or occluded and can lead to catastrophic consequences even with the best of algorithms. By treating drivers as complex, noisy sensors and using Bayesian estimation and dynamic potential fields for planning, we can plan for the possibility of occluded hazards, leading to improved perception and safety.
- Given the experimental nature of this work, we hope to develop tools to assess the data validity that are widely applicable for determining if enough samples have been collected for nonparametric models. By expanding tools from statistics and looking at rates of convergence, we would to quantify how representative the data is to its underlying (nonparametric) distribution.

References

[1] K. Driggs-Campbell, R. Dong, and R. Bajcsy. *Robust, Informative Human-in-the-Loop Predictions via Empirical Reachable Sets.* Under Review 2017.

[2] V. Govindarajan, K. Driggs-Campbell, and R. Bajcsy. *Robustness-Utility Tradeoff in Reachability Analysis for Human-in-the-Loop Systems.* Under Review 2017.

[3] K. Driggs-Campbell, V. Shia, and R. Bajcsy. *Improved Driver Modeling for Human-in-the-Loop Vehicular Control*, in International Conference on Robotics and Automation (ICRA). May 2015.

[4] K. Driggs-Campbell, V. Govindarajan, and R. Bajcsy. *Integrating Intuitive Driver Models in Autonomous Planning for Interactive Maneuvers.* To Appear in Transactions on Intelligent Transportation Systems, Oct. 2017.

[5] T. Rezvani, K. Driggs-Campbell, et al., *Towards Trustworthy Automation: User Interfaces that Convey Internal and External Awareness*, in IEEE International Conference on Intelligent Transportation Systems (ITSC), Nov. 2016.

[6] K. Driggs-Campbell and R. Bajcsy. *Identifying Modes of Intent from Driver Behaviors in Dynamic Environments,* in IEEE International Conference on Intelligent Transportation Systems (ITSC), Sept. 2015.

[7] K. Driggs-Campbell, *Experimental Design for Human-in-the-Loop Driving Simulations,* EECS Department, University of California, Berkeley, Technical Report UCB/EECS-2015-59, May 2015.