CPS Medium: Humanizing Autonomous Vehicles

May 4, 2018

CPS: Medium: Humanizing Autonomy

Autonomous cars are not just about the technology. They are about freedom of mobility, and a whole set of experiences that will literally and figuratively move people in new ways. While the promise of self-driving cars is attractive, applying it in a meaningful and coherent way still remains a major challenge. Recent accidents involving autonomous (or semi-autonomous) cars that resulted in fatalities of either the driver, or in one case the pedestrian, have exposed major holes regarding the interaction between the car and its driver/passengers. We know now, that it may not always be possible for the driver to take over the control from the autonomous vehicle (AV) at any stage.Therefore, before it's too late, we need to reevaluate our approach towards autonomy, by seeking answers to questions that put humans at the center stage. As journeys become fully automated, the experience itself will need to become more human.

The goal of the proposed research is to utilize both human and machine advantages to humanize autonomy ♠[1] by instilling the beneficial nuance of human behavior and trust, while exploiting technological and safety benefits of an autonomous vehicle. The proposed research aims to bring human factors such as trust, and emotional behavior into the autonomous loop. ♠[2] ♠[3] ♠[4]Trust is one of the core human needs that the autonomous vehicle must establish and defend if the technology is to be adopted at all. However, the nature of trust in the vehicle or the autonomous system varies from person to person. In this research, we will run experiments on a full-scale driving simulator, and real in real vehicles, to understand perspectives of trust in autonomous vehicles from real people, rather than working on assumptions. Transparency and communication are critical to building trust. To establish user understanding of the system and its capabilities the interface must communicate clearly, and transparently - by revealing what the car sees, what the system is currently doing, what it intends to do in response to environmental conditions and why. Safety is not primarily just a functional consideration, it is also emotional. We propose that the issues of functional and emotional design for autonomous vehicles should be tackled together. For example, emotional down-regulation could be used when passengers might be facing an upsetting or frustrating situation – for instance, a delay in travel. Here the AV could sense the frustration and then down-regulate through voice prompts. When you're jumping in and out of different AVs, a consistent and personal experience will be vital for successful adoption. We're concerned with a person's ability, and even right, to make their own decisions, and come and go as they please. Not about how clever cars are without human drivers.

Intellectual Merit: The contributions of this proposal are as follows:

1. Behavior guided autonomy

   Arsalan add

   Lu add

2. Reachability analysis and control synthesis for safety region estimation

   Nicola to add brief description.

3. Feedback design: We propose building a framework, that provides both the intent of the autonomous car and an explanation for its behavior to the driver/passenger by augmenting existing user interfaces. We propose a scenario-based trust modeling method in which by varying the degrees of contextual feedback provided to the user, we can measure how the human trust in the system varies under different

[1] NB: what does it mean to humanize autonomy? We need to be clear here

[2] NB: we need to clearly state the challenges and questions that we are trying to answer with this proposal.

[3] NB: How do we model trust? How do we increase trust? Would predicting help with trust?

[4] NB: The other big challenge that we plan to address is how can we close the loop with human and autonomy

traffic situations. We then create a 'trust' profile for the driver and the autonomous driving behavior can be molded to conform to the trust profile of the user, but only within the confines of overall safety. We show how local interpretability can be used to explain actions of an autonomous car, the operation of which is very complex and hidden from the driver/passenger. For example the AV's action of stopping suddenly is not enough – the user wants to know why. Without any explanation or context, the user will panic. Our proposed framework can generate such explanations.

Broader Impact: Autonomous control and decision systems are forming the basis for significant pieces of our nation's critical infrastructure. In particular, autonomous vehicles present direct, and urgent safety-critical challenges. If successful, the research outcomes will have the following impacts: (a) be a valuable contribution towards increasing the overall safety of fully autonomous vehicles, which are likely to become ubiquitous in the near future, (b) the underlying frameworks of generating local explanations from sensor data, and safe operation through reachability analysis can help enhance a large scope of autonomy including but not limited to autonomous vehicles, robotics, aircraft autopilots, and automatic surgery equipment, and (c) Leveraging human trust and emotional behavior to help enhance the capabilities of autonomous vehicles and also facilitate the deployment of autonomous vehicles in the real world.

Educational Impact: The PI's will develop curriculum including course lectures and hands-on projects related to autonomous driving. The PIs are very vested in promoting and employing undergraduate researchers. They will continue developing and participating in research programs to involve K-12 students into lab research and inspire their interests in autonomy and human factors based upon the hardware and software developed in the proposed research. The PIs will also actively disseminate the research outcomes through outreach in both academia and automotive industries.

Contents

CPS: Medium: Humanizing Autonomy

1 Introduction

Two fatal self-driving-car accidents in March 2018 (Uber and Tesla) have cast doubt on whether the self-driving car can become the future of personal transport. For reasons unknown, the Tesla accident happened in broad daylight in pretty much perfect driving conditions. Just days before this, a fully automated Uber, equipped with LIDAR(s) completely vehicle failed to detect a pedestrian crossing the road at night. The autonomous vehicles industry, and research community, faces an uphill battle in convincing the public that self-driving cars are safer than human drivers. As the framework for mobility in the United States begins to shift from one of personally-owned, manually-driven vehicles to one of a shared and perhaps partially automated fleet, established driver perceptions about their trust and comfort in various vehicle technologies is critical to our understanding as to how one needs to facilitate behavior change.

Everyone is talking about autonomous driving. From automotive manufacturers, to consumer electronic giants, to software engineers, and academic researchers, driverless cars are at the forefront of everyone's imagination. There were more autonomous driving concepts at CES 2018 than ever before [**?**]. AVs can make a meaningful difference to the world, enabling a new level of mobility, independence, and safety for all. This has been covered in reams of papers and many 1,000s of articles and news stories all over the globe. From questions of technological feasibility to thorny ethical dilemmas, it's been approached from many angles. But there are aspects that haven't yet been fully addressed – what do people want and need from AVs and how best to design for the many autonomous user experiences – what about those human factors? *Understanding why the vehicle is doing what it is doing will be a critical factor in trust and acceptance moving forth.* The industry is preoccupied with the race to make their cars be as smart as possible and more safe than human drivers. What's really important though, above all else, is how this technology can enable greater human mobility and hence improve their autonomy. The auto industry's approach to autonomy is imbalanced – there is too much focus on the discrete technologies that will enable it, with little regard for the powerful human factors involved. As the industry gets profoundly disrupted, we firmly believe that it's not just automotive insiders who have a valuable contribution to make. What's important is for car makers and service providers to embrace this moment to rethink the design process to transform the entire user experience for the betterment, and safety of everyone. With that in mind - the goal of the proposed research is to utilize both human and machine advantages to humanize autonomy by instilling the beneficial nuance of human behavior and trust, while exploiting technological and safety benefits of an autonomous vehicle.

Naturally, one of the most important things to get right in AVs is safety. These vehicles must be incredibly safe, in fact safer, statistically and practically, than manually driven cars on the roads today if they're to be of much use to us. Despite the advances in technology, there is a fundamental fear and distrust of autonomous vehicles. According to a Deloitte study, trust appears to be the biggest roadblock to adoption of self-driving cars in every country surveyed. South Korea ranks the highest with 81% of people expressing safety concerns about fully-autonomous vehicles. China has the lowest figures, with 62%, but that still represents a large majority of consumers. The US falls roughly in the middle, where nearly three-quarters of consumers (74%) believe that fully-autonomous vehicles will not be safe. For the other countries in the study the percentage of consumers who believe fully-autonomous vehicle will not be safe are: Japan at 79%, Germany at 72%, and India at 64% While everyone wants an AVs to be safe, the interpretation of safety, and the nature of trust in the vehicle or the system changes from person to person. Some perceive an AV in the same light as a malfunctioning robot, running amok around town, striking fear into passengers, fellow motorists, and pedestrians. On the other hand, even the sheer amazement experienced by those witnessing a fully

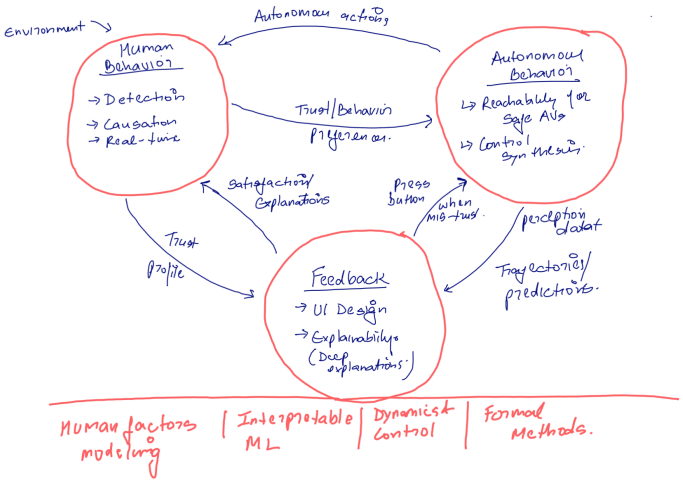

Figure 1: Placeholder overview figure

autonomous car could cause a distraction and lead to accidents. Then there are those who believe and expect that the AV will always "have your back" and is totally safe. *All of this boils down to one thing – trust*. Trust is one of the core human needs that the autonomous vehicle must establish and defend if the technology is to be adopted at all. For the technology to be adopted, it must be trusted, and to do that, the human factors must be taken into consideration.

Add motivation for the emotional behavior part

Beyond getting in and enjoying the ride – there are far more factors, details and nuances to be considered. What are the human factors at play here and how can we design the best end-to-end connected experience?

2 Research Description

2.1 Research Thrust 1: Behavior modeling

One of the most compelling benefits of emotion-aware vehicles is the ability to monitor drivers' behavior and address potential safety concerns associated with facial expressions and mood. Identifying fatigue, distraction, and frustration to prevent accidents before they happen. If a self-driving car perceived emotional distress from passengers, it could drive more slowly or play soothing music to assuage their anxiety. Autonomous driving machines adjust driving styles based on the occupants' non-verbal feedback.

Spill the magic dust Arsalan

2.2 Research Thrust 2: Reach-ability analysis based safety region estimation

This is all you Nicola

Lu adds subsection on control synthesis

The goal of this section is to define a method to quickly assess safety (and perhaps trust) online and then use this reachability analysis to close the loop and correct the system to maintain safety and increase trust

Here discuss recent work on reachability analysis for safety then add the new ideas on online reachability using machine learning and control to predict future states while guaranteeing safety and adapt.

Approach: ♠[5] This thrust is divided in two main parts, as outlined in Figure♠[6]. We start by proposing a reachability-based approach in which we leverage knowledge about the system dynamics and process, sensor, and environmental uncertainties to determine the future states over a finite time horizon. The computed reachable sets will be used to determine the first time that a safety-critical event may occur. More specifically, in this work we are interested in the fist time that the AV may: i) collide with any obstacle, both static and dynamics ii) deviate from the desired trajectory, and/or iii) suffer a failure. Based on these events, we will then design a safe reinforcement learning approach to adapt and correct the behavior of the system to maintain safety and increase trust ♠[7]. Here we propose also techniques for fast verification and monitoring ♠[8].

T1.1 - Online Prediction

Reachability-based Approach: Traditional model-predictive or finite horizon controllers (MPC) [**?**, **?**, **?**] can estimate future states and inputs of a system, given a correct knowledge of its model. On the other hand, reachability analysis (RA) [**?**, **?**, **?**] provides regions that can be reached by the system when applying a sequence of inputs and assuming different uncertainties like process noise, model uncertainties, sensor noise and jitter, delays, and external disturbances. For this reason RA is typically used to monitor and assess if and when safety conditions can be violated [**?**, **?**, **?**].

In this work we propose to research the use of RA to predict reachable states of vehicles under different uncertainties and assess safety constraints like collision avoidance and trajectory tracking. The key idea is that the context (i.e., obstacle configuration) in which the AV is operating dictates the reaction time to switch to a different mode of operation. Thus, predicting as accurately as possible the different situations that can occur in the near future will allow a better safety assessment, ♠[9] and will guide a better planning of actions to perform. Let us consider a vehicle modeled generally as a non-linear system of the form $\dot{\boldsymbol{x}} = f(\boldsymbol{x}, \boldsymbol{u}_m)$ with $\boldsymbol{x}$ the state vector of the system. Assume that the system can be linearized around a point $(\boldsymbol{x}^*, \boldsymbol{u}^*)$ (that can be the current operating point) and obtain the following, discretized with sampling time t_s, state space representation $\boldsymbol{x}(k+1) = \boldsymbol{A}\boldsymbol{x}(k) + \boldsymbol{B}\boldsymbol{u}(k) = \boldsymbol{A}\boldsymbol{x}(k) + \boldsymbol{B}_m\boldsymbol{u}_m(k) + \boldsymbol{B}_I f\left(\boldsymbol{x}^*, \boldsymbol{u}^*\right) + \boldsymbol{B}_d\,\boldsymbol{d}$ where $\boldsymbol{d}$ is an external disturbance.

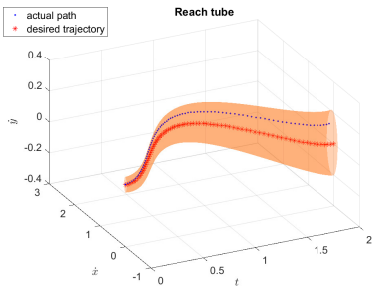

Figure 2: Velocity Reachable tube for a quadrotor following a straight line trajectory for 2s.

In this work we propose reachability analysis techniques for safety monitoring AVs and present a machine learning-based approach to minimize runtime computation time. A reachable set computed at time t_0

[5]NB: need to improve this section
[6]NB: add figure
[7]NB: it would be nice to connect trust here
[8]NB: I will elaborate and rephrase
[9]NB: THUS MORE TRUST

for a future time t_f and represented by $R(\boldsymbol{x}_0, \boldsymbol{u}(t), t_f)$ is an ellipsoid ϵ that contains all the states $\boldsymbol{x}$ reachable at a future time $t_f > t_0$ where the initial set $\boldsymbol{x}_0 \in \epsilon(\boldsymbol{x}_0, \boldsymbol{X}_0)$ is bounded by an ellipsoid with center $\boldsymbol{x}_0$ and shape matrix $\boldsymbol{X}_0$ and the input $\boldsymbol{u}(t) \in \epsilon(\boldsymbol{u}(t), \boldsymbol{U})$ is bounded by an ellipsoid with center $\boldsymbol{u}(t)$ and shape matrix $\boldsymbol{U}$. A reachable tube $R(\boldsymbol{x}_0, \boldsymbol{u}(t), [0, T])$ is then defined as the set of all reachable sets over the time interval $\Delta T = [0, T]$, [**?**, **?**]. The external bound for the reach set at time t starting from time t_0 is calculated based on the initial state ellipsoid, the plant model, and the input ellipsoid, as follows:

$$R^+(t, t_0, \epsilon(\boldsymbol{x}_0, \boldsymbol{X}_0)) = \boldsymbol{\Phi}(t, t_0)\epsilon(\boldsymbol{x}_0, \boldsymbol{X}_0) \oplus \int_{t_0}^{t} \boldsymbol{\Phi}(t, \zeta)\boldsymbol{B}(\zeta)\epsilon(\boldsymbol{u}(\zeta), \boldsymbol{U})d\zeta$$

where $\boldsymbol{\Phi}(t, t_0) = e^{\boldsymbol{A}(t-t_0)}$ and $\boldsymbol{A}$ and $\boldsymbol{B}$ are the state and input matrices related to the AV. Thus, with this approach, reachable sets can be calculated to capture the uncertain motion of the AV tracking a given trajectory. In Fig. 2, the $[\dot{x}, \dot{y}]$ velocity reachable tube for an autonomous quadrotor aerial vehicle following a straight line trajectory for 2s ♠[10], considering disturbances and measurement and input noise, is shown. The blue dotted curve shows the path of the quadrotor whereas the red star curve shows the desired trajectory. The desired trajectory is different from the actual one due to the presence of wind disturbance, however the actual trajectory is contained inside the reachable tubes, since system and sensor uncertainties as well as disturbances are considered when calculating such reachable tubes.

To deal with unmodeled and stochastic system dynamics, we propose a methodology to compute and maximize the probability of maintaining the state of the system within a certain safe region and decide to represent the system as a stochastic hybrid system whose dynamics can be influenced by a control input [**?**, **?**]. Unlike previous approaches, the safe set can be time-varying [**?**, **?**]. The proposed stochastic reachability methodology is based on formulating the reachability problem as a stochastic optimal control problem. Based on the expression of the probability that the state of the controlled system will evolve within the safe region as a multiplicative cost, dynamic programming (DP) can be used to compute the Markov policy maximizing the cost, and also the maximally safe sets corresponding to different safety levels. These are the set of initial conditions for the system, such that there exists a Markov policy capable of maintaining the state of the system within the safe set with a probability greater than a prescribed safety level [**?**, **?**, **?**]. If the objective is to minimize the probability that the system will exit the safe set, then we can formulate the reachability problem as a stochastic optimal control problem, but this time with a cost that is the maximum of a function of the state over the time horizon. Here, again, DP will be considered to determine probabilistic maximal safe sets for Markov policies, [**?**].

Runtime Verification for Safety-Aware Autonomous Vehicle Operations:

Offline verification of autonomous vehicles may not scale to larger systems where exploring all possible contingencies can be computationally prohibitive. For example, an autonomous car can continuously estimating clear roadways and potential obstacles for which the planned trajectory must be verified for safety. To perform offline mission verification would require considering all possible environments (roadways, weather conditions, obstructions, *etc*.). Moreover, virtually all prior research in the area of formally verifying complex mission planning (expressed using temporal logics) assumes the plan is developed off-line [**?**, **?**, **?**]. However, it is imperative to monitor – at runtime – the vehicle to check that the assumptions and properties are indeed satisfied and will continue to hold as the mission evolves. In this thrust, we propose to enable safety-aware learning by performing runtime verification of the control components for mission-critical scenarios where the systems may violate safety constraints.

Runtime verification [**?**] is a lightweight formal methods technique for online monitoring of evolving traces of observations from a system execution with respect to a formal specification. Like most other established formal methods, runtime verification techniques target traditional safety-critical systems, developed

[10]NB: I'll create one for a car

with respect to specifications, fixed during the system design. In AVs, it is possible that monitoring specifications will be changing dynamically. For example, in the presence of external disturbance (like changes in surface type for a terrestrial vehicle, or wind acting on an aerial vehicle), the system may need to switch mode of operation, reconfigure, adapt its control strategy, and replan its trajectory to fight the disturbance and achieve its desired goal, all while maintaining a certain level of safety assurance. Thus, online prediction for runtime verification is critical to AVs. The main challenge here lies on how to perform runtime monitoring online and efficiently maintaining computation time bounded while guaranteeing system's safety. To this end, we propose to develop techniques for prediction of autonomous vehicles safety-critical properties (e.g., probability of hitting an obstacle) that can be autonomically verified at runtime. In this thrust we consider two complementary approaches to runtime verification for safety-aware learning.

Our first approach will be to develop techniques for online verification of missions with AVs. While such problems are computationally very hard, we will leverage our prior work that relies on satisfiability modulo theories (SMT) solvers to address the system of constraints that mix Boolean variables with real variables [**?**]. One of the major challenges we will address is to verify reactive plans that dynamically compose primitives in order to quickly respond to changes in the environment or mission. In this approach, our aim is to generate guarantees based on a probabilistic temporal logic framework (such as the framework described in Thrust 1.3) so that we can achieve verified autonomy in unknown environments with learning in the control loop. Once completed, this will be the first blend of probabilistic formal methods that can reason about probabilistic abstractions of machine learning models in an autonomous system operation.

While recent advances in SMT solvers can be leveraged to perform *fast* verification of controllers, the worst-case complexity of such verifications remain troublesome. This motivates our second approach to runtime verification wherein we abstract the safety conditions such that monitoring can be performed efficiently using predictive techniques (e.g., reachability presented in the previous section). To enable online efficient prediction, hence minimizing computation time and allow real-time monitoring applications, we propose to cast the prediction problem as a two-point boundary value problem (2PBVP) [**?**] and use support vector machine learning algorithms to approximate the reachability boundary using a nonlinear classifier, separating training queries into reachable and non-reachable sets. The idea is to generate training data by solving a large number of 2PBVPs as presented in [**?**] for various randomly generated query pairs. The classifier is then used online to estimate if new query points are reachable. The advantage of such approach is that it can be applied to any system with minimal computation time at runtime since training can be executed offline a priori. ♠[11]

Safe Reinforcement Learning-based Adaptation:

♠[12] Reinforcement Learning (RL) is an area of machine learning concerned with learning how an agent should behave in a given environment in order to maximize some form of cumulative reward. To this end, an agent cycles through a series of transitions which consist of going from one state to the next state by applying actions to the environment and receiving rewards as a consequence of his actions. The goal of RL is to derive a policy, which, given a state, provides the action to take in order to maximize the cumulative reward. A central problem of RL is thus that of properly assigning the reward to the actions that lead to such reward (a notion known as credit-assignment [**?**, **?**]). Another main issue of learning behavioral policies in domains that are unknown at first (especially in autonomous vehicles) is that of efficiently bringing the agent from a tabula-rasa state to a condition where the agent is acting as close to optimality as possible. This notion is also known as regret minimization [**?**, **?**] and is closely related to the topic of trading off exploration of the environment (to sample previously unseen parts of the state-action space) with exploitation of the knowledge

[11] NB: add more discussion here

[12] NB: still working on this section...I'm going to use some of this material but the key idea is to be able to adapt your actions or the AV actions based on how safe the system believes to be

accumulated so far. An agent that enters the world would therefore need to explore the environment by applying actions and gather data which are as informative as possible so that a policy, encoding the knowledge accumulated so far, is derived. The goal is to go from a situation where the system is mostly exploring the environment to a situation where it is mostly exploiting the accumulated knowledge. When the transition probabilities and rewards of a Markov Decision Process (MDP) are known, an agent can obtain the optimal policy without any interaction with the environment. However, exact transition probabilities are difficult for experts to specify and may not be known completely a priori or change over a mission due to aging of the system or unforeseen disturbances and faults. With these considerations in mind, one option left to an agent is a long and potentially costly exploration of the environment. One such algorithm is the E3-algorithm presented in [**?**] in which the basic idea is to repeatedly apply an exploration policy, i.e., one that tries to visit state-action pairs whose transition dynamics are still inaccurately modeled. After a polynomial number of iterations, it will deem itself to have modeled enough of the MDP accurately. Then, it will apply an exploitation policy, which (given the current MDP model) tries to maximize the sum of rewards obtained over time . In [**?**] the authors demonstrate that the E3-family of algorithms [**?**] is often unacceptable because it requires executing policies that explore also unsafe parts of the workspace, including parts that would lead to a crash of the CPS (e.g. crash of the helicopter discussed in [**?**]). The same authors consider reinforcement learning in systems with unknown dynamics and propose an apprenticeship learning approach, in which an initial teacher demonstration of the task to be learned is presented and then by using only exploitation policies over the training data from the teacher to learn the optimal policy. Farther work from the same authors in [**?**] deal with scalability issues in reinforcement learning offering a hybrid algorithm that requires only an approximate model and a small number of trials to obtain a near-optimal performance in a real system. In [**?**], the authors propose another alternative to the suboptimal exploration of the state space presented in the previous works: given initial (possibly inaccurate) specification of the MDP, the agent determines the sensitivity of the optimal policy to changes in transitions and rewards. It then focuses its exploration on the regions of space to which the optimal policy is most sensitive. This technique named Active reinforcement learning enables this type of exploration: it uses sensitivity analysis to determine how the optimal policy in the expert-specified MDP is affected by changes in transition probabilities and rewards of individual actions. This analysis guides the exploration process by forcing the agent to sample the most sensitive actions first. It is shown that the proposed exploration strategy performs well on several control and planning problems. Reinforcement learning id particularly well-suited to problems with a long-term versus short-term reward trade-off. A key challenge is about enabling online learning while maintain safety.

If the uncertainties of the system's model are high, for example in the event of a failure like a motor outage, then we can think to use a machine learning approach to determine the optimal policy and control inputs to provide to the system to maximize a given reward (e.g. go-to-goal) while learning the new model. More specifically in this task we propose an iterative reinforcement learning (RL) framework to determine the control policy and refine the model of the system. The abstract setting can be described in the following terms: a AV is assumed to act according to an optimal policy for a Markov decision process M^{S}. The system knows its current state and action sets as well as the initial transition probabilities for M^{S}. Due to unforeseen disturbances and unpredicted behaviors (e.g., the loss of a motor, high wind) the transition probability model changes, but the reward function remains the same. Following the diagram in Figure **??** the model is not completely known and need to be refined as the system is running. This problem can be cast as an MDP problem in which transition probabilities are changed and adapted at every iteration to handle unknown environment and systems dynamics. Current and historic measurement and inputs will be used to refine the current model and update transition probabilities. We can think to leverage reachability analysis also in this task to incorporate uncertainties as the RL-based approach converges to a model and policy and to predict accordingly the possible future states that the robot may cover.

An even reacher extension to this planned work will consider disturbance and uncertainties in the model and environmental dynamics as an adversarial player with an unknown reward that we would like to learn. This specific problem can be solved with a number of inverse reinforcement learning algorithms. For this case we can still model the decision problem as an MDP M^{D} in which reward is structured around misclassification cost experienced in terminal states that correspond either to allowing the AV to proceed to its intended terminal state or to change control law. The state definition for M^{D} includes all of the robot's state elements as well as a belief function for the robot's reward function. Because it includes a belief state, MDP will generally suffer from the curse of dimensionality and be intractable from the perspective of full-width planning methods. To address this issue, we propose to make use of work on Monte-Carlo planning in large partially-observed MDPs (POMDPs) [**?**]. The basic idea is to use a simulator as a generative model of the decision process. The simulator is used to generate sequences of states, observations, and rewards that are used to update the value function. Following [**?**] and others, sampling will be done according to a Monte-Carlo trees search algorithm. Other methods, such as DESPOT [**?**], and recent POMDP computational methods [**?**, **?**, **?**, **?**] will also be considered.

2.3 Research Thrust 3: Feedback Design

The purpose of this research thrust is to develop and validate models to quantify trust of a driver in an autonomous vehicles. Trust in self-driving cars is one of the big discussion points in the public debate. Drivers who have always been in complete control of their car are expected to willingly hand over control and blindly trust a technology that could kill them. We hypothesize that trust is influenced by three components:

1. The person who trusts,
2. The system this person is supposed to trust, and
3. The driving situation.

In addition, the first component (i.e., the person), is characterized by a certain propensity to trust, which is influenced by different factors (e.g., gender, age, opinions, character traits). This is what we want to measure in the proposed research.

While cars have become significantly more usable — particularly with regard to reliability and safety over the past twenty years — thanks to the introduction of new technologies such as electronic fuel injection, the seat belt, crumple zones, ABS, airbags, electronic stability control and GPS satellite navigation, many of these technologies have succeeded out-of-sight of the humans behind the wheel. Yet when it comes to newer technologies - both on-board telematics,communication and the ADAS, we see a much less successful integration of technology, vehicle and user. At the broadest level, many of the technologies available in modern cars do not appear to have been developed with a particular user-centred approach. They exist because the technology has become available to perform a specific function.

As soon as driving "feels" even partly autonomous, people switch off, they become disengaged from the process of driving — and fail to monitor the system. A quick YouTube search, reveals many videos where people are aware the systems have limitations, but still push them further than their intended use, operating pilot assist systems on roads or situations when they shouldn't.

Trust comes from two factors: predictability and explainability. If a user expects a car to drive in a certain way in a certain situation, and the car conforms to his expectation, the user will tend to trust it more. On occasion, when the AV's action surprise/confuse the user- as long as there is an explanation provided for it, the user can again gauge her level of trust int he system. Our goal is :

1. Understand the set of trust expectations a driver has for their autonomous vehicle.

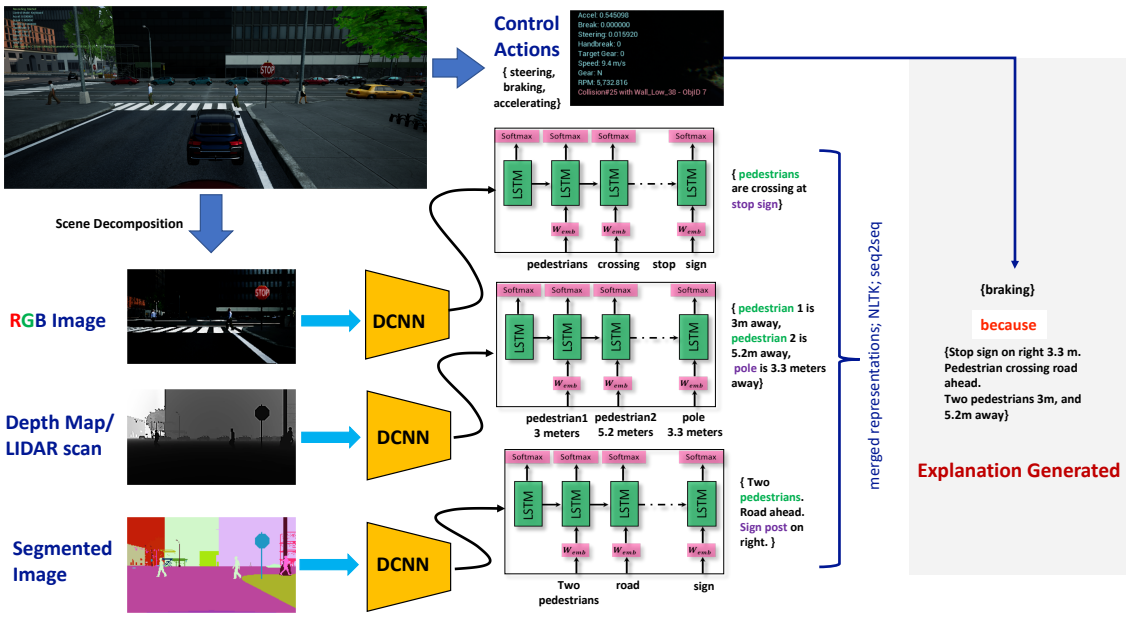

Figure 3: Deep-Explanation generation: Each dimension of the scene decomposition is used as an input to caption generation. Representation matching, and seq2seq are then used to generate a likely explanation for the predominant action stream.

2. Develop a set of explanations autonomous vehicles should provide to promote trust and create an algorithm that can provide these explanations.
3. Perform experiments to understand how content, timing and delivery impacts the effectiveness of explanations.

2.3.1 Scenario-based trust modeling

Madhur describes the scenario based experiments with prescan

I will add some text to this subsection as well.

2.3.2 Deep explanations

madhur to add.

Automatic image description generation is a challenging problem that has recently received a large amount of interest from the computer vision and natural language processing communities [**?**, **?**, **?**, **?**, **?**] Not only must caption generation models be able to solve the computer vision challenges of determining what objects are in an image, but they must also be powerful enough to capture and express their relationships in natural language. For this reason, caption generation has long been seen as a difficult problem. The task of automatic image description involves taking an image, analyzing its visual content, and generating a textual description (typically a sentence) that verbalizes the most salient aspects of the image. This requires the joint use of both Computer Vision and Natural Language Processing techniques. Yet despite the

difficult nature of this task, there has been a recent surge of research interest in attacking the image caption generation problem. In particular, deep neural networks have been shown to form new grammatically correct sentences as opposed to the template based models and their limited generalization capability to a novel image. To capture the correlation between two modalities i.e. visual and natural language we need to map both these to some same space so at learn the relation between them or say we need to learn the multimodal joint model. Models that uses different deep neural networks like convolutional neural network (CNN), long short term memory(LSTM) networks, recurrent neural network(RNN) to implicitly learn the common embedding. These by far gives the best result on all common datasets of caption generation Aided by advances in training deep neural networks and the availability of large classification datasets, recent work has significantly improved the quality of caption generation using a combination of convolutional neural networks (convnets) to obtain vector representation of images and recurrent neural networks to decode those representations into natural language sentences.

In the proposed research we extend attention-based image caption generators to work with multidimensional data-sets.

1. Instead of generating captions from RGB images alone, we will also generate captions from LIDAR data, depth sensor images, and segmented images.

2. The captions themselves, will be enhanced with information about the control decision (steering, acceleration, and braking) made.

3. We will gather and release a multidimensional caption data-set specifically for autonomous vehicles.

3 Evaluation/Experimentation Plan

- We need to describe in detail all the experiments and testbeds we will use/develop.
- this is emphasized int he solitication
- describe how we will recruit subjects
- emphasize the real world experimentation plan

This section should describe how the research concepts proposed will be demonstrated and validated. It should present metrics for success. It should identify critical experiments, and describe how the research will be demonstrated, including through simulation, prototyping, and integration with real (including sub-scale) cyber-physical systems. For Medium and Frontier projects, the validation plan must include experimentation on an actual cyber-physical system.

Lu describes the full scale driving simulator

Figure 4 shows the setup of an academic-scale testbed that co-PI Feng's group has partially built and will complete during the proposed effort. The hardware platform is based on the Force Dynamics 401CR driving simulator. This four-axis motion platform can pitch, roll, yaw, and heave, to simulate the experience of being in a vehicle. Thus, we expect to collect data about realistic human response during the driving. The human is interfaced to the hardware platform through PreScan, which is a software tool designed as a development environment for Advanced Driver Assistance Systems (ADAS) and Intelligent Vehicle (IV) systems. These are systems with sensors that monitor the vehicle's surroundings and that use the acquired information to take action. Such actions may range from warning the driver of a potentially dangerous situation to actively

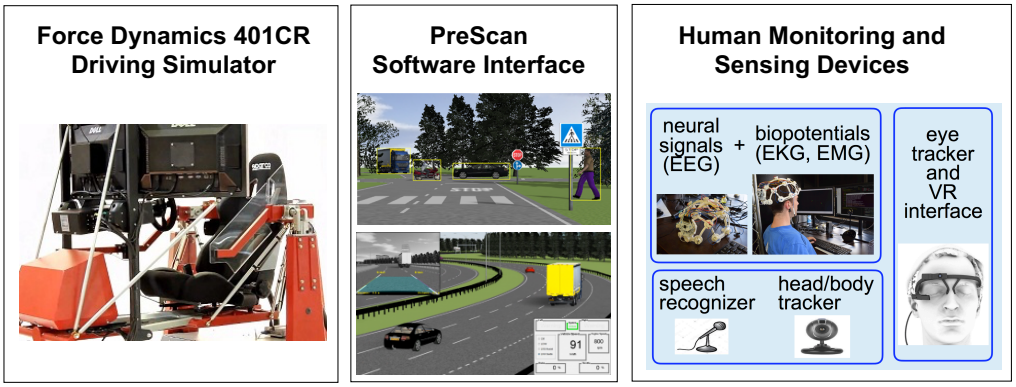

Figure 4: Our partially built testbed of human-autonomous vehicle interactions. The human is monitored by a number of sensors, and interacts with the driving simulator through the PreScan software interface.)

evading hazards by means of automatic braking or automatic steering. PreScan can be used for designing and evaluating ADAS and IV systems that are based on sensor technologies such as radar, laser, camera, ultrasonic, GPS and C2C/C2I communications. The "human monitoring and sensing" block in Figure 4 encloses the sensors that will be used for both high-level inference of human's intent and preferences and low-level monitoring of human behavioral, mental and physiological states. These sensors include EEG for neural signals, EKG for heart activity, EMG for muscle activity, a camera for head tracking, eye tracking suite and cloud-based speech recognizer.

Nicola describes the experiment setup for reachability analysis experimentation

Madhur describes F1/10 - could be used for reachibility + control synthesis

4 Project Management and Collaboration Plan

5 Broader Impacts

MB to add - others add too.- education, outreach, diversity

6 Team

Arsalan adds Bio

Nicola adds Bio

Lu adds bio

Madhur adds bio

E. Data Management Plan

Data generated as part of this project, including instrumentation code, modeling code, training data, model descriptions, etc. will be managed using existing University of Virginia infrastructure for administering and maintaining digital research, with automatic nightly back up of source-code and documentation repositories), at no additional cost to the project. Some of the resources and data management practices are already in place, and being utilized by the PI in his research group. We will adopt and extend these practices for short-term data collection, retention and management.

The PI, Madhur Behl, will have responsibility for coordinating and directing the retention and sharing of data generated through the proposed research activities. Because this work is collaborative across five PIs, many of the data management activities during the project period will reside with the coPIs and will be completed in adherence with NSF and university policies.

Expected Data

An outline of the data expected to be generated by the project is as follows:

1. *Human driving profile data* **Add explanation and address IRB**
2. *Modeling algorithms and code:* A second major code artifact of our project will be the modeling algorithms and toolchains used to build our scenario models. We will publish the mathematical formulations of these techniques for archival by the publisher, and will maintain implementations of these techniques (*i.e.*, code) in version control. Concurrent with the publication of these techniques, we will release code implementing them using open source licenses.
3. *User interface design data:* **Add explanation**

Period of Data Retention

During the project, the data associated with individual tasks will be stored by the PI using his laboratory's local servers and other resources. At regular intervals (quarterly), all the project related data will be copied and archived using the University of Virginia's digital repository.

Data Formats, Short-term storage and dissemination

All the data described above and its accompanying documentation will be incrementally made accessible to researchers and the general public as mentioned before. All the computer codes will be implemented and made accessible to the scientific community in the form live web-tools. Selected source code, associated input/output files and documentation will also be released as the project and this computational modeling technology matures. The researchers retain rights to access and utilize data in whatever format but will not limit requester's ability to re-use or re-distribute processed data or materials. The data will be deposited in established repositories, for example the UVA institutional repository Libra. Libra is an open repository with public access. Therefore, care will be taken to ensure confidential and sensitive data are not shared through Libra. We reserve the right to delay release of project data for a period of time to allow for publication of research results. This period will not exceed three years following the project end date.

Long-term Data Storage and Preservation of access

The UVA Libra system provides servers, backup procedures, and other policies to minimize the change of data loss. In accordance with the University of Virginia policy RES-002, "Policy: Laboratory Notebook and Recordkeeping," the data will be preserved for a minimum of five years upon completion of the project. However the current preservation plan for Libra will be to preserve the data indefinitely. The Libra backup plan provides for data redundancy including off-site storage. If the Principal Investigator resigns from the university, the department chairperson for the lead department will become the custodian of the data and will assume all the responsibilities for data management, control and dissemination on behalf of the University of Virginia.

F. Budget Justification

G. Facilities, Equipment, and Other Resources.

The research project will take place at the The Link-Lab - the Cyber-Physical Systems lab located on the University of Virginia's (UVA) campus. The facilities are partitioned into several laboratories that provide a complete environment for the design, fabrication, and testing of prototype hardware/software systems from initial concept to final implementation. These facilities include sufficient computing and prototyping resources for the proposed research, as described below:

A. UVA Link Lab - The PI and Co-PIs are members of the Link-Lab at UVA – this space is a new collaborative initiative on Cyber-Physical Systems (CPS) research and education at the University of Virginia. The lab is called "Link Lab" because it "links" multiple engineering departments through cross-cutting mechanisms such as shared lab space, staff, and conference rooms that house faculty and students from multiple disciplines. It houses approximately 20 faculty, 125 students, 3 research scientists, 3 staff, and 6 postdocs. An open floor plan promotes cross-pollination between research groups while at the same time using furniture and layout to provide sound insulation and reduce interruptions.The Link Lab has a large 3000 ft2 open space with moveable tables called the "Arena" that is designed for equipment staging, testbeds, experimental work, as well as a shared common space for seminars, presentations, or social events. It also includes a 2000 ft2 hardware prototyping lab. By locating the space immediately adjacent to the student and faculty desks, and close to the door, this layout will facilitate daily collaboration on this project. Figure 1, illustrates the Link-Lab floor plan.

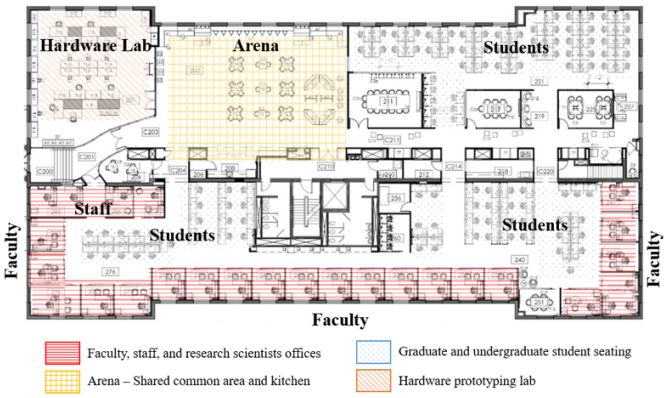

Figure 5: Link-lab floor plan

B. UVA Viz Lab The Viz lab is a facility at UVA designed to help faculty, staff and students explore various 3D visualization tools, such as virtual and augmented reality head mounted displays, for research and education purposes. The staff at Viz lab provide assistance with developing and evaluating virtual and augmented reality environments.

C. Autonomous Mobile Robots and CPS Lab: Part of the proposed autonomous vehicles testing will be carried out using the testbed available in Co-PI Bezzo's "Robotic and CPS Lab" which has a large, dedicated, state-of-the-art facility for mobile robotic systems development, prototyping, and control. The lab has been recently renovated and includes support for diverse system development activities

with electronics workbenches, flexible space for assembling and experimenting with large demo systems, and secure storage in addition to the computing resources. This space covers an area of more than 900 ft2 with high ceilings and includes:

- Latest generation Vicon Motion Capture System: 8 Vantage cameras (@350 fps) with Lock system and Vicon Tracker software. This system allow sub-millimeter precision localization and tracking of multiple objects moving within the volume of the Lab space
- 2 Ascending Technologies Pelican quadrotors with carbon fiber body and equipped with 3rd generation Intel Core i7 CPU, IMU, pressure sensor, GPS, Lidar, stereo cameras, 2.4 GHz XBee link and WiFi - max airspeed 16 m/s, max climb rate 8 m/s, max thrust 36N, max payload 650g.
- 1 Ascending Technologies Firefly hexarotor with carbon fiber body and equipped with 3rd generation Intel Core i7 CPU, IMU, pressure sensor, GPS, Lidar, stereo cameras, 2.34GHz XBee link and WiFi (max airspeed 15 m/s, max climb rate 8 m/s, max thrust 36N, max payload 600g.)
- 3 Ascending Technologies Hummingbird quadrotors with carbon fiber body and equipped with IMU, pressure sensor, GPS, and 2.4 GHz XBee link - max airspeed 15 m/s, max climb rate 5 m/s, max thrust 20N, max payload 200g.
- Several Crazyflie 2.0 nano quadrotors equipped with IMU, pressure sensor, bluetooth, and Qi inductive charger.
- Several Parrot Bebop quadrotors equipped with IMU, two cameras, GPS, and sonar.
- 2 Clearpath Jackal unmanned ground vehicles, equipped with 3rd generation Intel Core i7 CPU, IMU, NovAtel SMART GPS, Velodyne 3D Lidar, Point Grey Flea3 camera, and WiFi - max speed 2 m/s, max payload 20 kg.
- 1 Black-i LandShark military grade 6 wheels UGV equipped with Intel Core i7 CPU, automated turret, Moog Quickset GeminEye, 100x zoom camera, thermal imager, 2 fisheye cameras, 1 camera, 2 Microstrain IMUs, 12 sonar rangers, 12 IR rangers, 2 Hokuyo UTM-30LX Lidars, GPS, and OCU (max speed 10 mph, max payload ¿ 200 lbs)
- 1 Stratasys uPrint SE Plus 3D printer
- 14 Cisco Systems, standard and high definition dome surveillance cameras with power- over-ethernet capability, along with hardware and software to support surveillance management and leading-edge video analytic completed with high-quality teleconference system and cloud capabilities

D. Computational Resources and Rivanna Compute Cluster - The university operates a Linux-based commodity cluster with a frontend named Fir. This cluster is managed by UVa Advanced Computing Services and Engagement and is open to faculty, staff, and graduate students at the University. Undergraduate students and university affiliates are eligible for accounts under faculty sponsorship. Fir is a large-memory cluster consisting of 92 nodes. Twelve of the nodes contain one dual-core 3-GHz Intel dual-core Xeon cpu with 32GB of RAM per node. Another 56 nodes are 8-core servers, with 48 GB per server. There are also 24 12-way nodes with 96 GB per node. Most of these cores are hyperthreaded, bringing the total number of logical cores that are eligible for no preemption to 1496. The interconnect for all these nodes is GigE.

Rivanna, launched in Fall 2014, is the Cray CS300, a 4800-core, high-speed interconnect cluster, with 1.4 PBs of storage. It is composed of 240 compute nodes, each with two 10-core processors and FDR Infiniband interconnect along with a parallel filesystem capable of providing about 25Gb/sec

bandwidth. The Cray cluster combines large amounts of processing with large amounts of memory to provide a significant new resource for computationally-intensive research at UVA.

E. Laser Cutters - Campbell Hall has two Universal Laser Systems CO2 lasers. The 50 watt X660M has an 18"x32" bed. The 25 watt M-300 has a 12" x 24" bed capacity. Both can cut virtually any material other than metal, PVC plastics, or anything reflective. These machines can cut and engrave using both vector (lines/shapes) and raster (pixels) modes from virtually any software program.

F. 3D Printer / Rapid Prototyping - The Stratasys Dimension SST 3D printer uses Fuse Deposition Modeling (FDM) technology to build solid ABS plastic model prototypes from 3D stereolithography (stl) files. The machine has an 8"x8"x10" build envelope and builds models in layers down to 0.010 in. thickness.

G. 3-axis Miller and Routing - The MicroMill 2000 and MicroRouter from Denford, Inc. provide full 3-axis CNC and CAM machining capability. Using CAD/CAM software (EdgeCAM / MasterCAM) to generate G-code instructions for the machine, we can translate 2D profiles or 3D solid/surface geometry into machined parts. The router supports 12"x24"x2.5" travel with the ability to feed in and clamp longer stock materials from the side, while the mill supports approximately 9"x3.5"x6" of travel for a single machining operation. Common materials include wood, foam, plastic, aluminum, brass, copper, and mild steel. The machine is also capable of milling marble

H. 3D Digitizer / 3D Laser Scanner - Using technology from MicroScribe and NextEngine, the fabrication facility can both digitize and scan 3D objects and models into CAD systems. The MicroScribe point digitizer captures point, line, spline and surface information using common 3D modeling and CAD software. These can be used to generate surface and solid models of objects, models, topography, and reliefs. The NextEngine 3D Laser Scanner can scan up to a full 360 degree revolution around a small object, creating a full polygon mesh model for export into any 3D CAD software.

I. Software Resources - The University of Virginia has site licenses for a variety of software for basic computing needs as well as modeling and data analysis, including ANSYS, LabView, Mathcad, and MATLAB.