Toward Closing the Loop on Human Values

Sarah M. Thornton, Benjamin Limonchik, Francis E. Lewis, Mykel J. Kochenderfer, and J. Christian Gerdes

***Abstract*—Human drivers navigate the roadways by balancing values such as safety, legality, and mobility. An automated vehicle driving on the same roadways as humans likely needs to navigate based on similar values. The iterative methodology of value sensitive design (VSD) is used to formalize the connection of human values to engineering specifications. A modified VSD methodology is used to develop an automated vehicle speed control algorithm to safely navigate a pedestrian crosswalk. Two VSD iterations are presented that model the problem as a partially observable Markov decision process and use dynamic programming to compute an optimal policy to control the longitudinal acceleration of the vehicle based on the belief of whether a pedestrian is crossing. The speed control algorithms were tested in real time on an experimental vehicle on a closed-road course.**

***Index Terms*—Automated vehicles, motion planning, value sensitive design, human values.**

I. Introduction

THE roadways are populated by many stakeholders, such as pedestrians, bicyclists, and vehicle occupants. Automated vehicle designers have not only to solve the problem of smoothly navigating through the environment with these various road users but also navigating their expectations of appropriate vehicle behavior. Such expectations arise from the values these stakeholders attach to their experience on the road, incorporating such elements as mobility, safety, and legality [1], [2]. The challenge is to connect these more abstract human values to engineering specifications. One way to address this challenge is to integrate the stakeholders and their values directly into the design process of algorithms for automated vehicle motion planning.

The process of engaging stakeholders and obtaining their input on the design is broadly known as human-centered design (HCD) [3], [4]. Practitioners of HCD engage with a group of stakeholders, largely direct users of the technology, in order to get feedback on how to improve designs. Generally, these interactions are structured around making the design more usable. Both Millar [5] and Niemelä *et al.* [6] suggested that a focus on usability can detach this process from broader ethical considerations and prevent understanding a designs potential ethical implications. Given the broad implications of automated vehicles for society, it is important to consider a wider range of values. A related approach that takes such a broader look at ethics or values is known as life-based design (LBD) [7]. LBD approaches the design task by investigating the needs of stakeholders through their quality of life. LBD entails describing the human requirements in the design activity, then identifying the users and technology requirements, and finally determining if the humans quality of life improves based on the designed technology. While HCD and LBD are both iterative design processes that attempt to improve the design of technology for the respective users, HCD focuses on values relating to usability and LBD focuses on values relating to quality of life. Both approaches have merits, but automated vehicle design requires a process that includes a broader set of values than traditional HCD while bringing more specificity than the abstract quality of life focus LBD provides.

Value sensitive design (VSD) [8], [9] satisfies these criteria. VSD is a methodology that addresses ethical considerations by explicitly incorporating values (usually emphasizing those with ethical import [10]) throughout the entire design process. VSD consists of a tripartite methodology of iterating over conceptual, technical, and empirical stages of a design. At every stage of the design process, human values are connected to the designed technology. VSD is most applicable to a design task in which value conflicts exist for ethical issues. Friedman *et al.* indicated VSD has been found to be widely useful in the human-computer interaction community to help balance privacy concerns with usability for end-users [9], [11]. In the design of an office space with a virtual window viewing a public plaza, Friedman *et al.* [9] demonstrated that recognizing indirect stakeholders (a component of the conceptualization phase) uncovered privacy concerns for passersby. Denning *et al.* used VSD to construct a list of specifications to guide future designs of a security system in implantable medical devices [12]. Furthermore, the generality of VSD allows for modification to be in line with certain design tasks. Wynsberghe appended to VSD the moral theory of "care ethics" in the design of health care robots to ensure the robots reflect stakeholder values [13].

In designing algorithms for automated vehicle motion planning, engineers already account for some human values. Many algorithm designs focus on the values of safety and efficiency as is demonstrated by Chen *et al.*'s evaluation framework of an automated vehicle approaching an unsignalized pedestrian crosswalk [14]. Bandyopadhyay *et al.* [15], [16] also focused

Manuscript received September 16, 2018; revised December 18, 2018 and December 20, 2019; accepted March 8, 2019. Date of publication May 28, 2019; date of current version August 23, 2019. This work was supported by the Ford-Stanford Alliance. *(Corresponding author: Sarah M Thornton.)*

S. M. Thornton, B. Limonchik, F. E. Lewis, and J. C. Gerdes are with the Dynamic Design Lab Department of Mechanical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: smthorn@alumni.stanford.edu; ben.limonchik@gmail.com; glewis17@stanford.edu; gerdes@stanford.edu).

M. J. Kochenderfer is with the Stanford Intelligent Systems Lab, Department of Aeronautics and Astronautics, Stanford University, Stanford, CA 94305 USA (e-mail: mykel@stanford.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIV.2019.2919471

on safety and efficiency in the creation of the reward function of a partially observable Markov decision process (POMDP) for speed control in pedestrian environments. Brechtel *et al.* similarly considered safety and efficiency in the construction of the reward function of a speed control POMDP for entering occluded intersections [17]. These examples demonstrate that connecting human values to automated vehicles is not new. The focus on safety and efficiency in the evaluation framework and motion planning policies, however, highlights the difficulty of designing for conflicting values. Brechtel *et al.* additionally considered occupant comfort in the reward function and suggested that traffic rules can be included in future iterations of the POMDP design. Their discussion indicated a desire to account for the various human values at stake in the design of a motion planning policy. To account for these values, it would be useful to have a methodology that can help determine which values to include because humans value more than just safety and efficiency. Having a list of identified values is also useful to determine conflicts between stakeholders and values. Establishing value conflicts early in the design process can help engineers design technology that explicitly resolves them early on rather than requiring patchwork solutions to dissolve value tensions after a system deploys.

Here, it is proposed that VSD can help fill the gaps in the design process of motion planning algorithms for automated vehicles. VSD is used to formalize the connection of human values to engineering specifications by enumerating the human values that are at stake in the design problem and by resolving value conflicts through justification of design choices. This paper demonstrates a continued application of VSD to automated vehicle motion planning with the design task of a speed controller for the scenario of a pedestrian crosswalk. The speed is controlled by acceleration commands from POMDP policies designed using VSD. The VSD speed controllers presented in this paper are the first two iterations from the design process and are not final products. The contributions are as follows:

- A second iteration of VSD for the design of a speed control algorithm to navigate a pedestrian crosswalk:
  - Identification of the stakeholders and human values implicated in the design task given the insights from the first iteration (conceptualization)
  - Prototype of a new technology given insights from the first iteration, which is a modified version of the first design (technical implementation)
  - Experimental results of the new technology (empirical analysis)
- Explicit documentation of how values are incorporated into the speed control design and how tensions between values are resolved.
- An example of how to close the loop between the implementation and the identified human values is conducted by multi-objective optimization.

This paper proceeds as follows. Section II outlines the three phases of the VSD methodology: conceptualization, technical implementation, and empirical analysis. For the scenario of an occluded pedestrian crosswalk, Section III presents the first iteration of VSD for the speed control design task. Section IV describes the second VSD iteration. An important part of designing a motion planning algorithm with human values is ensuring the realization of said values. Section V elaborates on one technique that can assist with closing the loop on the integration of human values into the technology. A summary is provided in Section VI.

II. Value Sensitive Design

The methodology of value sensitive design (VSD) consists of three phases: conceptual, technical, and empirical [8], [9]. During the conceptual phase, the methodology involves identifying the values encompassed by the technology. Additionally, the conceptualization phase determines the direct and indirect stakeholders of the technology. A feature of VSD holds that some technological implementations are better suited to uphold certain values than other implementations. For the technical phase, the technical solutions most in line with the identified values (from the conceptual phase) are used to develop the technology being designed. Finally, the empirical phase allows for quantitative and qualitative analyses of the developed design, such as data analysis or observations from human-user studies. This period allows for inspection of how successfully the designed technology meets the conceptualization. Throughout the design development, the designer iterates over the various phases until all three align. Engineers already implicitly iterate over conceptual, technical, and empirical phases as they design new technology. VSD provides a tool to help formalize the engineering process to explicitly account for values embedded in the technology by identifying the values and tracking these values throughout the iterations.

III. The First Iteration

Designing an automated vehicle motion planning algorithm requires addressing a broad array of situations the vehicle may encounter on the roadways. With such broad impact, the list of stakeholders and values may be untenable to design for. To simplify the design task, this paper will focus on a particular scenario as a form of case-study in order to confine the stakeholder and value consideration space.

The first iteration considers the scenario of a two-lane roadway with a single, dashed yellow line. The roadway also includes a marked pedestrian crosswalk. In front of the crosswalk is a large, illegally parked van. From the perspective of the automated vehicle approaching the crosswalk, the crosswalk is partially occluded due to the van. The steering controller from Thornton *et al.* [18] laterally controls the vehicle around the van along an obstacle-free path. The design task is to develop a speed control algorithm along the given path such that the automated vehicle safely navigates the scenario. The conceptualization phase is summarized in Table I, where v_t is the vehicle speed, d_t is the vehicle distance to the crosswalk, c_t is a Boolean for whether or not the pedestrian is crossing, a_t is the acceleration command, and Δt is the change in time. Details of the conceptualization, technical implementation, and empirical analysis of the first iteration of VSD are presented in Thornton *et al.* [19].

TABLE I
SUMMARY OF HUMAN VALUES MAPPING TO ENGINEERING SPECIFICATIONS FOR THE FIRST VSD ITERATION

Human Value	Engineering Specification	Information
Safety Legality Care and respect for others Respect for authority Fairness and reciprocity	Safety and Legality	v_t d_t c_t
Mobility Individual autonomy	Efficiency	v_t
Trust Transparency	Smoothness	a_t Δt

The first iteration of the speed control design demonstrates the difficulty of designing an algorithm when there are competing values. There are some components of the implementation to highlight and some components to improve.

Positive Outcomes:

- Accounting for the pedestrian uncertainty allowed the vehicle to successfully yield to the pedestrian. This effect largely came from the vehicle approaching the crosswalk at a "reasonable speed" because the POMDP anticipated future state information.
- The only information used about the pedestrian was whether they were detected. This largely upheld the values of fairness and reciprocity but was not explicit.
- With proper choice of weights, the tension between safety/legality and efficiency/mobility can be balanced.
- Modeling the problem as a POMDP and solving for an offline policy helped with investigating and balancing some of the value tensions in this design task.

Outcomes to Improve:

- Remove the limitation on braking authority. In the experiment, this led to occupant comfort being prioritized above safety.
- Although the POMDP formulation is intentionally designed for occupant comfort, it optimized only for smoothness in velocity and did not account for the jerk vehicle occupants experience due to choppy acceleration commands. The in-vehicle experiments indicate that smoothness may not have been properly accounted for with this first iteration.
- This particular scenario is not generalizable to non-occluded crosswalks.
- Pedestrian modeling is key to the value tension, so more focus is needed there.

The next iteration will explore how to maintain these positive attributes while addressing some of the downsides of this implementation.

IV. THE SECOND ITERATION

The iterative process of value sensitive design is not only helpful to identify how to improve the technical implementation but can also be used to re-evaluate the design task. In the second iteration, the scenario is revised to focus on the uncertainty of pedestrian behavior. By eliminating the occluding vehicle, the design task can investigate how pedestrian intent affects the vehicle behavior and vice versa. The pedestrian behavior introduces a lot of uncertainty in crosswalk scenarios, and the inclusion of an occluding vehicle obfuscates the pedestrian-vehicle interaction. Hence, the occlusion is removed and a pedestrian is positioned at the side of the road as shown in Fig. 1. As the iterations progress, it is reasonable to assume the designer will gain a better understanding of the pedestrian-vehicle interaction. At this point the occlusion can be re-introduced to the design task or used as a test case in the analysis phase.

Fig. 1. Experimental scenario of pedestrian crosswalk.

A. Conceptualization

The design task still involves various stakeholders and touches upon many human values. The direct stakeholders are now the occupants in the automated vehicle, the pedestrian potentially crossing the street, and the authority of traffic laws. Even with the removal of the occluding vehicle, the values at stake in the scenario are the same as the first iteration: mobility, safety, legality, care and respect for others, fairness and reciprocity, respect for authority, trust and transparency, and individual autonomy. In lieu of engaging with actual stakeholders, this list of human values stem from Haidt [20] and Choi and Ji [21]. Haidt suggested there is a set of values (or moral foundations) inherent to human beings, such as care and respect for others, fairness and reciprocity, respect for authority, and individual autonomy, while Choi *et al.* indicated trust and transparency are important for the acceptance of automated vehicles. The values take on the same definition as the first iteration:

- *Care and respect for others* manifests by the desire to not harm other persons.
- *Fairness and reciprocity* affect both the vehicle occupants and pedestrian stakeholders in that the automated vehicle should not take biased or discriminatory actions based on information about the stakeholders. The automated vehicle should treat all agents equally.
- *Respect for authority* engages the relationship between the automated vehicle and its adherence to traffic laws.
- *Trust* emerges when the pedestrian assumes an oncoming vehicle yields to his or her right-of-way while crossing within the crosswalk. *Transparency* occurs when the automated vehicle's actions facilitate this trust.
- *Individual autonomy* of the vehicle occupants acknowledges the desire to get from one destination to another with little impedance.

For this second iteration, how the values translate to an engineering specification is going to be refined.

TABLE II
SUMMARY OF HUMAN VALUES MAPPING TO ENGINEERING SPECIFICATIONS FOR THE SECOND VSD ITERATION

Human Value	Engineering Specification	Information
Fairness and reciprocity	Do not use discriminatory information.	
Legality Respect for authority	Legality	v_t d_t c_t p_t
Safety Care and respect for others	Safety	v_t d_t c_t p_t
Mobility Individual autonomy Trust Transparency	Mobility	v_t a_{t-1} a_t

In the last iteration, the only human values explicitly considered were those that related to an engineering objective. This iteration serves to clarify how each identified value is to be captured in the technology. In particular, the value of fairness and reciprocity does not translate directly to an engineering objective. Instead, it becomes a higher-level design constraint that limits the information to be non-discriminatory, e.g., no age or gender information.

The other values are addressed by relating them to specifications that can be captured by engineering terms. These are summarized in Table II.

1) Legality and Respect for Authority: In the California Vehicle Code §21950, safety and legality are not strictly the same requirement. The vehicle code only requires drivers to exhibit "due care" to be safe, which is not the same requirement to actually be safe. The vehicle code further specifies reducing the vehicle speed and taking actions as necessary to safeguard the safety of the pedestrian. The key pieces of information necessary for legal decision-making are vehicle speed (v_t), vehicle distance to crosswalk (d_t), and pedestrian behavior. In order to safeguard the pedestrian, the automated vehicle must have information about whether the pedestrian is going to transition from the sidewalk to the crosswalk. The pedestrian behavior is assumed to be captured by the pedestrian position (c_t) and pedestrian posture (p_t), which are non-discriminatory.

2) Safety, Care and Respect for Others: The value of safety is a more strict interpretation of the vehicle code. Safety focuses on harm and injury reduction. To address this value, the same information as for legality is needed: vehicle speed (v_t), vehicle distance to crosswalk (d_t), pedestrian position (c_t), and pedestrian posture (p_t).

3) Mobility and Efficiency: The metric of time efficiency is captured by the value of mobility, and it is directly related to the speed of the vehicle (v_t) for a straight path.

4) Mobility and Smoothness: An additional aspect of mobility is smooth driving, which still affects occupant comfort and interjects trust and transparency between the stakeholders. In this iteration, the value of mobility is intended to improve by using both the previous acceleration (a_{t-1}) and current acceleration command (a_t) for smooth change in actions.

TABLE III
PEDESTRIAN TRANSITION MODEL FOR THE SECOND VSD ITERATION

Pedestrian posture	Transition probability $\Pr(c_t \mid \neg c_t)$
Distracted	0.5
Stopped	$0.523^{\dagger}(d_t/d_{\max})$
Moving	$0.867^{\dagger}$

†calculated from yield event statistics [22].

B. Technical Implementation

A new iteration provides an opportunity to choose a different technique or algorithm that better aligns with the defined values. Since the POMDP helped illuminate value tensions in the previous iteration, the POMDP is kept in this iteration since it offers potential resolution. Dynamic programming is used again to compute an optimal policy to control the longitudinal acceleration of the vehicle based on the belief of a pedestrian crossing.

Given the engineering specifications of legality, safety, and mobility, the information necessary to address their respective values in the objective function is captured by the state vector

$$x = [v_t \quad d_t \quad c_t \quad p_t \quad a_{t-1}]^{\top} \tag{1}$$

and the control input

$$u_t = a_t, \tag{2}$$

where v_t is the vehicle speed, d_t is the vehicle distance to the crosswalk, c_t is the pedestrian position captured as a Boolean value because the pedestrian is either crossing or not, p_t is the pedestrian posture, and a_{t-1} and a_t are the previous and current longitudinal acceleration, respectively. The top speed of the roadway is assumed to be 10 m/s, so the vehicle speed is upper bounded by the speed limit to coincide with both the legality and safety objectives. The pedestrian position is either in the crosswalk or on the sidewalk, and the pedestrian posture is either stopped while the pedestrian makes eye contact with the vehicle, is distracted, or is in motion. In order to continue to uphold the values of fairness and reciprocity, the pedestrian states do not rely on other information about the pedestrian that may be discriminatory. Previously, the control input was limited to ± 3 m/s^2 to provide comfortable acceleration values, but this impeded the vehicle's ability to be safe [19]. Here, the control algorithm allows the vehicle to use its full braking authority by allowing deceleration up to -10 m/s^2.

The dynamics (or state transitions) still use a point mass model of the vehicle to calculate the distance to the crosswalk and vehicle speed. A new model for the pedestrian is developed in order to further investigate the value tensions for the design task (Table III). The likelihood of the pedestrian transitioning from the sidewalk to the crosswalk is a function of the pedestrian posture. The likelihood is 50% when the pedestrian is distracted and 86.7% [22] when the pedestrian is in motion. (The probability of 86.7% is calculated from Schroeder and Rouphail's [22]

statistics on yield and non-yield events for an assertive pedestrian at site B.) When the pedestrian is stopped while making eye contact with the vehicle, the probability of transitioning is a function of the vehicle's distance to the crosswalk

$$\Pr(c_t \mid \neg c_t; p_t = \texttt{STOPPED}) = (p_{\text{xing}}/d_{\max})d_t, \tag{3}$$

where p_{xing} is 52.3% [22] and $d_{\max}$ is the maximum distance the vehicle is defined to be away from the crosswalk. (The probability of 52.3% is calculated from Schroeder and Rouphail's [22] statistics on yield and non-yield events for a pedestrian waiting on the near side at site B.) Once within the crosswalk, the pedestrian is assumed to stay in the crosswalk for the next time step. The control loop assumes perfect information for the vehicle distance to the crosswalk, vehicle speed, and, for simplicity, the pedestrian posture. However, there is observation uncertainty for the pedestrian position with a false positive of 5%, which captures sensor uncertainty. These false positive rates were again chosen arbitrarily small but, in practice, would come from the perception system's capability of detecting pedestrians.

The goal is still for the automated vehicle to smoothly drive safely and efficiently through the crosswalk while adhering to the relevant traffic laws. The reward function defines the stage (or immediate) reward $g(x_t, u_t)$ for being in state x_t and taking action u_t, which again further connects the conceptualization values to the technical implementation. The reward for a state-action pair involves adding stage rewards (4), (5), and (6) for that state and action.

The stage reward for legality derives from the constant acceleration point mass equations relating the constant deceleration needed to come to a complete stop given the distance to the crosswalk and vehicle speed, and is as follows

$$g_{\text{legality}}(x_t, u_t) = -\zeta \frac{v_t^2}{d_t + \epsilon} \mathbf{1}(c_t), \tag{4}$$

where $\epsilon > 0$ is a buffer in the denominator to soften the constraint, and $\zeta > 0$ is a weight on the penalty incurred by driving quickly as the vehicle gets closer to the crosswalk.

The stage reward for safety is

$$g_{\text{safety}}(x_t, u_t) = -\eta \mathbf{1}(c_t \wedge d_t < 0), \tag{5}$$

where $\eta > 0$ is a terminal penalty independent of velocity to encourage the vehicle to stop when the pedestrian is crossing.

For mobility, the stage reward takes the form of

$$g_{\text{mobility}}(x_t, u_t) = \lambda v_t \mathbf{1}(\neg c_t) - \xi(a_{t-1} - a_t)^2, \tag{6}$$

where $\lambda > 0$ is a reward weight to encourage higher speed when the pedestrian is not crossing, and $\xi > 0$ is a penalty on large changes in acceleration.

To solve the POMDP, the QMDP approximation [23] is used again. In this iteration, vehicle speed increments in steps of 0.5 m/s, vehicle distance to crosswalk increments by 1 m, and accelerations are quantized by 0.5 m/s^2 intervals. These discretizations were chosen such that the sizes of the state and action spaces in the POMDP were kept small: 142,884 total states (including terminal states) and 27 possible actions.

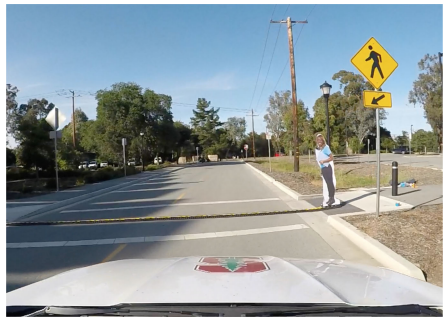

Fig. 2. Experimental setup of pedestrian crosswalk using a cardboard cutout for the pedestrian that moves along a track. Depicts the pedestrian posture of stopped.

C. Empirical Analysis

The empirical analysis for the second iteration focuses on experimental results. A policy comparison is not included in this analysis because, unlike the baseline, the state space of this POMDP cannot be fully represented in three dimensions. This is because the POMDP policies are conditioned on the previous acceleration command.

1) Experimental Results: Once again, in-vehicle experiments are conducted using a fully automated Ford Fusion. However, instead of using the lidar sensors and retro-reflective material for pedestrian detections, these experiments use computer vision to identify a pedestrian bounding box [24]. With the definition of a static polygon for the shape of the road, the pedestrian detection is used to determine whether the pedestrian is in the crosswalk influence area [22] (i.e., the sidewalk) or in the crosswalk. The vehicle is tasked with following a straight line path down the road still using a deterministic model predictive steering control [18]. The experimental scenario involves a pedestrian crosswalk on a two-lane roadway (Fig. 2). The vehicle is at speed when the pedestrian enters the crosswalk influence area, at which point the policy starts executing. As the vehicle approaches the crosswalk, the pedestrian may or may not transition into the crosswalk. The control algorithms have no prior knowledge as to whether or when the pedestrian will transition. The control algorithms also run at a constant rate of 100 Hz.

The baseline from the first iteration is used again for comparison against the new POMDP policies. It is considered an aggressive baseline because it will not yield to the pedestrian until they enter into the crosswalk. As an alternative, a conservative baseline is also considered, where the vehicle starts to yield to the pedestrian once they enter the crosswalk influence area. The policies are the same, except for what is considered to be the crosswalk influence area, which determines when to switch between cruise control and braking. However, they do not account for the pedestrian posture. Figures 3 and 4 depict the overhead driven trajectory, acceleration commands, and speed profile for the aggressive and conservative baselines, respectively. The circles indicate when the pedestrian was detected to be in the crosswalk by the computer vision algorithm. Since there is no circle in the aggressive baseline, the pedestrian never entered the crosswalk. The vehicle continued at the speed limit,

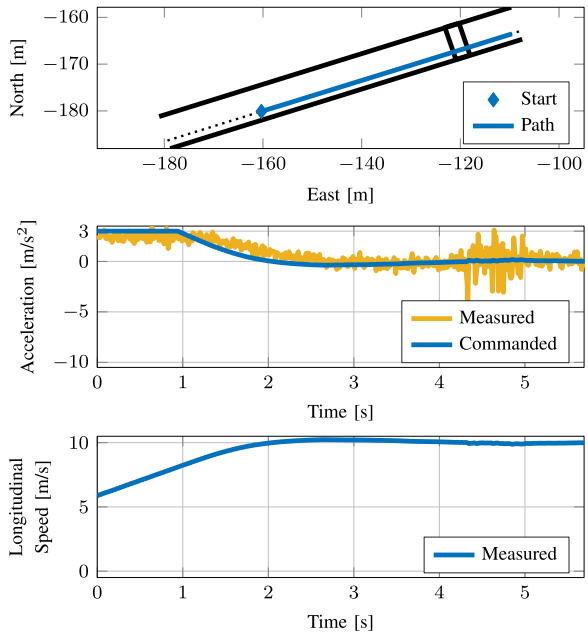

Fig. 3. Aggressive baseline trajectory overhead, acceleration command, and speed profile using deterministic speed control. There is no red circle because the pedestrian does not enter the crosswalk.

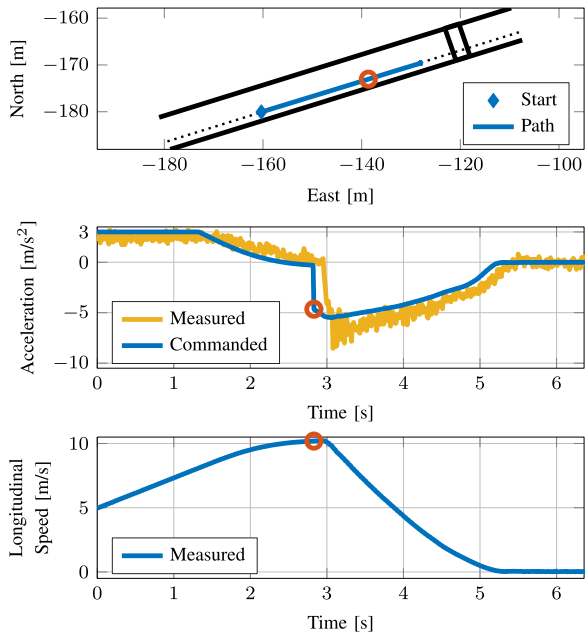

Fig. 4. Conservative baseline trajectory overhead, acceleration command, and speed profile using deterministic speed control (circle indicates when the pedestrian was detected).

never yielding to the pedestrian, because the pedestrian did not enter the crosswalk. For the conservative baseline, the perception system detected the pedestrian to be in the crosswalk influence area when the vehicle was 12.99 m away from the crosswalk, and the vehicle successfully yielded to the pedestrian.

For the policy execution of the POMDP, an observation of the vehicle speed, vehicle distance to crosswalk, pedestrian posture, and pedestrian location are used to update the belief of the pedestrian location with a Bayesian filter similarly to the previous iteration. Figures 5–8 depict the overhead driven trajectory, acceleration commands, and speed profile for the POMDP policies. The circles indicate when the pedestrian was detected in the crosswalk by the computer vision algorithm.

TABLE IV
WEIGHTS OF THE REWARD FUNCTION WITH RESPECT TO PEDESTRIAN POSTURE (p_t)

Variable	Pedestrian Posture (p_t) Weights			Unit
	DISTRACTED	WALKING	STOPPED	
Legality (ζ)	0.01	0	0.01	s^2/m
Buffer (ϵ)	8	8	8	m
Safety (η)	0.5	0.5	0.5	–
Mobility (λ)	0.05	0.1	0.03	s/m
Mobility (ξ)	0.003	0.01	0.003	s^2/m^2

In this second POMDP implementation, the pedestrian postures are independent of each other. Hence, a different set of weights are used in the reward function for each posture (Table IV). This partition is reasonable because each pedestrian posture is a unique sub-scenario that requires a different vehicle response. The numerical value for the buffer was chosen such that the numerator does not evaluate to zero and to limit the magnitude of the constant deceleration term. The other weights can be further tuned with additional analysis through Pareto optimization (see Section V) but are chosen preliminarily here to see if this design is satisfactory.

For the scenario of the distracted pedestrian (Fig. 5), the weights are chosen such that safety, efficiency, and smoothness are prioritized to similar normalized values: $\eta_{\text{n}} = 0.5$, $\lambda_{\text{n}} = 0.5$, and $\xi_{\text{n}} = 0.507$, respectively, at the extreme states and actions when $v_t = 10$ m/s, $d_t < 0$ m, and $a_{t-1} - a_t = 13$ m/s^2. The legality term is normalized to $\zeta_{\text{n}} = 0.125$ when $v_t = 10$ m/s and $d_t = 0$ m, suggesting lower prioritization. Fig. 5 shows that once the pedestrian enters the crosswalk influence area, the policy executes small negative accelerations to slow the vehicle down to around 1.5 m/s and to make it coast until the pedestrian enters the crosswalk. Once the pedestrian enters the crosswalk, the vehicle comes to a complete stop.

When the pedestrian is walking (Fig. 6), the terms for efficiency and smoothness increase to normalized values of $\lambda_{\text{n}} = 1$ and $\xi_{\text{n}} = 1.69$. With higher efficiency, the vehicle drives faster down the road and more smoothness is needed to smoothly decelerate the vehicle from the faster speed in case the pedestrian enters the crosswalk. Because of the high probability the pedestrian transitions to the crosswalk, there is consequently a high belief of 0.36 that the pedestrian is crossing. The impact of the safety and legality terms largely influence the vehicle behavior well before the pedestrian is physically within the crosswalk. To reduce the impact of the safety and legality terms, one of the terms is reduced to 0 (legality, in this instance) to allow the vehicle to progress towards the crosswalk. With these weights, the vehicle again coasts until the pedestrian is detected. As the vehicle gets closer to the crosswalk, it smoothly decelerates to a full stop.

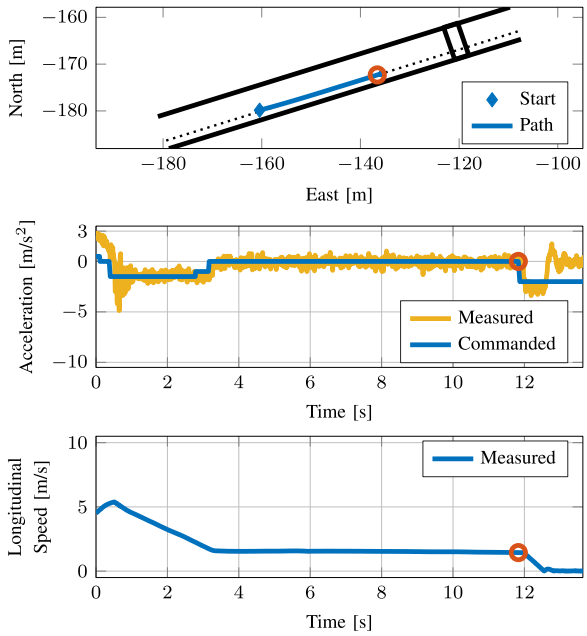

Fig. 5. Distracted pedestrian POMDP trajectory overhead, acceleration command, and speed profile using the belief of the pedestrian crossing (circle indicates when the pedestrian was detected).

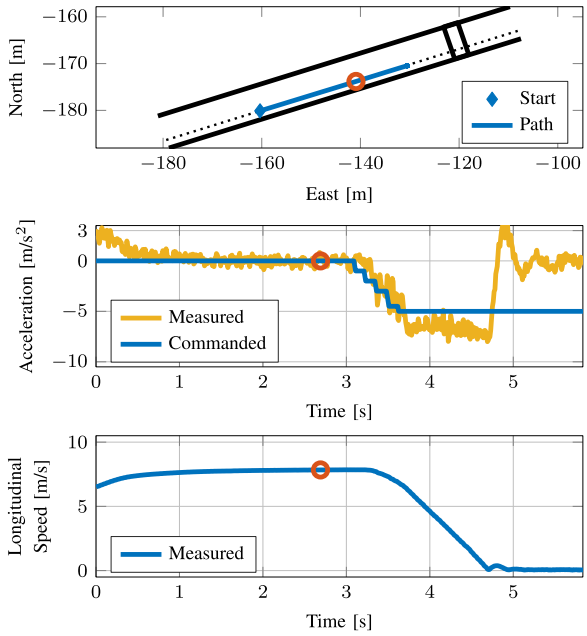

Fig. 6. Walking pedestrian POMDP trajectory overhead, acceleration command, and speed profile using the belief of the pedestrian crossing (circle indicates when the pedestrian was detected).

For the stopped pedestrian, the normalized legality term increases back to $\zeta_n = 0.125$ while efficiency and smoothness decrease to $\lambda_n = 0.3$ and $\xi_n = 0.507$. Smoothness is especially important in this scenario for the automated vehicle to demonstrate transparency about its intentions to move through the environment, and hence is chosen to have the highest prioritization. Using these weights, two different scenarios were tested. The scenario depicted in Fig. 8 shows the vehicle gradually

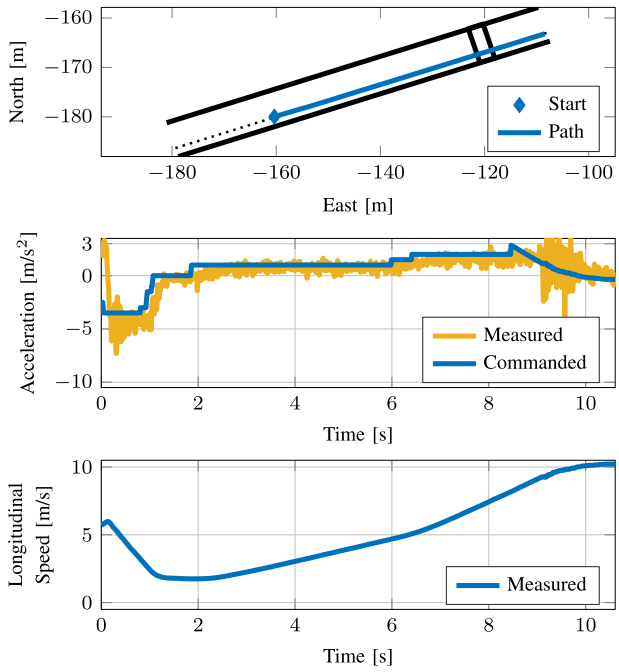

Fig. 7. Stopped pedestrian POMDP trajectory overhead, acceleration command, and speed profile using the belief of the pedestrian crossing. There is no red circle because the pedestrian does not enter the crosswalk.

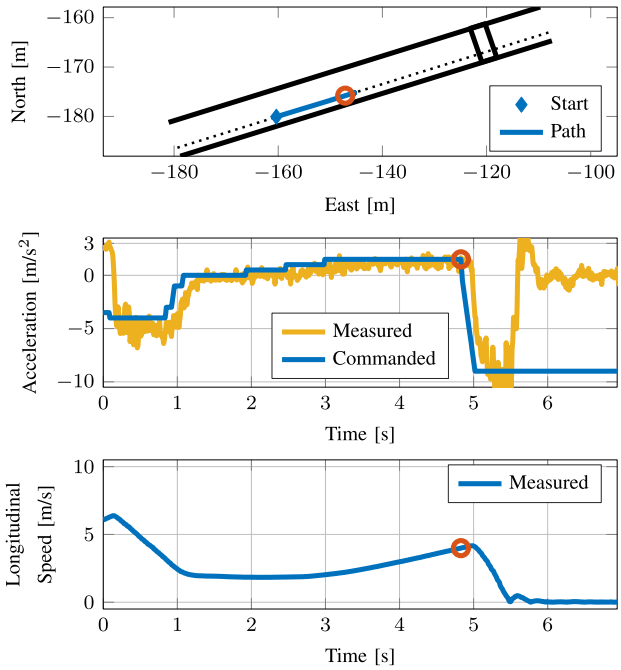

Fig. 8. Stopped pedestrian POMDP trajectory overhead, acceleration command, and speed profile using the belief of the pedestrian crossing (circle indicates when the pedestrian was detected). Pedestrian takes right of way.

accelerating as the vehicle gets closer to the crosswalk because of decreasing belief that the pedestrian will cross the street. Since the pedestrian does not want to cause an immediate hazard, it does not enter the crosswalk. In the second stopped pedestrian scenario, Fig. 8 portrays the pedestrian crossing the street before

the vehicle accelerates too much. Once the pedestrian is in the crosswalk, the vehicle comes to a complete stop.

With the focus on pedestrian behavior in this second iteration, the scenario inherently gains complexity around the value tension between the pedestrian's intent and the automated vehicle's desire to travel down the road. The weights chosen here still represent arbitrary trade-offs over the value statements. Hence, a deeper analysis is likely appropriate at this point to better determine if a particular design point can resolve the value tensions. For example, a Pareto optimization would be useful to simulate over many scenarios as the choice of weights change as demonstrated in Section V.

D. Lessons Learned

This second iteration of the speed control design demonstrates improvements in handling the value conflicts as the engineering specifications refine in terms of the identified values. There are still components of the implementation to highlight and some aspects to improve because this is not the final product.

Positive Outcomes:

- In all scenarios, accounting for pedestrian uncertainty allowed the vehicle to successfully yield to the pedestrian. The design choice of modeling the problem as a POMDP meant dynamic programming could be used to account for future state information in the policy.
- The only information about the pedestrian used was whether he or she was in the crosswalk and what posture he or she composed. This continued to uphold the values of fairness and reciprocity.
- The continued design decision to model the problem as a POMDP and solve for an offline policy helped to investigate and balance some of the value tensions in this design task. Although the actual choice of weights were arbitrary, it demonstrated the potential for the value tensions to be resolved.
- Smoothness improved by penalizing change in acceleration, and its influence corresponded directly with ξ.
- Efficiency continued to correspond to the term λ.
- Modeling the pedestrian as a function of posture gave insight into pedestrian intent and crossing the street.
  - For the stopped pedestrian, the vehicle slowly increased its speed as it approached the crosswalk to indicate to the pedestrian it will yield if they enter the crosswalk while also indicating to the pedestrian that the vehicle wants to travel down the road.
  - The random probability for the distracted pedestrian allowed the vehicle to approach the crosswalk at a very cautious speed.

Outcomes to Improve:

- Pedestrian modeling needs to be improved
  - Pedestrian posture and position did not capture all the attributes relevant to the pedestrian's intent to cross the street. Other parameters could be considered while keeping in mind the values of fairness and reciprocity to mitigate use of biased information or discriminatory actions by the vehicle.
  - There is likely some correlation between distraction and motion for the pedestrian. Other pedestrian models could be considered to further study nuances in pedestrian posture and motion.
  - The pedestrian transitions assume the pedestrian will stay within the crosswalk once they enter the crosswalk. More exploration into modeling the transition from the crosswalk to the sidewalk (or safe distance from the automated vehicle's traveling lane) should be considered.
- The vehicle tended to stop short of the crosswalk, which could be a result of poor choice of weights. Alternatively, the reward function could adjust, i.e., add a slight penalty on large decelerations so the vehicle only comes to a full stop when necessary. Stopping short impacts the engineering specification of mobility and efficiency.
- In the situation with the moving pedestrian, the high likelihood of transitioning greatly increased the influence of the safety and legality terms on the policy. These weights either need to be tuned down significantly or an alternative formulation should be considered to better isolate the impact of safety and legality.
- Future iterations should consider analyzing scenarios when the pedestrian is not present or not detected.
- Future iterations should also consider aspects to deploy on public roads, such as scalability of the technology.

If another iteration were to occur (this paper only presents two iterations), then it would investigate how to maintain these positive attributes while addressing some of the downsides of this second implementation. Further investigation into the choice of weights in the reward function is also needed in order to determine how well mobility, safety, and legality can be realized with this implementation. Such an investigation could be done with a Pareto, or multi-objective, optimization over the weights, for example, and is demonstrated in Section V for the first iteration.

V. Closing the Loop on Human Values

The policy comparison and experimental results demonstrate a potentially reasonable speed control algorithm design, but only for a particular set of weights. The behavior of the vehicle can greatly vary depending on the choice of weights in the reward function. To analyze how well the designed technology aligns with the stakeholder values, an analysis technique is needed. One way to perform this analysis is with the technique of Pareto (or multi-objective) optimization in order to determine which set of weights best resolve the value tensions. A design is Pareto optimal if one objective cannot improve without worsening at least one other objective. To construct a frontier of Pareto optimal points, the design objectives map to a criterion space using evaluation criteria. The determination of Pareto optima serves to "close the loop" on the design process: human values map to engineering objectives, engineering objectives map to evaluation criteria, and evaluation criteria map to human values. Thus, engineers can focus on Pareto optimal designs without committing to a particular prioritization between objectives ahead of time.

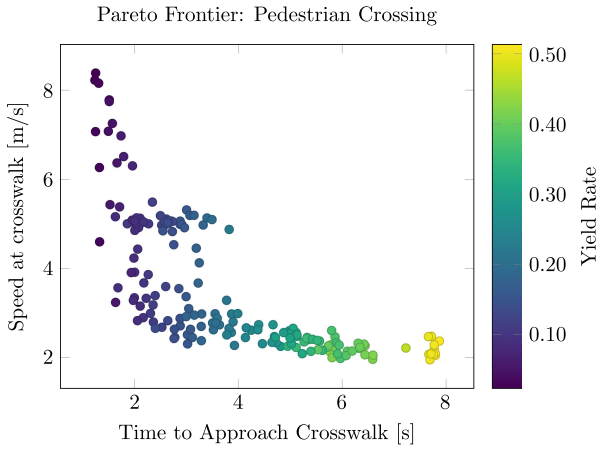

Fig. 9. Pareto frontier of POMDP for various weights mapped to evaluation criteria.

An example of a Pareto frontier for the first VSD iteration is constructed by varying the weights in the reward function that correspond to the engineering objectives. For each combination of weights in the reward function, a different optimal policy is generated. For each given optimal policy, Monte Carlo simulations [25] are run, and the simulation results are averaged to calculate the evaluation criteria. The Monte Carlo simulations have the pedestrian suddenly appear from behind the occluding vehicle at random times whenever the automated vehicle is within 20 m of the crosswalk. The simulations include the assumption that the pedestrian requires about 4 s to cross the street, meaning the simulation terminates when the vehicle passes the crosswalk or the pedestrian completes crossing. For the evaluation criteria, the objective of safety and legality maps to the criterion of the vehicle velocity at the crosswalk. The objective of efficiency maps to the criterion of average time to complete maneuver. The objective of smoothness maps to the criterion of average maximum change in acceleration.

The resulting Pareto frontier can then be brought back to a larger group of stakeholders, such as policymakers, lawyers, and public interest groups, to determine which set of weights to deploy on the automated vehicle. Figure 9 shows an example of a slice of the Pareto frontier for safety and legality (average speed at the crosswalk) vs. mobility (average time to approach the crosswalk). It is additionally colored by the yield rate for the simulated sudden pedestrian scenarios. The larger group of stakeholders can confer the Pareto frontiers to additional information, such as injury curves [26], user studies, emissions curves [27], and congestion studies [28].

Pareto optimization is not the only tool that can help close the loop on human values. Another utility-based analysis tool could be a risk management or cost-benefit analysis for a set of outcomes [29]. Or maybe a deontological-like analysis [30] is more desirable where thresholds or conditions are determined by policymakers or by re-engaging with stakeholders. This Pareto analysis is a first step for demonstrating how some analysis tools can help determine how successfully a technical implementation embodies the human values identified in the conceptualization phase.

VI. Conclusion

This paper demonstrates the formal connection of human values into the design of a speed control algorithm through the conceptualization and technical implementation phases. The empirical analysis phase helps identify areas of improvement for subsequent iterations. In the first iteration, a POMDP is chosen to help realize the values of safety and legality, efficiency and mobility, and smoothness for a scenario with a large vehicle parked in front of a pedestrian crosswalk. The POMDP helped represent the uncertainty in the situation and allowed the vehicle to be proactive by approaching the crosswalk at a reasonable speed, which resulted in adequate yielding to pedestrians that appear suddenly. The pedestrian model in the first iteration is very simple, but the second iteration improves the pedestrian model in order to analyze the value tension between the pedestrian and automated vehicle. In the second iteration, the values of legality, safety, efficiency, and smoothness were refined in terms of the technical implementation. Additional analysis with Pareto optimization provides further insight into how well an implementation aligns with the identified values. Iterating through VSD helps engineers study how human values are implicated in the technology as it develops.

There are of course many applications of this methodology to automated vehicle design beyond speed control at crosswalks. Problems such as designing automated maneuvers around bicyclists, setting safe following distances for platooning trucks, designing equitable routing algorithms for dispatching automated vehicles or balancing human and machine inputs in driver assistance systems could potentially benefit from such an approach.

Acknowledgment

The authors would like to thank Dr. J. Millar for encouraging the pursuit of value sensitive design, L. Cathey for helping construct the pedestrian crosswalk rig, and members of the Dynamic Design Lab for helping run the in-vehicle experiments on campus.

References

[1] NHTSA's National Center for Statistics and Analysis, "How vehicle age and model year relate to driver injury severity in fatal crashes," U.S. Dept. Transp, Washington, DC, USA, Tech. Rep. DOT HS 811 825, 2013.
[2] *Federal Automated Vehicles Policy*, U.S. Department of Transportation, Washington, DC, USA, 2016.
[3] M. Maguire, "Methods to support human-centred design," *Int. J. Human Comput. Studies*, vol. 55, no. 4, pp. 587–634, 2001.
[4] J. Giacomin, "What is human centred design?" *Des. J.*, vol. 17, no. 4, pp. 606–623, 2014.
[5] J. Millar, "An ethics evaluation tool for automating ethical decision-making in robots and self-driving cars," *Appl. Artif. Intell.*, vol. 30, no. 8, pp. 787–809, 2016.
[6] M. Niemelä, V. Ikonen, J. Leikas, K. Kantola, M. Kulju, A. Tammela, and M. Ylikauppila, "Human-driven design: A human-driven approach to the design of technology," in *Proc. IFIP Int. Conf. Human Choice Comput.*, 2014, pp. 78–91.
[7] J. Leikas, P. Saariluoma, J. Heinilä, and M. Ylikauppila, "A methodological model for life-based design," *Int. Rev. Soc. Sci. Humanities*, vol. 4, no. 2, pp. 118–136, 2013.
[8] B. Friedman and P. H. Kahn, "Human values, ethics, and design," in *The Human-Computer Interaction Handbook*, J. A. Jacko and A. Sears, Eds. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 2003, pp. 1177–1201.

[9] B. Friedman, P. H. Kahn, and A. Borning, "Value sensitive design and information systems," in *Human-Computer Interaction and Management Information Systems*, P. Zhang and D. Galletta, Eds. Armonk, NY, USA: M.E. Sharpe, 2006, vol. 5, pp. 348–372.
[10] A. Borning and M. Muller, "Next steps for value sensitive design," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2012, pp. 1125–1134.
[11] B. Friedman, P. H. Kahn, and A. Borning, "Value sensitive design: Theory and methods," Univ. Washington, Seattle, WA, USA, Tech. Rep. 02-12-01, 2002.
[12] T. Denning, A. Borning, B. Friedman, B. T. Gill, T. Kohno, and W. H. Maisel, "Patients, pacemakers, and implantable defibrillators," in *Proc. Int. Conf. Human Factors Comput. Syst.*, 2010, pp. 917–926.
[13] A. van Wynsberghe, "Designing robots for care: Care centered value-sensitive design," *Sci. Eng. Ethics*, vol. 19, no. 2, pp. 407–433, 2013.
[14] B. Chen, D. Zhao, and H. Peng, "Evaluation of automated vehicles encountering pedestrians at unsignalized crossings," in *Proc. IEEE Intell. Vehicles Symp.*, 2017, pp. 1679–1685.
[15] T. Bandyopadhyay, C. Z. Jie, D. Hsu, M. H. Ang, D. Rus, and E. Frazzoli, "Intention-aware pedestrian avoidance," in *International Symposium on Experimental Robotics*, J. P. Desai, G. Dudek, O. Khatib, and V. Kumar, Eds. Berlin, Germany: Springer, 2013, pp. 963–977.
[16] T. Bandyopadhyay, K. Won, E. Frazzoli, D. Hsu, W. Lee, and D. Rus, "Intention-aware motion planning," in *Algorithmic Foundations of Robotics X*, Berlin, Heidelberg: Springer-Verlag, 2013, pp. 475–491.
[17] S. Brechtel, T. Gindele, and R. Dillmann, "Probabilistic decision-making under uncertainty for autonomous driving using continuous POMDPs," in *Proc. IEEE Conf. Intell. Transp. Syst.*, 2014, pp. 392–399.
[18] S. M. Thornton, V. Zhang, S. Varnhagen, and J. C. Gerdes, "Comparative analysis of steering system models in model predictive control of automated vehicles," in *Proc. Int. Symp. Adv. Vehicle Control*, 2018.
[19] S. M. Thornton, F. E. Lewis, V. Zhang, M. J. Kochenderfer, and J. C. Gerdes, "Value sensitive design for autonomous vehicle motion planning," in *Proc. IEEE Intell. Vehicles Symp.*, 2018, pp. 1157–1162.
[20] J. Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. New York, NY, USA: Vintage, 2012.
[21] J. K. Choi and Y. G. Ji, "Investigating the importance of trust on adopting an autonomous vehicle," *Int. J. Human Comput. Interact.*, vol. 31, no. 10, pp. 692–702, 2015.
[22] B. J. Schroeder and N. M. Rouphail, "Event-based modeling of driver yielding behavior at unsignalized crosswalks," *J. Transp. Eng.*, vol. 137, no. 7, pp. 455–465, 2011.
[23] M. J. Kochenderfer, *Decision Making Under Uncertainty: Theory and Application*. Cambridge, MA, USA: MIT Press, 2015.
[24] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc, IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6517–6525.
[25] N. T. Thomopoulos, *Essentials of Monte Carlo Simulation: Statistical Methods for Building Simulation Models*. Berlin, Germany: Springer Science & Business Media, 2013.
[26] B. C. Tefft, "Impact speed and a pedestrian's risk of severe injury or death," *AAA Found. Traffic Safety*, 2011.
[27] P. S. Bokare and A. K. Maurya, "Study of effect of speed, acceleration and deceleration of small petrol car on its tail pipe," *Int. J. for Traffic Transport Eng.*, vol. 3, no. 4, pp. 465–478, 2013.
[28] F. Soriguera, I. Martínez, M. Sala, and M. Menéndez, "Effects of low speed limits on freeway traffic flow," *Transp. Res. Part C, Emerg. Technol.*, vol. 77, pp. 257–274, 2017.
[29] N. J. Goodall, "Away from trolley problems and toward risk management," *Appl. Artif. Intell.*, vol. 30, no. 8, pp. 810–821, 2016.
[30] S. M. Thornton, "Autonomous vehicle motion planning with ethical considerations," Ph.D. dissertation, , Stanford University, Stanford, CA, USA, 2018.

Sarah M. Thornton received the B.Sc. degree in mechanical engineering from UC Berkeley, Berkeley, CA, USA, in 2011, the M.Sc. degree in mechanical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2013, and the Ph.D. degree in mechanical engineering from Stanford University, Stanford, CA, USA, in 2018. Her research focused on developing an adaptive shift control algorithm for automatic transmissions and in the area of ethical and uncertain decision making for automated vehicles.

Benjamin Limonchik was born in Jerusalem, Israel. He received the B.S. degree in science and computer science and the M.S. degree in management science and engineering with a concentration in artificial intelligence from Stanford University, Stanford, CA, USA. Between 2010 and 2013, He was with the Intelligence in the Israeli Defense Forces. He is currently working as a Software Developer with ThoughtSpot's Search and Data Analytics team.

Francis E. Lewis received the B.S. degree in computer science from Stanford University, Stanford, CA, USA, in 2017, where he is currently working toward the M.S. degree in computer science. He is affiliated with the Dynamic Design Lab, where his current research interests focus on exploring the interplay between multi-modal perception and vehicle control.

Mykel J. Kochenderfer is an Assistant Professor of Aeronautics and Astronautics and Assistant Professor, by courtesy, of Computer Science with Stanford University, Stanford, CA, USA. He is the Director of the Stanford Intelligent Systems Laboratory, conducting research on advanced algorithms and analytical methods for the design of robust decision making systems. Of particular interest are systems for air traffic control, unmanned aircraft, and automated driving where decisions must be made in uncertain, dynamic environments while maintaining safety and efficiency.

J. Christian Gerdes received the Ph.D. degree from UC Berkeley, Berkeley, CA, USA, in 1996. He is a Professor with the Department of Mechanical Engineering and, by courtesy, with the Department of Aeronautics and Astronautics, Stanford University, Stanford, CA, USA. He is the Director of the Center for Automotive Research, Stanford University. Prior to joining Stanford, he was the Project Leader for virtual proving grounds development with the Vehicle Systems Technology Center, Daimler-Benz Research and Technology, Portland, OR, USA.