REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)
24-12-2019

2. REPORT TYPE
Final Report

3. DATES COVERED (From - To)
26-Sep-2018 - 25-Sep-2019

4. TITLE AND SUBTITLE
Final Report: Developing and Signaling Trust in Synthetic Autonomous Agents (SAAs)

5a. CONTRACT NUMBER
W911NF-18-2-0300

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES

Arizona State University
ORSPA
P.O. Box 876011
Tempe, AZ 85287 -6011

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
74016-LS-DRP.2

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Kathryn Johnson
a. REPORT	b. ABSTRACT	c. THIS PAGE			
UU	UU	UU	UU		19b. TELEPHONE NUMBER 480-965-7598

RPPR Final Report
as of 02-Jan-2020

Agency Code:

Proposal Number: 74016LSDRP **Agreement Number: W911NF-18-2-0300**

INVESTIGATOR(S):

Name: Kathryn Ann Johnson
Email: Kathryn.A.Johnson@asu.edu
Phone Number: 4809657598
Principal: Y

Organization: **Arizona State University**
Address: ORSPA, Tempe, AZ 852876011
Country: USA

DUNS Number: 943360412 EIN: 860196696
Report Date: 25-Oct-2019 Date Received: 24-Dec-2019
Final Report for Period Beginning 26-Sep-2018 and Ending 25-Sep-2019

Title: Developing and Signaling Trust in Synthetic Autonomous Agents (SAAs)

Begin Performance Period: 26-Sep-2018 **End Performance Period:** 25-Sep-2019
Report Term: 0-Other
Submitted By: Kathryn Johnson Email: Kathryn.A.Johnson@asu.edu
Phone: (480) 965-7598
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 4 **STEM Participants:** 6

Major Goals: Goal 1. The primary goal of this one year research project was to draw on social psychological research in order to specify the morals and values of good drivers that may be available for programming SAAs to make decisions and behave with moral integrity.
Goal 2. Our second goal was to begin to test the feasibility of programming value-governed parameters of SAAs, in a newly developed, four-wheel, skid-steer robotic car that resembles a 1:28 scale self-driving car which we refer to as a Go-CHART.
Goal 3. Our third goal was to identify the most efficacious signal of programmed moral integrity in order to garner appropriate trust from human operators and the general public.

Synthetic Autonomous Agents (SAAs; e.g., self-driving cars, unmanned search and rescue vehicles, lethal autonomous weapons) can accomplish tasks too difficult or risky for humans and we must not fail in preparing for this advancing technology. Yet opponents argue that SAAs should never be developed and, instead, humans must maintain meaningful human control (Roff & Moyes, 2016) in every case because SAAs may fall into enemy hands, become disconnected from their human counterparts, or may initiate undesirable outcomes. One way to overcome this distrust of autonomous agents is to ensure that SAAs behave with moral integrity.

Whether or not SAAs are deemed to be true moral agents, we contend they can be programmed to make decisions and to behave as responsible moral agents. To date, morality has generally been conceptualized as either deontological (following rules regardless of the outcome) or utilitarian (accomplishing a worthy goal). However, the two systems often conflict, require the programming of all possible rules or outcomes, and people rarely agree about which system is best (Awad, et al., 2018) (Conway & Gawronski, 2013). As one example, people agree that self-driving cars should never drive on sidewalks (deontological). Given the classic trolley car problem, empirical studies show nearly all people agree that a child's life should be spared over an adult's life (utilitarian). However, it is not clear whether a self-driving car should drive onto a sidewalk and, incidentally, kill an adult to avoid killing a child darting into the street (Bergmann, et al., 2018).

More importantly, the philosophical debate over deontological versus utilitarian ethics has neglected the extensive work in social psychology regarding moral priorities (Graham, et al., 2011) and values (Schwartz, et al., 2012) that ground human decision-making processes. We proposed that SAAs programmed with the values and moral priorities of the most trust-worthy drivers among us would provide safeguards when meaningful human control is lacking (e.g., in complex or dynamic environments where human decision-making may be too slow) and, consequently, increase trust.

RPPR Final Report
as of 02-Jan-2020

Accomplishments: Goal 1 Accomplishments

We surveyed over 1,000 automobile drivers in the US using existing measures to identify the underlying values (Schwartz, 1992; Schwartz, et al., 2012) and moral priorities (Graham et al., 2011) associated with driving style (Ozkan & Lajunen, 2005; Reason, Manstead, Stradling, Baxter, & Campbell, 1990), number of accidents and traffic citations, and self-sacrifice vs. self-preservation decisions in various crash scenarios (Awad, et al., 2018). In Study 1, we found that a moral priority of caring for others and the values of benevolence and self-directedness (e.g., autonomy and adaptability) were positive predictors of good driving practices, whereas the value of power was a positive predictor of aberrant driving (i.e., errors, violations, aggressive driving).

In Study 2, we confirmed that people honestly report having received traffic citations and that aberrant driving practices predicted having received traffic citations. In Study 3, we replicated our finding that the value of benevolence was associated with good driving and also found that benevolence predicted self-sacrifice in trolley-dilemma crash scenarios. Also replicating Studies 1 and 2, we found that the value of power was associated with aberrant driving and traffic citations. Self-directedness was also positively associated with good driving and negatively with aberrant driving. A manuscript presenting the results of these three studies and a statistical model of our results (see Figure 1, above) have been submitted for publication.
Additionally, in follow up studies, we found that individuals who valued benevolence say they would drive at significantly lower speeds relative to those who value power when driving through a neighborhood when children are present.
In summary, we found that benevolence, self-directedness, and power were critical values in predicting driving style and crash decisions among humans, and these values became the basis for pilot testing moral decision-making in simulated and robot driving scenarios.

Goal 2 Accomplishments

Some have suggested that SAAs programmed with artificial neural networks could learn how to respond in diverse situations by training the neural networks on hundreds of driving and crash scenarios (Gerdes & Thornton, 2015). Training would include human feedback regarding the appropriateness of the AV's response. This bottom-up learning approach would be constrained by hard-coded (deontological) rules for driving and crashing. However, as previously discussed, there is no agreement among US citizens—much less the global community—regarding best driving and crashing behaviors (Awad et al., 2018; Basu et al., 2017; Bergmann et al., 2018; Bonnefon et al., 2016). We propose that one solution would be to reinforce training behaviors to align with the values of ideal operators (e.g., good drivers) in relevant cultures rather than the general public; the values of ideal operators could and should inform reinforcement learning strategies in training any type of SAA.
A second possibility is that, in ambiguous or uncertain circumstances, the values of benevolence and power can be used to set decision-making parameters, which we will refer to as value-governed parameters, in SAA programs (e.g., weights on objectives or constraints in optimization problems, gains in feedback controllers) that cause the vehicle to exhibit the behavioral characteristics and decision-making strategies that we have come to expect from good drivers as responsible moral agents. The value-governed parameters might vary depending upon the type of vehicle. For example, Gerdes and Thornton (2015) suggest that an automated rideshare may prioritize benevolence for the comfort of the passengers, whereas an ambulance may prioritize power for a quick response to an emergency.
We recognize that value-governed parameters will need to be constrained by legal limits, liability concerns, overriding goals of avoiding collisions, and the prime directive to do no harm to humans. Similarly, Gerdes and Thornton (2015) recommended that an AV operate according to a hierarchy of constraints according to the priorities of the vehicle (a deontological ethics system) until a dilemma, such as a crash scenario, is reached. In the case of a dilemma, certain constraints may then be violated, according to their predetermined hierarchy, if the predicted outcome is sufficiently beneficial to justify their violation (a consequentialist ethics system).

To begin to test the feasibility of programming value-governed parameters of SAAs, we developed a novel four-wheel, skid-steer robot that resembles a 1:28 scale standard commercial sedan, which we refer to as a Go-CHART (Kannapiran and Berman, 2019a and 2019b). The Go-CHART is equipped with onboard sensors and both onboard and external computers that replicate many of the sensing and computation capabilities of a full-size AV. It can autonomously navigate our small-scale traffic testbed (see Figure 2, uploaded), an updated version of our previous testbed (Subramanyam, 2018), by responding to its sensor input with programmed controllers. Alternatively, it can be remotely driven by a user who views the testbed through the robot's four camera feeds, which facilitates safe, controlled experiments on driver interactions with driverless vehicles. We have demonstrated the Go-CHART's ability to perform lane tracking and detection of traffic signs, traffic signals, and other Go-CHARTs in real-time,

utilizing an external graphics processing unit (GPU) that runs computationally intensive computer vision and deep learning algorithms.
In summary, the Go-CHART can be used to investigate the transferability of human morals and values to programs that control AVs. For example, we can design Go-CHART controllers that reproduce human driver tendencies to speed up when task completion is prioritized (a proxy for power) and slow down when care for pedestrians is prioritized (a proxy for benevolence).

Goal 3 Accomplishments

Our ultimate goal is to facilitate appropriate trust in SAAs by providing assurances of moral integrity. Therefore, it will be important for morally programmed cars (and other types of SAAs) to signal trustworthiness to operators and observers. One solution may be to brand AVs with a logo signaling trustworthiness. To that end, we tested the efficacy of various brands, badges, and icons as visible signals of trust-worthiness.

Participants were shown a self-driving car with no logo followed by five randomly selected images with various icons: certification checkmark (see Figure 3 below and other exemplars, uploaded), an olive branch, a watching eye, a globe, and the Waymo logo. Participants rated the trustworthiness (i.e., competence, benevolence, and integrity; Mayer, Davis, & Schoorman, 1995) of the vehicle on a Likert scale rating of 1 to 7. Planned contrasts revealed that the car with no logo (M = 4.10, SE = .097) was rated as significantly more trustworthy than the checkmark (M = 4.04, SE = .103), olive branch (M = 3.96, SE = .099), eye (M = 3.82, SE = .100), globe (M = 3.99, SE = .101), or Waymo logo (M = 3.96, SE = .103).
We found that external cues (e.g., car logos and brands) may not be useful for conveying trustworthiness of AVs. One hypothesis that should be investigated in future research is that individuals rely on behavior and the attribution (whether accurate or not) of internal states (e.g., a mind) when evaluating the trustworthiness of non-human agents (e.g., AVs).

Training Opportunities: During the course of the project, Drs. Johnson and Berman mentored three graduate students. The students did not receive direct financial support from project funds.

Immanuella Kankam adapted an open source computer simulation program (CARLA) in order to conduct the driving simulation and crash scenarios used to investigate values as a predictor of moral decision-making. This work led to the successful defense of her master's thesis, Design of an Immersive Virtual Environment to Investigate How Different Drivers Crash in Trolley-Problem Scenarios, and partial fulfillment of the requirements of her master's degree in engineering.

Sangeet Sankaramangalam Ulhas investigated the use of cross platform training for neural networks transferring an object detection model from a virtual to a physical environment. This work led to the successful defense of his master's thesis, Cross Platform Training of Neural Networks to Enable Object Identification by Autonomous Vehicles, and partial fulfillment of the requirements of his master's degree in engineering.

Shenbagaraj Kannapiran developed the Go-CHART used in pilot testing the utility of programming human values of benevolence and power as value-governed parameters in self-driving cars. His work resulted in a paper presentation, Go-CHART: A Miniature Remotely Accessible Self-driving Car Robot, currently under review at the International Conference on Robotics and Automation (ICRA).

In addition, three undergraduate research assistants helped with various aspects of the project: Natalie Beaulieu, Daniel Shuster, and Hunter Larkins.

Professional development

Dr. Johnson conducted three colloquia presenting the results of this research and discussing future directions and related research with attendees: Social and Personality Research Institute Meeting (October, 2018), PolyTechnic Brown Bag (March, 2019), and the Aviral Shrivastava Weekly Lab Meeting (May, 2019).

Additionally, we added an important member to our interdisciplinary research team of social psychologists and engineers, Dr. Ted Pavlic, Assistant Professor in ASU's School of Computing, Informatics, and Decision Systems Engineering and ASU's School of Sustainability.

Results Dissemination: We want to ensure that our findings from our surveys and experiments on the small-scale testbed will transfer to full-scale vehicles on community roadways and meet the needs and desires of community stakeholders. Therefore, we have submitted a proposed planning grant to the National Science Foundation (Smart and Connected Communities) that would facilitate our meeting with industry representatives and community leaders to (1) present our research findings, (2) better understand the enthusiasm and/or concerns about SAAs within the community, and (3) obtain feedback regarding the scalability of programming moral integrity in autonomous systems such as autonomous vehicles.

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Kathryn A. Johnson
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Co-Investigator
Participant: Spring Berman
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Co-Investigator
Participant: Erin Chiou
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Co-Investigator
Participant: Adam B. Cohen
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

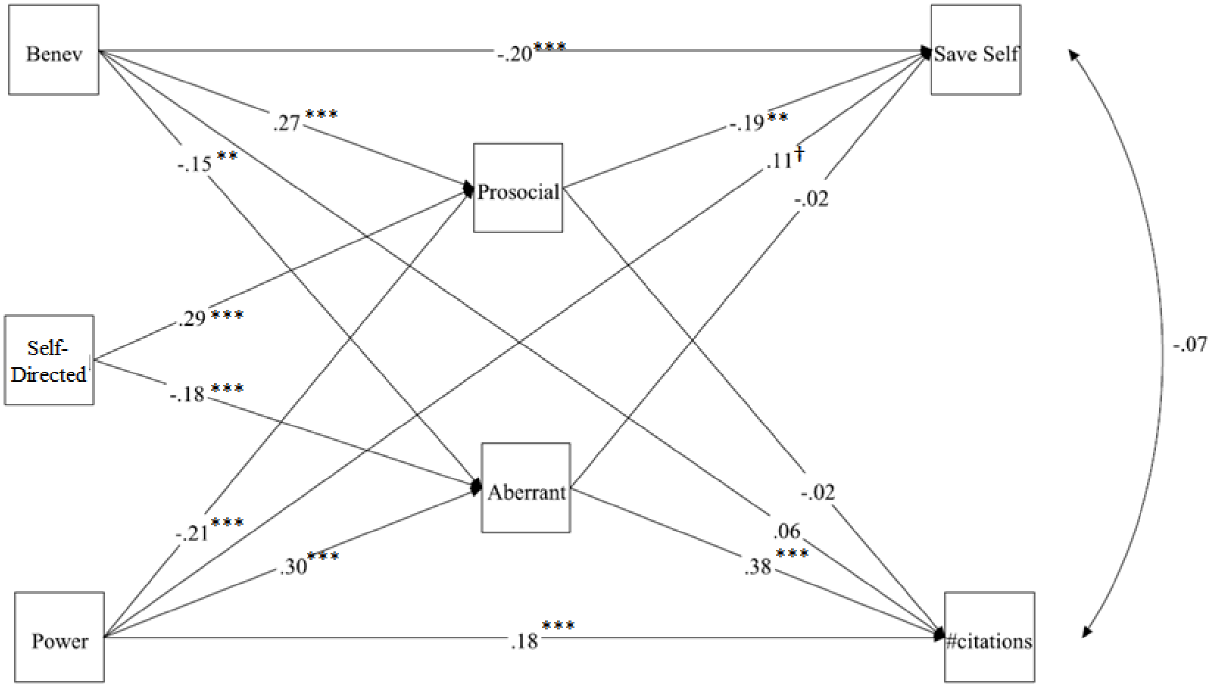
Participant Type: Faculty
Participant: Theodore P. Pavlic

Person Months Worked: 12.00

Funding Support:

Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Figure 1. Summary model and path analysis demonstrating the effects of the values of benevolence, self-directedness, and power on driving practices, crash decisions, and citations



Note: $^{***}p \leq .001$; $^{**}p \leq .01$; † $p = .066$; Benev = priority of the value of benevolence; Self-directed = priority of the value of self-directedness (adaptability); Power = priority of the value of power; Prosocial = Positive driving practices; Aberrant = Errors, Violations, and Aggressive driving; Save Self = Self-preservation in crash dilemmas; #citations = Number of traffic citations in past three years.

Figure 2. *Small-scale CHARTOPOLIS driving testbed and a Go-CHART robot*

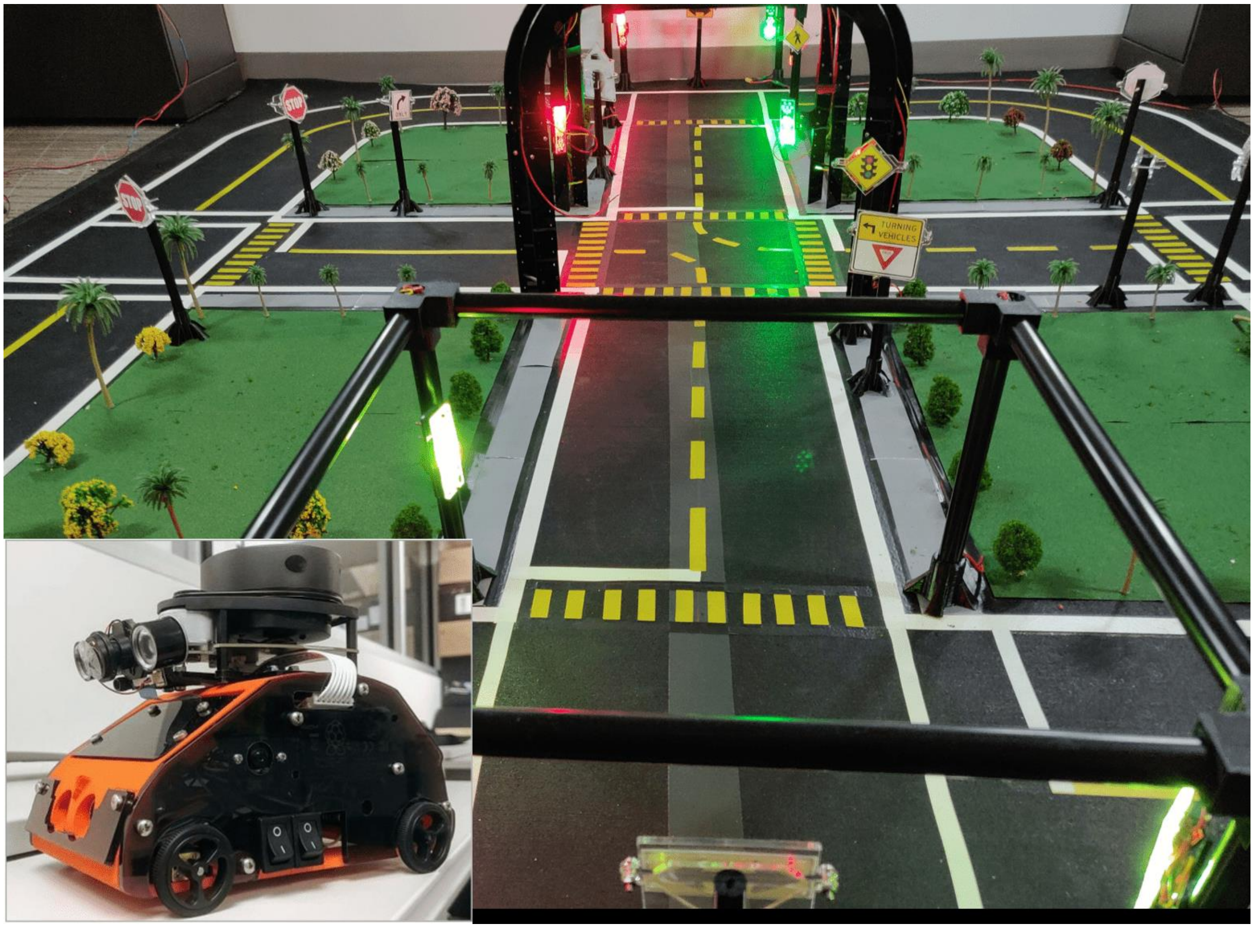


Figure 3. *Potential icons signaling the trustworthiness of self-driving cars*

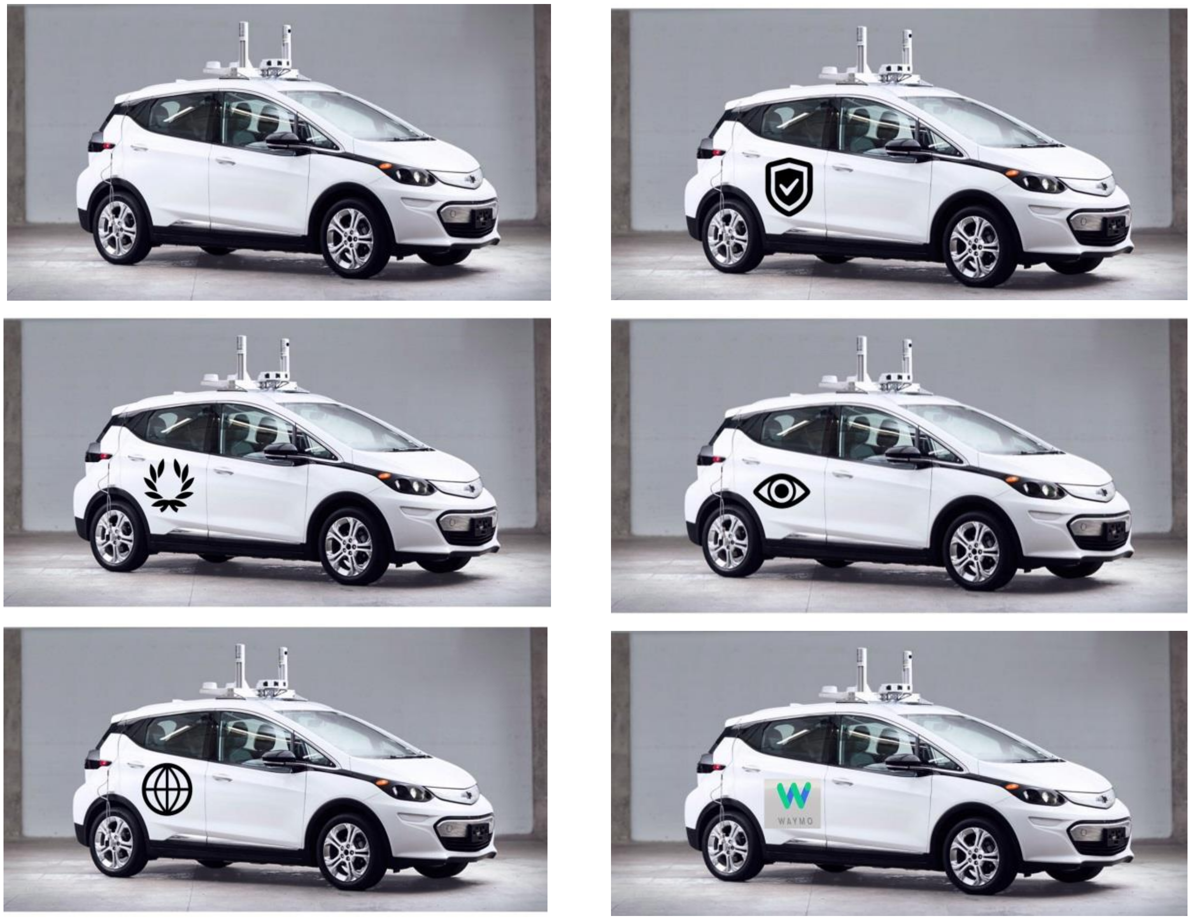


Figure 4. *Driving simulator*