Human Drivers' Situation Awareness of Autonomous Driving Under Physical-world Attacks

Katherine S. Zhang
Purdue University
zhan3461@purdue.edu

Claire Chen
The Pennsylvania State University
ckc5857@psu.edu

Aiping Xiong
The Pennsylvania State University
axx29@psu.edu

Abstract—**Artificial intelligence (AI) systems in autonomous driving are vulnerable to a number of attacks, particularly the physical-world attacks that tamper with physical objects in the driving environment to cause AI errors. When AI systems fail or are about to fail, human drivers are required to take over vehicle control. To understand such human and AI collaboration, in this work, we examine 1) whether human drivers can detect these attacks, 2) how they project the consequent autonomous driving, 3) and what information they expect for safely taking over the vehicle control. We conducted an online survey on Prolific. Participants ($N = 100$) viewed benign and adversarial images of two physical-world attacks. We also presented videos of simulated driving for both attacks. Our results show that participants did not seem to be aware of the attacks. They overestimated the AI's ability to detect the object in the dirty-road attack than in the stop-sign attack. Such overestimation was also evident when participants predicted AI's ability in autonomous driving. We also found that participants expected different information (e.g., warnings and AI explanations) for safely taking over the control of autonomous driving.**

I. Introduction

To achieve autonomous driving in complex and dynamic driving environments, a collection of artificial intelligence (AI) systems have been designed and developed to handle the core functions such as perception, localization, prediction, and planning. For example, the perception module usually takes camera and sensor data (e.g., LiDAR and radar) as inputs, perceives the surrounding environments, and extracts the information for driving (e.g., road obstacles). However, those AI systems are known to be vulnerable to adversarial attacks [9], [20], [25], making autonomous driving susceptible to safety- and security-critical errors that can cause road hazards and even fatal consequences.

Human drivers are still required to take over control from autonomous driving when the system fails or is about to fail (e.g., with SAE Level 3 automation [15]). Thus, it is essential to increase their awareness of those AI vulnerabilities in autonomous driving and design a safe and secure system for them to use. While previous studies have focused on the technical aspects of those AI systems (e.g., [3], [20], [25]), recent work has started to understand human drivers' detection of physical-world attacks (e.g., classification of malicious stop-sign images [10]). However, little research has investigated whether human drivers are able to identify physical-world attacks as the sources of errors for autonomous driving. Moreover, it is unclear what information human drivers expect to receive such that they can increase their awareness of those attacks and take over the control if needed.

We conducted an online survey ($N = 100$) on Prolific to understand human drivers' situation awareness of autonomous driving under two physical-world attacks. Participants answered questions about object classification and autonomous driving projection using images of the two attacks. We also presented videos of the simulated driving for both attacks and assessed participants' satisfaction, take-over intent, and comprehension of the situation in the videos. In addition, we asked open-ended questions to elicit the participants' expected information to understand the situation and take over the control if needed. Moreover, we asked questions to assess participants' knowledge, awareness, and trust of AI systems in autonomous driving before and after our study.

We found that participants could differentiate the benign and adversarial objects (i.e., STOP signs and road lanes). They also perceived that AI was less capable than human drivers in object detection. Yet, the participants overestimated the AI's capability of lane detection with the dirty-road patch compared to the adversarial stop-sign classification. Their projection of autonomous driving with and without the physical-world attacks showed similar results as the detection tasks. The participants also overestimated the AI's ability to drive safely on the dirty road. The quantitative and qualitative evaluations of the videos revealed that the participants' unawareness of physical-world attacks (i.e., overestimation of the AI ability) is likely due to their driving experience or mental models of driving situations (e.g., ice/wet patches on the road due to accidents rather than dirty-road attacks). We suggest researchers consider various tasks (e.g., detection and projection) for human drivers when evaluating physical-world attacks in autonomous driving. We also recommend exploring different information (e.g., warnings and AI explanations) human drivers expected to afford safe take-over control of autonomous driving.

II. Related Work

In the following, we discuss related work of physical-world attacks on AI perception, human situation awareness and take-over control, and trust in autonomous driving.

A. Physical-world Attacks on AI Perception

Multiple AI components are required to complete the complex tasks of autonomous driving, including perception,

Symposium on Vehicles Security and Privacy (VehicleSec) 2023
27 February 2023, San Diego, CA, USA
ISBN 1-891562-88-6
https://dx.doi.org/10.14722/vehiclesec.2023.23002
www.ndss-symposium.org

localization, prediction, and planning. Yet, AI components of autonomous driving are known to be vulnerable to adversarial attacks [11]. In the past decade, extensive efforts have been conducted to understand the attack spaces of autonomous driving. The majority of the existing attacks on autonomous driving systems have focused on physical-layer attacks, especially physical-world attacks (see [25] for a review). Physical-world attacks refer to modifying the physical-world driving environment, and consequently tampering with the sensor inputs to AI systems. Among the physical-world attacks, about half of them are specific to object texture for AI perception [25]. For example, previous efforts have leveraged malicious object texture (e.g., robust physical perturbation [8] and representative ShapeShifter [4]) to make a STOP sign undetected by AI systems [9], [27]. Sato et al. deployed a physical-world adversarial attack using dirty-road patch and obtained a very high success rate [24].

While AI systems are vulnerable to perturbed STOP signs, human drivers can identify the STOP signs before and after a malicious manipulation. Recent work has started to investigate whether human drivers can understand that AI systems are vulnerable to the perturbed STOP sign images. For example, Carcia et al. [10] found that participants revealed a higher-than-expected belief that AI systems would be able to identify a perturbed STOP sign in their study, indicating human drivers' unawareness of the AI vulnerability.

B. Human Situation Awareness and Take-over Control

Since the AI perception systems of autonomous driving are known to be generally vulnerable to adversarial attacks, human drivers' unawareness of such attacks can expose them to safety-critical situations. Moreover, in SAE Level 3 automation, drivers have to be available to take over vehicle control detected and announced by the automated system in "situations that exceed the operational limits of the automated driving system" [15]. Thus, it is essential to understand the proper in-car communication to increasing drivers' awareness of the AI vulnerability and facilitating the take-over control.

Taking over vehicle control before the physical-world attacks causing harm can be demanding for drivers. First of all, they need to detect the occurrence of an attack (e.g., a perturbed STOP sign), assess the risk impacts (e.g., AI perception system could fail to detect the STOP sign), and estimate further effects (e.g., a STOP sign violation or causing a traffic accident). Drivers' situational awareness [7] (i.e., *perception*, *comprehension*, and *projection* of the situation), have been extensively studied with respect to take-over control across different automation levels [1]. However, the domain of physical-world attacks on AI system constitutes a special case for situation awareness in autonomous driving, as it involves how AI systems' capabilities and vulnerabilities should be conveyed to human drivers. To provide human drivers with helpful information, we propose to examine what information they expect for safely taking over the control in those situations.

C. Human Trust in Automation

Humans usually assume that machines (e.g., an AI system) function flawlessly, and consequently encounter them with a trust advance [18], [19]. However, over-trusting the AI system could result in AI's vulnerabilities being neglected and potential risks going through unmitigated. The human-AI trust is expected to be calibrated to an appropriate level through the communication and collaboration between humans and the AI systems [14], [18]. Our study also examines human drivers' knowledge, awareness, and trust of AI systems in autonomous driving as a function of physical-world attacks.

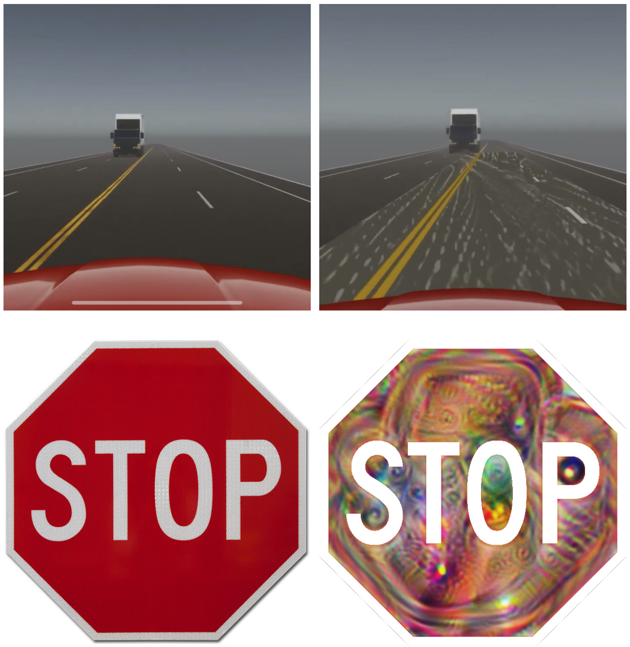

Fig. 1: Images for object detection and autonomous driving projection. Top row shows images for road-surface scenario and bottom row shows images for stop-sign scenario. Left column shows the benign images and right column shows the adversarial images.

III. METHODOLOGY

We examined whether human drivers can *detect* the objects and how they *project* autonomous driving using two scenarios (i.e., dirty-road and stop-sign attacks). We designed our study around SAE Level 3 automation [15], in which participants were required to take over control from the autonomous-driving system when prompted by a take-over request.

A. Materials

To maintain adequate ecological validity of the experiment, we first asked participants to evaluate *images* of both scenarios. We also presented two *videos* of simulated driving showing physical-world attacks (i.e., physically perturbed objects in the environment of autonomous driving).

1) Images: We generated the images by varying two independent factors within subjects: *scenario* (STOP sign vs. road surface) and *image type* (benign vs. adversarial). The image type reflected whether the STOP sign or road surface had been tampered with or not. The benign image was either a standard STOP sign or a local road with a clean surface (see Figure 1). The adversarial image was the same stop-sign image but perturbed by ShapeShifter (SS) [4] or the same local road with a dirty patch [24]. Both attacks have been demonstrated feasible in prior work [24], [25].

2) Videos: After the image evaluation, we asked participants to view a randomly selected video using a between-subject design. The dirty-road attack video was created by Sato et al. [24] and the stop-sign attack was produced by Shen et al. [25]. Each video was about 11 s, presenting a simulated driving of the attack.

Dirty-road attack. In this simulated scenario (i.e., local-road scenario [24]), the autonomous vehicle (AV) first drives toward the lane center along a local road. After 5 s, a dirty-road patch down the road starts to take effect from the driver's perspective. The AV fails to identify road lanes with such dirty-road patch, and causes the vehicle to deviate to the left significantly and hit the truck from the opposite direction. Thus, the simulation scenario shows that the safety impacts of the attack can be severe.

Stop-sign attack. In this simulated scenario, the AV first drives toward an intersection with a STOP sign. The AV fails to recognize the perturbed STOP sign and overshot the STOP sign near the intersection. We chose this scenario because traffic-sign attacks have been presented and investigated recently [9], [25]. Also, the simulated scenario does not result in an accident, showing less-critical safety impacts than the dirty-road attack.

B. Procedure

The online survey was designed using Qualtrics. The survey consisted of four parts. After participants indicated their informed consent, the survey started. *Part 1* was designed to obtain a baseline of human drivers' awareness and trust of AI systems in autonomous driving. We asked participants' agreement on six statements about the AI system for autonomous driving. Question 1 asked participants whether AI is used for extracting important driving information from the environment, which was meant to gauge their knowledge of AI function. Questions 2-6 asked participants to judge their own awareness and trust in AI. All six quantitative questions asked for a response on a 7-point Likert scale, with "1" indicating "completely disagree" and "7" indicating "completely agree." See Appendix A for details of those and the following questions.

In *Part 2*, the participants were exposed to images of the two physical-world attacks (see Figure 1). For each attack, we had one benign setting, in which a clean STOP sign or a clean road was presented. There was also an adversarial setting for each scenario, in which we presented images of the perturbed STOP sign [25] or a dirty-road patch [24].

The participants reported their perception of how an agent would *classify* objects in each image. The agent was either human drivers (e.g., the participants themselves) or an AI system. For the human drivers, the statement asked participants' agreement level with, "I think this image shows a STOP sign/I think this image shows lane lines of the road clearly." on the 7-point Likert scale. For the AI system, the statement asked about their perception of the current AI technology's ability to classify the image, "I think the current AI system in AVs will classify this as an image of a STOP sign/I think the current AI system in AVs will detect the lane lines of the road in the image." using the same 7-point scale.

Moreover, we asked the participants to *project* how the agent would drive using the same design. For example, the statement asked about their agreement level with AI system, "I think the AI system in AVs will navigate the above road condition safely." Participants were not told that the perturbed images were under physical-world attacks. We counterbalanced the order of the two scenarios, as well as the benign and adversarial settings in each scenario between subjects.

In *Part 3*, we first asked participants to imagine that they were driving in an AV with activated AI systems to contextualize the scenario in the video. We also made it clear that they need to supervise the AVs' behavior. We then randomly presented one of the videos with the physical-world attacks. After the video presentation, participants were prompted to evaluate the scenario with two open-ended questions regarding their 1) understanding of the presented scenario and 2) expected information about the situation such that human drivers can safely take over the control. Participants also rated five statements about autonomous driving regarding their 1) satisfaction, 2) perceived safety, 3) take-over intention, 4) perceived cause due to accident, and 5) perceived cause due to intentional attack. The same 7-point Likert scale as the image evaluation was used. To ensure participants' viewing of the video's content, we asked an attention-check question. Following the video presentation, the participants were asked to select the traffic sign in the stop-sign video or the color of the truck from the opposite direction in the dirty-road video.

Lastly, participants filled in their demographic information and answered questions about their driving experience in *Part 4*. We also asked the six questions of Part 1 to examine the impacts of the physical-world attacks on participants' knowledge, awareness, and trust of AI systems in autonomous driving.

C. Pilot Study

Prior to the online survey's deployment, we conducted one pilot study ($N = 20$) on Prolific [23] to evaluate the study's procedure, determine the average duration, and gather feedback on the survey questions for clarity and comprehensibility. Participants took about 10.5 min to complete the survey. We compensated participants with $2 (based on recommended payment rate on Prolific $12/h). Nine participants commented on the study, but none of them experienced any issues.

D. Recruitment and Participants

We recruited another 100 participants on the Prolific. To participate in our study, Prolific workers needed to own a car and be located in the US. We selected car ownership as a criterion to ensure that participants have experience with regular cars. Additionally, we required a 95% human intelligence task (HIT) Approval Rate for all Prolific workers and that they have more than 100 approved HITs.[1] All participants were compensated $2 for completing the study ($T_{Mean} = 11.5$ min, $T_{Median} = 9.6$ min). This experiment was approved by the Institutional Review Board (IRB) at the authors' institution (IRB #: STUDY00021615). Informed consent was obtained from each participant.

A total of 44 males and 56 females took part in our study. Eighty participants do not have a degree or job in computer

[1] The same criteria were also implemented in the pilot study.

science or a related field. About half of the participants have used connectivity functions or driving assistance functions. Only 8 participants reported having previous experience in autonomous driving. Participants' demographics are shown in Appendix Table I.

IV. Results

There was an approximately equal number of participants in each video condition, STOP sign (57) and dirty road (43). Only six participants failed the attention check in the dirty-road condition. We checked responses to the open-ended questions from the six participants. Their responses were relevant to the questions and all looked reasonable. We included their results in the data analysis.

A. Analysis Plan

1) Statistical Analysis: Our statistical analysis focused on quantitative measures in each part. We examined participants' *detection* of adversarial objects and *projection* of autonomous driving by manipulating three within-subject factors: *scenario* (STOP sign vs. road surface), *agent type* (human drivers vs. AI) and *image type* (benign vs. adversarial) in Part 2. To quantify the effect, we conducted repeated-measure analysis of variances (ANOVAs). We also conducted a two-sample *t*-test to compare participants' evaluation of two physical-attack videos. At Parts 1 and 4, we measured participants' knowledge, awareness, and trust of AI system in autonomous driving before and after the study, respectively. A 2 (*study*: before vs. after) × (*video*: STOP sign vs. dirty road) mixed ANOVA was conducted to determine how those two factors and their interaction affected participants' ratings.

We conducted null hypothesis testing (α = 0.05) for those measures. The null hypothesis was rejected when the obtained results among the conditions were significantly different from each other. We implemented parametric tests such as ANOVA because they are robust to yield the right answer even when distributional assumptions are violated [22].

2) Thematic Analysis: We use open coding [26] to evaluate the responses to the two open-ended questions. Two researchers independently coded answers to open-ended questions for each scenario in two iterations. Initially, one of them coded the first half of the dataset, the other one coded the other half, and constructed an initial version of the codebook. After this first iteration, the two researchers discussed their codes and adapted the codebook accordingly. During the second iteration, the two researchers swap the data.

B. Detection and Projection of Physical-World Attacks

1) Detection.: As shown in Figure 2 top panel, the participants can differentiate the benign images from the adversarial images ($F_{(1,99)} = 336.34$, $p < .001$). They gave lower ratings for the adversarial images than for the benign images. The participants also gave lower ratings for AI agent than for human drivers ($F_{(1,99)}$ = 36.18, $p < .001$), indicating that they were less certain about AI's capability of classifying the images in general. Moreover, the 2-way interactions of *image type* × *scenario* ($F_{(1,99)} = 26.33$, $p < .001$) and *agent type* × *scenario* ($F_{(1,99)}$ = 7.29, p = .008), as well as the 3-way interaction of *image type* × *agent type* × *scenario* ($F_{(1,99)}$ = 15.22, $p < .001$) were all significant. Specifically, the participants showed uncertainty in AI's object detection in both scenarios with benign images. Yet, such uncertainty was only evident for the adversarial stop-sign images.

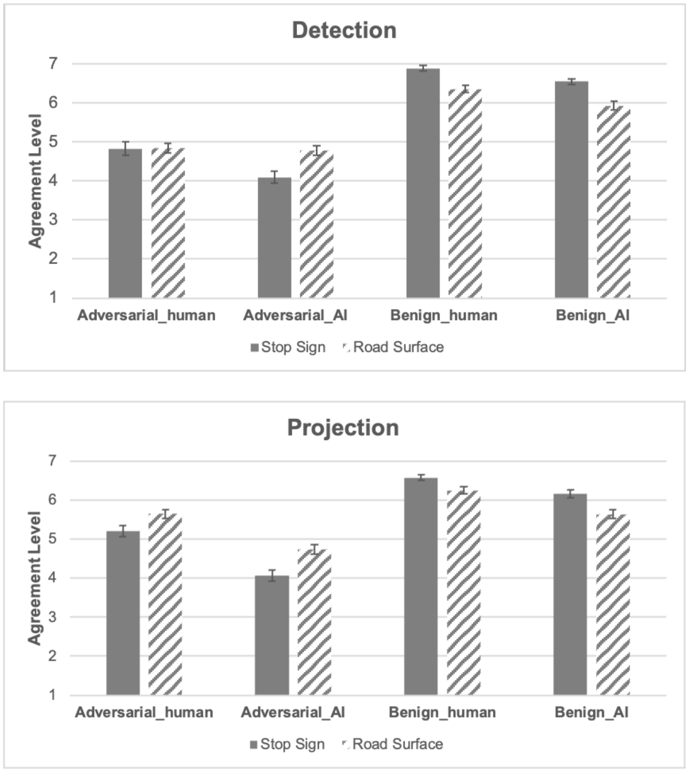

Fig. 2: Results of object detection and autonomous driving projection of images in different conditions at Part 2. The error bars represent ± one standard error.

2) Projection.: The participants were less certain about AI's capability of safely driving in the scenario ($F_{(1,99)} = 110.92, p < .001$). They rated both AI and themselves (human drivers) to be less capable of driving in the adversarial settings than for the benign settings ($F_{(1,99)} = 275.1, p < .001$), and the reduction was greater for AI agent than for human drivers ($F_{(1,99)} = 31.55, p < .001$). Moreover, the 2-way interaction of *scenario* × *image type* ($F_{(1,99)}$ = 51.99, $p < .001$) and the 3-way interaction of *scenario* × *image type* × *agent type* ($F_{(1,99)}$ = 8.95, p = .004) were significant. Specifically, participants' uncertainty of AI' capability was similar between the two scenarios in the benign settings. However, they were more uncertain about AI's capability of driving safety in the stop-sign scenario than the dirty-road scenario in the adversarial settings (see Figure 2 bottom panel).

Overall, the results were similar for the detection and projection measures. Also, participants' gave an average rating above 4 for adversarial images evaluated by AI agent (see Figure 2), indicating that they did not seem to be aware of the physical-world attacks, especially the dirty-road attack.

C. Quantitative Measures of Videos

Figure 3 shows the results of quantitative measures of the two videos. The participants were not satisfied with the behavior in the dirty-road video, which was significantly lower

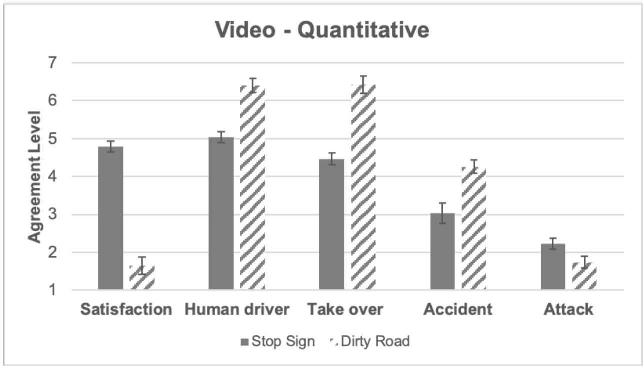

Fig. 3: Results of quantitative measures for videos. The error bars represent ± one standard error.

than that in the stop-sign video ($t_{(1,98)} = -10.63$, $p < .001$). They believed that human drivers would handle the situation better than AI in each video, and the rating was higher after viewing the dirty-road video than the stop-sign video ($t_{(1,98)} = 5.79$, $p < .001$). The participants would like to take over the control in both scenarios. The rating was also higher for the dirty-road video than for the stop-sign video ($t_{(1,98)} = 6.66$, $p < .001$). Participants believed that the dirty-road scenario was more likely due to accidents than the stop-sign scenario ($t_{(1,98)} = 3.89$, $p < .001$). Those results are possibly due to the collision shown in the dirty-road video.

Although the participants gave higher ratings for the stop-sign video than that for the dirty-road video ($t_{(1,98)} = -2.16$, $p = .033$), they did not believe that either scenario was due to attacks. Such results are in agreement with the results of detection and projection tasks.

D. Human Drivers' Knowledge, Awareness, and Trust of AI Systems in Autonomous Driving

Figure 4 shows the mean values of the ratings across conditions. The participants agreed that AI is essential for perceiving the driving environment and extracting relevant information in autonomous driving in general. Such agreement was not dependent on the video scenarios ($F < 1.0$) or before/after the current study ($F_{(1,98)} = 2.62$, $p = .108$). The participants became more aware of using AI for perceiving the driving environment and extracting relevant information after our study ($F_{(1,98)} = 20.73$, $p < .001$). Their basic understanding of the concepts and technology that allow AI to work showed a trend to increase after viewing the stop-sign video but not after viewing the dirty-road video ($F_{(1,98)} = 5.76$, $p = .018$).

The participants who viewed the stop-sign video showed higher trust than those who viewed the dirty-road video ($F_{(1,98)} = 6.14$, $p = .015$). The main effect of *video* was qualified by *study* ($F_{(1,98)} = 20.72, p < .001$), indicating that such difference became more evident after the current study. In agreement with the trust measure, participants became more worried about the use of AI in autonomous driving after the current study ($F_{(1,98)} = 6.73$, $p < .011$), mainly for those who viewed the dirty-road video but not for those who viewed the stop-sign video ($F_{(1,98)} = 12.36$, $p < .001$). Even if the participants could know more about how AI perceives the environment and extracts information in autonomous driving, they reduced their trust in AI for autonomous driving after the current study ($F_{(1,98)} = 4.56$, $p = .035$).

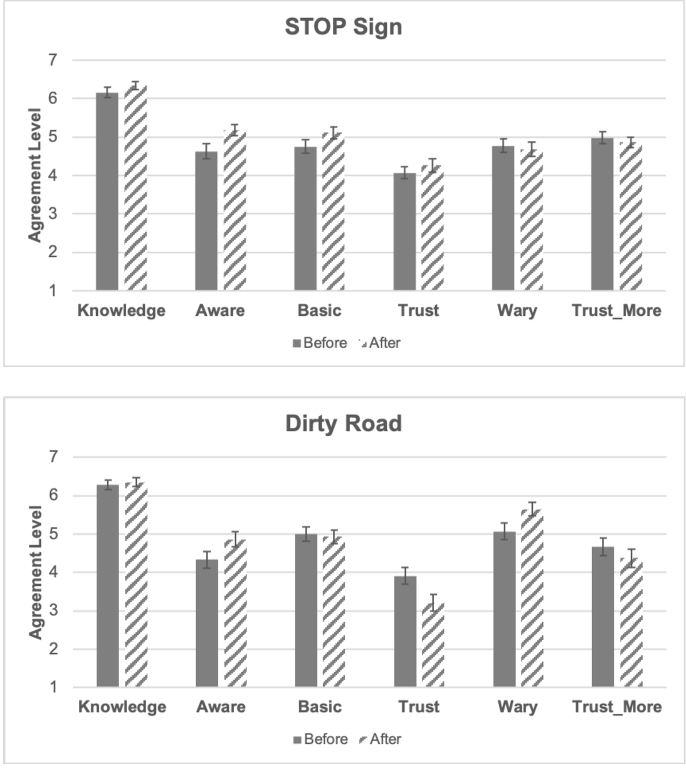

Fig. 4: Average ratings of the six statements of human drivers' knowledge and trust of AI systems in autonomous driving at Parts 1 and 4. The error bars represent ± one standard error.

E. Thematic Analysis

The inter-rater reliability of Cohen's Kappa is 0.87 [17], indicating a satisfactory agreement. The two coders resolved the discrepancies, discussed the results, and finalized the thematic analysis together. Note that a single response may have multiple themes. We describe a few major themes and the numerical difference across scenarios.

1) Comprehension of the Video Contents: We report the results of the dirty-road video first and then the STOP-sign video. For participants ($n = 43$) who viewed the dirty-road video, the most common themes are:

- **AI malfunction, error, or confusion:** 28 participants (65.1%) mentioned the scenario was caused by some problem in the AI, like **P100**: *"The AI confused the blurriness on the ground and did not stay in its lane."*
- **Incorrect AI Lane Detection:** 19 participants (44.2%) specifically mentioned that the AI did not correctly read the lane lines on the road (e.g., **P19** said that *"The AI thought the marks on the road were the lanes ..."*, and **P43** felt that *"... the AI could not detect the street lines appropriately and swerved out of bounds as a result."*)
- **Road Condition:** 24 participants (55.8%) mentioned something about the condition of the road, like the dirty markings. Interestingly, several participants interpreted the dirty patch as ice, like **P40**, who said that *"... an icy/wet patch was hit ..."*, and **P76** responded, *"... [the] mark on the ground could have been a patch of black ice and it made the car swerve into the oncoming lane and hit the truck."*

- **Collision:** 22 participants (51.2%) explicitly mentioned the car's collision with the truck. For example, **P3** responded: *"The AI was not able to detect the line on the ground ... so it crashed into the oncoming truck."*

For participants ($n = 57$) who viewed the STOP sign video, most ($54, 94.8\%$) answered that the car or AI stopped at the sign. Only 13 (22.8%) participants believed that AI did not stop or human drivers had intervened. Of those responses, **P2** described that "*... the AI ignored the stop sign and the human had to stop the car themselves.*" In contrast to the dirty-road condition, only three participants explicitly mentioned that the STOP sign was malignant.

2) Expected Information for Safe Take-over Control: We first report the results of participants who viewed the dirty-road video. They expected the following information.

- **An Alert:** 34 participants (79.1%) wished for some kind of alert (e.g., audio or a combination of audio and visual). An example is **P76**, who stated that the vehicle *"... should make a chime ... stating that it cannot detect part of the road ahead ..."*
- **An explanation of the AI's decision-making:** 20 participants (46.5%) wanted to see the AI's reasons for deciding how the vehicle will move next. For example, **P56** wanted the vehicle to explain that the AI *"... had detected something impairing its ability to make judgment on the road condition,"* and **P47** wanted the AI to indicate that it *"... cannot detect road lines so the driver can take over."*
- **An explanation of AI errors:** Eighteen participants (41.9%) wanted to see some explanation of why a driving AI made a faulty decision. Five other participants specifically wanted to see a description of lane violation. For example, **P100** wanted the AI to *"... notify the [driver] that it cannot identify what it sees"*, and **P66** stated that *"It should provide that it is about to swerve ..."*
- **Request the driver to take over:** Eleven participants (25.6%) wanted the vehicle to indicate that the human drivers should take control (e.g., **P34** wanted the vehicle to indicate *"That the AI is unable to safely navigate and the human will need to interact immediately."*)

Participants who viewed the stop-sign video expected the following:

- **Explicit mention of the stop sign:** Twenty participants (35.1%) wanted the vehicle to explicitly point out the stop sign to the driver. **P81** responded that *"The AI system should acknowledge there's a stop sign approaching ..."*, and **P32** wanted a *"... visible indicator of stop sign."*
- **An alert:** Nineteen participants (33.3%) expected some kind of warning alert. Like the dirty-road condition, the warnings could be audio, or both visual and audio. For example, **P95** responded that *"... the AI system could have some sort of notification pop up on the dashboard that says 'Upcoming: Stop Sign'..."*
- **An explanation of the AI's decision-making:** Eleven participants (19.3%) would like to see an explanation of the AI's decisions. For example, **P8** wanted *"... some sort of screen that shows the human what the AI was detecting - like if there was a little message on the dash board that said 'stop sign approaching' ... so that humans know that the AI knows there is a stop sign."*
- **An explanation of AI errors:** Eight participants (14%) expected descriptions of AI error, confusion, or uncertainty (e.g., **P80** stated, *"[The AI] should tell the human that it is having trouble telling if there is a stop sign."*

Different from the dirty-road condition, 25 (43.9%) participants who viewed the stop-sign video indicated that they might not have understood why human drivers need to take over the control, indicating their unawareness of the malicious STOP sign and stop-sign violations in the video.

V. Discussion

The present study investigated human drivers' detection of physical-world attacks and projection of autonomous driving in two scenarios (i.e., dirty-road and stop-sign attacks). We found that participants were able to differentiate the benign and adversarial images for both attacks. They also rated AI agent to be less capable of such detection and projection than human drivers. While such uncertainty about AI's capability was evident for the benign images of both scenarios, it was only evident for the adversarial stop-sign images. After viewing the videos, participants did not believe either scenario was due to an attack. As revealed in their responses to the open-ended questions, participants' unawareness of the physical-world attacks could be due to them having experienced a similar situation without an attack (e.g., ice/wet patches on the road due to bad weather/an accident).

The participants rated AI agent and human drivers to be less capable of detection and projection tasks in the adversarial settings than in the benign settings, but the reduction for the AI was greater than that for human drivers. Such results are in agreement with prior work showing that humans have higher self confidence than confidence in automation [14]. Compared to the general knowledge question in Section IV-D, the detection and projection tasks are more specific. Thus, while human drivers are aware of the use of AI in autonomous driving, they lack knowledge about *how* AI system functions in specific autonomous driving tasks.

Participants in our study believed that the lane-detection and lane-keeping tasks were more difficult than classifying STOP signs and driving in the stop-sign scenarios using benign images. Such results make perfect sense since lane-detection and lane-keeping tasks involve lateral control, which is more attention-demanding or effortful than speed control (e.g., STOP sign compliance) [13]. Interestingly, we also found that participants did not believe that the dirty-road patch could cause problems for either human drivers or AI agent. Responses to the open-end questions revealed that participants' prior driving experience may have made them more confident in driving on the dirty road (e.g., ice/wet road). In the case of the stop-sign video, the participants might not be able to identify the violation since overshooting a STOP sign is a very common offense for a lot of human drivers.

Mental models are human's internal representations of the external world, which is dynamic and based on individual experience [16]. Our findings indicate that human drivers' mental models of driving situations are based on their experience (e.g., dirty-road attack → ice/wet road), which could impact their perception and projection of autonomous driving and result in misconceptions. Previous studies showed that users' mental models could be improved with increased transparency of intelligent systems [6]. Explainable AI (XAI) is a research field that aims to make AI systems' decisions more transparent and understandable to humans [21]. Previous work has started to explain autonomous driving behaviors [2], [12]. When asked about their expected information for safely taking over the control, a lot of participants expected explanations of the AI decision, as well as AI errors, confusion, or uncertainties. In terms of trust in AI, participants gave a lower rating for the dirty-road scenario than for the stop-sign scenario. A possible reason is that participants viewed the collision in the dirty-road video, which is safety-critical. The results are similar to prior work that shows participants' trust was corrected downwards with every mistake the machine made [5]. Besides explaining capabilities, it is essential to communicate vulnerability of AI systems, which can enhance mental models of driving situations, calibrate trust in AI, and enable human-AI collaboration.

After viewing the videos, few participants mentioned the perturbed STOP sign, and more than half of them even responded with confusion when asking for their expected information for taking over the control. Thus, although human drivers are capable of detecting the attacks, they probably will miss the attacks in actual driving or think the malignant marks do not matter. Such results underline the importance of constructing ecologically valid experiments to evaluate human drivers' susceptibility to physical-world attacks.

A. Limitations

There are a few limitations in the current study. *First*, we only evaluated two cases of physical-world attacks, which limited us to discussing further implications of other physical-world attacks from human drivers' perspective. Future work can systematically assess the capability of human drivers to identify physical-world attack vectors (e.g., object texture, object shape, and object position [25]) in autonomous driving. *Second*, we recruited participants on Prolific, who are usually between 18 and 44 years old and have at least some level of college education. Thus, the obtained results might not be representative of the general public. *Third*, participants made responses with hypothetical scenarios, images, and videos from simulated drivings. Thus, the experience could be different from actual driving. *Lastly*, we noticed extra differences between the benign and adversarial images in the dirty-road attack, such as the trunk of the ego vehicle and the truck size (see Figure 1). However, we expect minimal impacts of those factors on the detection and projection results since participants were instructed to pay attention to the road lanes. However, future work can consider better controlling those factors and further evaluating the scenarios. Moreover, we note that the size, viewing angle, and viewing distance of the STOP sign in the video might have resulted in participants' unawareness of the perturbed sign. Future work could consider conducting a study using a driving simulator, which could offer more real driving experience and driving-related tasks.

VI. Conclusion

A clear understanding of human drivers' detection of physical-world attacks and projection of consequent driving allows us to better comprehend their situation awareness of autonomous driving under such attacks. Our study shows that human drivers' awareness of those attacks (e.g., the dirty-road attack) is limited, which is likely due to their prior driving experience. Moreover, our work demonstrates the importance of constructing different tasks for human drivers to evaluate various physical-world attacks in autonomous driving. Our study also provides useful information to in-car risk communication, which could afford safe take-over control of autonomous driving and appropriate calibration of trust in AI.

References

[1] L. Avetisyan, J. Ayoub, and F. Zhou, "Investigating explanations in conditional and highly automated driving: the effects of situation awareness and modality," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 89, pp. 456–466, 2022.

[2] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[3] Y. Cao, N. Wang, C. Xiao, D. Yang, J. Fang, R. Yang, Q. A. Chen, M. Liu, and B. Li, "Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks," in *2021 IEEE Symposium on Security and Privacy*. IEEE, 2021, pp. 176–194.

[4] S.-T. Chen, C. Cornelius, J. Martin, and D. H. P. Chau, "Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2019, pp. 52–68.

[5] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *International Journal of Human-computer Studies*, vol. 58, no. 6, pp. 697–718, 2003.

[6] M. Eiband, H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, and H. Hussmann, "Bringing transparency design into practice," in *23rd International Conference on Intelligent User Interfaces*, 2018, pp. 211–223.

[7] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," in *Situational awareness*. Routledge, 2017, pp. 9–42.

[8] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramer, A. Prakash, T. Kohno, and D. Song, "Physical adversarial examples for object detectors," in *12th USENIX Workshop on Offensive Technologies*, 2018.

[9] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[10] K. R. Garcia, S. Mishler, Y. Xiao, C. Wang, B. Hu, J. D. Still, and J. Chen, "Drivers' understanding of artificial intelligence in automated driving systems: A study of a malicious stop sign," *Journal of Cognitive Engineering and Decision Making*, vol. 16, no. 4, pp. 237–251, 2022.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[12] J. Haspiel, N. Du, J. Meyerson, L. P. Robert Jr, D. Tilbury, X. J. Yang, and A. K. Pradhan, "Explanations and expectations: Trust building in automated vehicles," in *Companion of the 2018 ACM/IEEE International Conference on Human-robot Interaction*, 2018, pp. 119–120.

[13] J. He, J. S. McCarley, and A. F. Kramer, "Lane keeping under cognitive load: Performance changes and mechanisms," *Human Factors*, vol. 56, no. 2, pp. 414–426, 2014.

[14] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors*, vol. 57, no. 3, pp. 407–434, 2015.

[15] S. International, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles,," 2021, available at https://www.sae.org/standards/content/j3016_202104/.

[16] P. N. Johnson-Laird, *Mental models: Towards a cognitive science of language, inference, and consciousness.* Harvard University Press, 1983, no. 6.

[17] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, pp. 159–174, 1977.

[18] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[19] P. Madhavan and D. A. Wiegmann, "Similarities and differences between human–human and human–automation trust: an integrative review," *Theoretical Issues in Ergonomics Science*, vol. 8, no. 4, pp. 277–301, 2007.

[20] Y. Man, M. Li, and R. Gerdes, "{GhostImage}: Remote perception attacks against camera-based image classification systems," in *23rd International Symposium on Research in Attacks, Intrusions and Defenses*, 2020, pp. 317–332.

[21] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein, "Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai," *arXiv preprint arXiv:1902.01876*, 2019.

[22] G. Norman, "Likert scales, levels of measurement and the "laws" of statistics," *Advances in Health Sciences Education*, vol. 15, no. 5, pp. 625–632, 2010.

[23] S. Palan and C. Schitter, "Prolific. ac—a subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, 2018.

[24] T. Sato, J. Shen, N. Wang, Y. Jia, X. Lin, and Q. A. Chen, "Dirty road can attack: Security of deep learning based automated lane centering under {Physical-World} attack," in *30th USENIX Security Symposium*, 2021, pp. 3309–3326.

[25] J. Shen, N. Wang, Z. Wan, Y. Luo, T. Sato, Z. Hu, X. Zhang, S. Guo, Z. Zhong, K. Li *et al.*, "Sok: On the semantic ai security in autonomous driving," *arXiv preprint arXiv:2203.05314*, 2022.

[26] A. Strauss and J. M. Corbin, *Grounded theory in practice*. Sage, 1997.

[27] Y. Zhao, H. Zhu, R. Liang, Q. Shen, S. Zhang, and K. Chen, "Seeing isn't believing: Towards more robust adversarial attack against real world object detectors," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 1989–2004.

APPENDIX

Survey questions about physical-world attacks. [Part 1&4: Knowledge and Trust in AI] *Participants were asked to rate their agreement level with the following statements using a 7-point Likert scale (Completely disagree (1), Strongly Disagree (2), Disagree (3), Neither Agree Nor Disagree (4), Agree (5), Strongly Agree (6), Completely agree (7)).*

Q1. Autonomous driving requires the use of AI to perceive the environment and extract information for driving, such as detecting pedestrians or other vehicles. **Q2.** I am aware of how autonomous driving systems use AI to perceive the environment and extract information for driving. **Q3.** I have a basic understanding of the concepts and technology that allow AI to work. **Q4.** I can trust AI in autonomous driving. **Q5.** I am wary of the use of AI in autonomous driving. **Q6.** If I know more about how AI perceives the environment and extracts information, I would trust it more in autonomous driving.

[Part 2: Images] *The participant was presented with both the benign and adversarial images for each scenario in a random order. The questions for the benign and adversarial images in each scenario were the same.*

Stop Sign Questions: **Q1.** I think this image shows a STOP sign. **Q2.** I think the current AI system in AVs will classify this as an image of a STOP sign. **Q3.** I think a human driver will navigate a driving situation with the above STOP sign on the road safely. **Q4.** I think the current AI system in AVs will navigate a driving situation with the above STOP sign on the road safely. *Road Surface Questions:* **Q1.** I think this image shows lane lines of the road clearly. **Q2.** I think the current AI system in AVs will detect the lane lines of the road in the image. **Q3.** I think a human driver will navigate the above road condition safely. **Q4.** I think the current AI system in AVs will navigate the above road condition safely.

[Part 3: Video] *The participant was randomly shown one of the two videos. They were then asked a multiple choice question to gauge their attention. Afterwards, they were presented with two open-response questions.*

Stop Sign Video: **Q1.** What traffic sign was featured in the video? (*Options:* Stop sign, Traffic Light, Yield Sign, Prefer not to answer) **Q2.** What do you think happened in the video? **Q3.** What information should the AI system provide about the situation so that human drivers can *avoid* the STOP sign violation? *Road Surface Video:* **Q1.** What colors were the truck? (*Options:* Blue and red, Green and yellow, Black and white, Prefer not to answer) **Q2.** What do you think happened in the video? **Q3.** What information should the AI system provide about the situation so that human drivers can *safely* take over the control?

Participants were then asked to rate their agreement with the following statements using the 7-point Likert scale as Part 1. These questions were the same regardless of what video the participant saw.

Q4. I am satisfied with the AV's behavior in the situation. **Q5.** I would drive more safely than AI in this situation. **Q6.** I would take over the AI's driving in this situation. **Q7.** I believe this situation was caused by accident. **Q8.** I believe this situation was caused by intentional attack.

TABLE I: Demographic Information of Participants

Item	Options	Percentage
Gender	Male	44%
	Female	56%
Age	18-24	12%
	25-34	23%
	35-44	24%
	45-54	24%
	55 or older	17%
Ethnicity	American Indian/Alaskan Native	1%
	African/African American	8%
	Hispanic/Latino	4%
	Caucasian	73%
	Asian	14%
Education	No high school	1%
	High School	19%
	Some College	5%
	Associate's	11%
	Bachelor's	43%
	Masters/PhD	21%
CS Related Experience	No	80%
	Yes	19%
	Prefer not to answer	1%
Driver's License	Yes	100%
Avg. Annual Mileage (mi.)	$<$ 2,000	14%
	2,000–5,000	26%
	5,000–10,000	32%
	10,000–20,000	21%
	$>$ 20,000	4%
	Prefer not to answer	3%
Connectivity Function Experience	Not at all	35%
	Rarely	14%
	Sometimes	30%
	Quite often	21%
Driving Ass. Function Experience	Not at all	37%
	Rarely	14%
	Sometimes	34%
	Quite often	15%
Driven AVs	Not at all	88%
	Rarely	4%
	Sometimes	6%
	Quite often	2%