Human Perception of AI Capabilities at Classifying Perturbed Roadway Signs

Katherine R. Garcia, Jing Chen, *Member, IEEE*, Yanru Xiao, Scott Mishler, Cong Wang, *Member, IEEE*, and Bin Hu, *Member, IEEE*

***Abstract*—Artificial Intelligence (AI) is crucial to numerous functions required for driving automation systems, including the computer vision techniques used to detect the roadway environment and make real-time decisions. However, the images used as inputs to the AI system may be maliciously perturbed, or manipulated, causing the AI system to make an incorrect classification. In this study, we examined humans' perception of the AI's computer vision capability of classifying various road sign images, including the original images, images with two different types of malicious attacks, and images that are scrambled randomly at the pixel level. Our results showed that participants rated the AI agent to be less capable than themselves of classifying the road signs. However, they overestimated the AI's computer vision capability for correctly classifying images with malicious attacks that should cause the AI system to misclassify the image. These findings suggest that people lack an accurate understanding of the vulnerabilities of AI computer vision technologies and tend to overtrust AI in driving automation systems.**

***Index Terms*—Cyber security, human factors, human-automation interaction, machine learning, man–machine systems.**

I. Introduction

AUTOMATED driving systems utilize sensors to transform environmental data, such as roadway hazards and the detection of other vehicles, into information that may be used for real-time perception and decisions by the system. Numerous systems and components must work together for the automated driving system to operate successfully. Artificial Intelligence (AI) oversees these processes and combines the information in higher level automation systems, such as SAE Levels 2–5 of driving automation systems (DAS) [1]. In order to avoid dangerous situations, the AI systems in higher levels of DAS must learn to operate the vehicle and monitor the roadway for any potential hazards.

Currently, at SAE Level 2, human drivers are still required to attend to the vehicle and the driving environment so that they can take over control when the system fails; at Level 3, drivers are not required always to pay attention, but they also need to take over control when prompted [1]. As such, for human drivers to take over in time, it is critical for them to recognize and predict when the DAS will fail before it causes any harm. To predict when the DAS will falter, human drivers must understand its capabilities and limitations. However, there is still a lack of understanding of the users' perception regarding the capabilities of AI technologies in DAS [2]. This study investigated human perception of the AI's computer vision capabilities in correctly classifying various road signs by manipulating the type of potential cyber-attacks on the signs. This article makes the following contributions.

1) Highlights a critical gap between human perceptions and AI capabilities in the DAS cybersecurity context.
2) Implements a large-scale user study to understand how humans perceive AI capabilities and trust it in classifying perturbed road signs.
3) Demonstrates drivers lack an accurate understanding of the limitations of computer vision techniques.
4) Emphasizes the importance of explaining the AI system's capabilities for users to better understand it to aid appropriate trust calibration.

II. Related Works

A. AI in DAS

AI can now be seen everywhere today, from facial recognition on social media platforms to voice-activated home assistants [3]. AI is a broad field that focuses on intelligent machines that mimic human abilities to achieve goals and can range from specific functions to complex processes [4], [5], [6]. With the advancement of technology, AI is expected to lead to more efficiency, safety, and higher quality services than humans are able to provide [7]. However, the more complex the system is, the higher the likelihood that AI will falter and result in an incorrect classification. As AI advances and becomes more prominent, the human that once completed the task transitions into the role of a supervisor rather than an operator [8]. Whenever the AI system

Received 22 January 2024; revised 21 March 2025; accepted 28 April 2025. This work was supported by the National Science Foundation [Awards #2007386 and #2245055]. This article was recommended by Associate Editor M. Dorneich. *(Corresponding author: Jing Chen.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board at Old Dominion Univeristy, and performed in line with the American Psychological Association's ethical guidelines for human research.

Katherine R. Garcia and Jing Chen are with the Department of Psychological Sciences, Rice University, Houston, TX 77005 USA (e-mail: krg5@rice.edu; jingchen@rice.edu).

Yanru Xiao is with the Computer Science Department, Old Dominion University, Norfolk, VA 23529 USA (e-mail: yxiao002@odu.edu).

Scott Mishler is with the Psychology Department, Old Dominion University, Norfolk, VA 23529 USA (e-mail: smish001@odu.edu).

Cong Wang is with the Department of Cyber Security Engineering, George Mason University, Fairfax, VA 22030 USA (e-mail: cwang.odu@gmail.com).

Bin Hu is with the Department of Engineering Technology, University of Houston, Houston, TX 77204 USA (e-mail: bhu11@uh.edu).

Digital Object Identifier 10.1109/THMS.2025.3573173

makes an error, such as misunderstanding a voice command, the human monitor may intervene to correct the error [9].

In the context of DAS, incorrect decisions by the AI system may lead to harm to the driver and those in the driving environment, including both humans and properties [7]. AI is involved in multiple functions within the DAS, such as adaptive cruise control, lane departure sensor, collision avoidance, parking assist, and many others. All of these functions are present in today's DAS, as well as object recognition, which requires AI computer vision, which is a technique to mimic the human visual system [10], [11], [12]. As the level of automation increases, the driver becomes more of a supervisor in comparison to lower levels when the driver engages in more manual driving [1]. It is important for human drivers to understand the capabilities and limitations of AI to best perform their new role and correct these errors when they occur.

AI systems are vulnerable to adversarial inputs, such as maliciously manipulated inputs that cause the AI algorithms to come to an erroneous decision [13]. In machine learning, a type of AI, the system learns to complete a specific task with a dataset that is used for training [5]. If the AI system is given limited training data, it may develop biases in its future decisions. For example, training a facial recognition program with only individuals of a certain skin tone will lead to more errors for skin tones that were not included in the training data [14], [15]. In addition, since future decisions are made based on the given training data, the AI system may produce an error if a given input does not resemble the training data [15]. Although AI can complete complex processes and functions, it is vulnerable to adversarial inputs [16]. It has been shown that purposed perturbations that slightly modify the inputs, such as maliciously added image pixels or minimally perturbed added noise, maximize the error rate of the AI algorithm, causing a misclassification [13], [17]. In comparison, these maliciously manipulated images can typically be easily identified by human eyes.

B. Computer Vision in DAS

In DAS, AI computer vision and the vehicle's sensors work together to collect and process data, relay information about the environment, plan a path for the vehicle, execute the navigation path by manipulating the vehicle, and monitor changing conditions within the vehicle to make sure everything is functioning properly [7], [18], [19]. AI computer vision can classify a large range of objects in the driving environment, such as road signs, road markings, other vehicles, pedestrians, obstacles, and more.

Random noise inevitably occurs from the sensor data collection and must be processed by computer vision, which causes uncertainties in decisions [20]. Noise can be from conditions in the naturalistic environment, such as fog or snow [18], human-made markings and stickers [21], or just natural wear and tear [22]. Malicious attackers may intentionally perturb the road-sign images by taking advantage of the AI system's vulnerabilities, leading to misclassification and preventing the DAS from functioning as intended [23], [24]. However, small perturbations to road-sign images, such as the addition of noise, would not normally affect human vision from recognizing the sign, but it would for the AI system. The AI system works to minimize the error of its output, but perturbed noise causes this system to maximize its error [25], [26]. The human driver would need to take over for the vehicle when the AI system inevitably fails to correctly classify a sign that has been maliciously perturbed. To ensure a rapid takeover, the human would need to be aware of the AI computer vision's vulnerabilities regarding these manipulated inputs to predict when the AI system may fail.

C. Human Perception of AI in DAS

It is important for users to understand the capabilities and limitations associated with automated systems, which are prone to making errors [27], [28]. For many automated systems, the human is still needed in the process to oversee the system, make sure it is functioning properly, and intervene if an error occurs [8]. Since the malicious attacks applied to road signs usually do not affect human vision, it is important for human drivers to be aware of AI's vulnerabilities to appropriately predict when the system may fail. Studying how people perceive AI can help predict how users will use and rely on it.

Accurate calibration of trust by the user is a crucial factor in ensuring a functional human-automation system [29], [30]. Trust calibration refers to the process of individuals assessing the capabilities of automated systems and adjusting their level of trust accordingly [29], [30]. Over-trusting the system may lead to undetected or uncorrected errors moving forward, whereas under-trusting the system may lead to not using the system to its fullest extent, and losing out on its potential efficiency [31]. Similar to trust in human-automation systems, human–AI trust is key to developing successful collaborations between the two parties and taking full advantage of all AI has to offer [32]. Part of this collaboration is understanding how the AI system works and what it is capable of. As AI systems become more advanced, the human's role transitions to one that provides more oversight over the completion of the task than one who works on the task itself [8]. AI systems are prone to making errors, and it is important for users to understand the system's capabilities and limitations to make up for these errors [27], [33].

A recent study by Garcia et al. [2] investigated how human drivers perceive the AI's computer vision capabilities in classifying a stop sign that was compromised by a malicious attack. They found that participants overestimated the AI's ability to correctly classify the maliciously manipulated stop sign. This finding suggests that the general public does not accurately understand the AI system enough, which leads them to overestimate the AI's computer vision capabilities. However, Garcia et al. [2] only used one type of road sign (a stop sign), and only one type of malicious attack (the physical attack). Consequently, it is difficult to generalize the results to other potential roadway signs and how the contents of the signs may affect the results. In addition, the physical attack that was used in the study is discernible to the human eye. This makes it unclear how other perturbations, such as one that is less discernible to the human eye, would affect the driver's perceived AI capabilities.

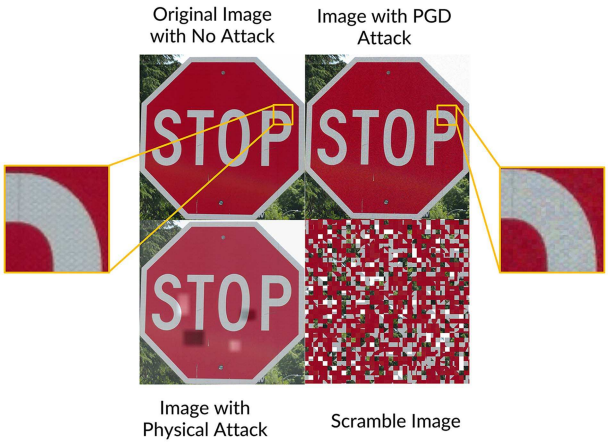

Fig. 1. Illustration of four attack types on a stop sign.

D. Focus of Current Study

Drivers need to monitor DAS to take over in time whenever it fails. Computer vision is used in many AI systems within DAS, and understanding its capabilities and vulnerabilities will aid drivers in anticipating potential errors and cyberattacks. The current study aimed to examine humans' perception of the AI's computer vision capabilities of classifying various road sign images that have been maliciously perturbed through different adversarial attacks. This study used 11 unique road sign types with four different attack types to collect both the human's perception of the attacks and their perception of the AI's computer vision ability to correctly classify road signs under different attacks. Participants were tasked to rate their agreement with statements that reflected either their own or the AI system's ability to correctly classify the road-sign images. By including different road signs, we also tested how the safety criticality of the road signs played a role in people's perception by including it as a covariate in the analyses.

A total of four different attack manipulations were used, including the no-attack, the projected gradient descent (PGD) attack, the physical attack, and the scrambled manipulation (see Fig. 1). The no-attack images were the original, non-manipulated, images of road signs, and served as a baseline condition in which both the human and the AI's computer vision should be able to recognize the images. The PGD attack, physical attack, and scrambled manipulation were each applied to the no-attack image. The PGD attack images were generated by using the algorithms [34] proposed, where the gradient perturbation is embedded in the background of the image. The noise added by the PGD attack was difficult for the human eye to detect [35]; due to the subtleness of the added noise, the no-attack image and PGD-attack image were almost visibly identical to the human eye. The physical attack images were generated using the algorithm [36] proposed, where the current AI technology has a 0% accurate classification rate. The added noise by the physical attack was more visible to the human eye [2], [36]; the no-attack image and physical-attack image were visibly distinct. Both the PGD and physical attack images should be misclassified by the AI's computer vision while the human should correctly classify both [16]. Lastly, the scrambled manipulation images were generated by mixing individual pixels of the original images to make it difficult for both the human and AI to recognize the images, serving as a control condition.

Participants rated how well both they themselves and the AI system could classify each road-sign image. We hypothesized that participants would overestimate the AI's computer vision capabilities for the maliciously attacked images, based on the finding in [2]. It was also expected that this overestimation would be greater for the PGD attack that is less visible to the human eye than for the physical attack. The AI system would have a difficult time classifying both the images with the PGD attack and those with the physical attack because the noise added, regardless of visibility to the human eye, would disrupt the computer vision techniques [20]. However, average users may be unaware of this knowledge and expect the AI system to be able to correctly classify the image with the PGD attack, as they would be able to, more than the images with the physical attack. We also expected that people would perceive both the human and AI agent to be able to correctly classify the images with no-attack, and neither would be able to correctly classify the images with the scrambled manipulation.

This study contributes to the understanding of how humans perceive and trust the AI system in classifying maliciously manipulated images. It also sheds light on how humans connect the perceived safety criticality of various road signs to the capabilities of AI computer vision techniques. This study emphasizes the importance of explaining the AI's computer vision capabilities to its users so they may have an accurate understanding of the system [37].

The the rest of this article is structure as follows: Section III explains the methods employed for this study, including the participants, materials, experimental design, and procedures. Section IV first describes the analyses of the participant's ratings of trust in the AI's computer vision system, followed by the results of the participants' perception of the human's and AI's ability to classify the different road-sign images under the four attack types. Section V discusses the findings of the study, along with their implications and limitations, as well as future research ventures. Section VI concludes the article.

III. Method

A. Participants

A total of 230 participants were recruited through Amazon Mechanical Turk (MTurk), an online crowdsourcing site. The mean reported age was 36.53 years ($SD = 11.06$). Participants reported their gender as either female ($n = 66$) or male ($n = 164$). Participants reported an average of 15.23 years of driving experience ($SD = 12.24, n = 215$), or responded with "N/A" ($n = 15$). Participants reported their race as White ($n = 192$), Black or African American ($n = 18$), American Indian or Alaska Native ($n = 3$), Asian ($n = 11$), or other/mixed ($n = 6$). Participants were required to be between the ages of 18 and 89, have a driver's license, be located in the United States, and have a HIT approval rating of at least 95% (i.e., they have successfully completed and got approved for at least 95% of the tasks they performed on MTurk in the past). The experiment took an average of 55 min,

and each participant was compensated $6 for their participation. This study was approved by a southeast university's Institutional Review Board and complied with the APA ethical guidelines.

B. Materials

The study was presented using Qualtrics (Qualtrics.com) on participants' own laptops or desktop computers. This study contained a pretask questionnaire, an experimental task with two blocks, and a post-task questionnaire. The purpose of the pretask questionnaire was to measure participants' trust in AI-in-general, in AI in autonomous driving, and their self-confidence in their own driving abilities, and thus it was implemented before the experimental task. The pretask questionnaire contained a total of 17 statements participants were required to rate their agreement using a 7-point Likert scale, where 1 represents *strongly disagree* and 7 represents *strongly agree*. Seven of the 17 statements measured the participants' trust in AI-in-general, and seven measured their trust in AI in autonomous driving (AI-in-AV). These 14 statements were adapted from [38]. Then three statements measured the participants' self-confidence in their own driving ability.

The experimental task included two blocks, one with questions on the AI agent, and the other with questions on the human agent. For example, a human agent statement would state, "I think this image shows a 'stop' sign," while an AI agent statement would state, "I think the current AI technology will classify this as an image of a 'stop' sign." We did not elaborate on what the "current AI technology" referred to avoid biasing participants. The order of the two blocks was counterbalanced between participants to control for potential order effects. The road sign specified in each statement matched that shown in the corresponding image, but participants were not told this.

Each block contained 1 introduction page, 1 practice question, followed by 88 experimental questions and 2 attention-check questions. The introduction page specified whether the questions were asking for the participants' perceptions about themselves or the AI for the road signs. All the images shown were of road signs. Each question was presented on a separate page, showing an image of a road sign and either the human or AI agent statement for the participant to report how capable they thought the agent was at classifying that road sign. Participants had the option of including comments about the question, such as explaining why they gave the ratings that they did, before moving onto the next question. There were two unique images for each of the 11 road signs (stop, do not enter, pedestrian crossing, school zone, speed limit, yield, no turn on red, lane reduction, signal ahead, keep right, and roundabout), which resulted in a total of 22 original images. For each original image, the three types of manipulations (PDG attack, physical attack, and scrambled) were then applied separately, resulting in another 66 images. Participants were not informed about these manipulations in the instructions.

These 88 images were randomly ordered and presented within each of the two experimental blocks (the AI block and the human block). The two additional attention checks were also included in each block and were placed to be the 30th and the 60th questions in each block to evenly split the 90 total questions into three parts. Instead of an image of a road sign, each attention check had an image of a vehicle, with a statement, "For this question, you are required to choose the *strongly disagree* option below." If participants missed one or more of the attention checks, they were removed from the data analyses.

The post-task questionnaire asked about the safety criticality of the 11 road signs and demographic questions. For the safety criticality questions, the original 22 road-sign images were displayed in a randomized order, and participants rated the safety-criticality of each sign on a 7-point scale with the following statement, "This sign is safety-critical." As before, participants had the option to explain why they gave the rating that they did for each question. The nine demographic questions asked about the participant's age, gender, race, education, and driving experience.

C. Experimental Design

This study was comprised of two within-subjects independent variables (IVs). Each participant answered all three parts of the questionnaire. The first IV was the agent type, which was either the human or the AI. The agent type represented the entity that was classifying the image. This variable allowed for the comparison between how participants perceived their own abilities and their perception of the AI's computer vision capabilities for classifying road sign images.

The second IV was the attack type on the image, which consisted of four different attacks: no-attack (original image), PGD attack, physical attack, and scrambled manipulation. As described earlier, the original images should be correctly identified by both the human and AI agents. In contrast, the scrambled images should be difficult for both the human agent and AI agent to correctly classify the signs. The PGD attack and physical attack both projected masks over the original images, which the AI system should falter and fail to correctly classify [16]. For the human agent, the PGD attack was less visible, or detectable, to the human eye compared to the physical attack.

The dependent variables were the participants' agreement with each agent's statement for each image. The rating was on a scale of 1–7 for each statement (1 representing *strongly disagree*, 7 representing *strongly agree*), which reflected how the participant perceived their own or the AI's computer vision capability of classifying the road signs. In addition, participants subjectively rated how safety-critical each of the 11 road signs was (also on a 7-point Likert scale), and this safety-criticality rating was included as a covariate in the analyses of how agent type and attack type affected participants' perceived capabilities of the agent.

D. Procedures

Participants were recruited through MTurk and were all individuals who had an MTurk account. Upon starting the study, they read and acknowledged the online consent form. The introduction of the study explained that they would answer questions about AI and related applications, that it was about their opinions and personal ideas rather than a test of their knowledge, and

that they should not use any search engines to try and answer any of the questions. They were then directed to the pretask questions, measuring their trust in AI-in-general, trust in AI in autonomous driving, and self-confidence in their own driving skills. Next, they began the experimental task, either starting from the AI-agent block or the human-agent block. After the two experimental blocks were finished, participants completed the post-task questions, including the safety-criticality questions and the demographics questions. Once they had finished the last demographic question, they were debriefed and given a unique code to enter into MTurk to receive compensation.

IV. Results

A total of 64 (27.8%) participants were excluded from the data analyses. Among those participants, 14 did not have a driver's license, 34 were suspected of using automated techniques to complete the survey, and 16 did not pass one or more attention checks. For the remaining 166 participants, their average reported age was 37.04 years ($SD = 12.11$). Participants reported their gender as either male ($n = 116$) or female ($n = 50$). Participants reported an average of 16.22 years of driving experience ($SD = 13.04$, $n = 164$) or responded with "N/A" ($n = 2$). Participants reported their race as White ($n = 141$), Black or African American ($n = 12$), American Indian or Alaska Native ($n = 3$), Asian ($n = 4$), or other/mixed ($n = 6$). The following analyses were performed on these 166 participants.

A. Trust

Spearman's rank correlation analyses [39] with a Benjamini–Hochberg correction [40] were conducted among the trust measures, including trust in AI-in-general, trust in AI-in-AV, self-confidence in one's driving ability, and participants' rating on their perceived capability of the human agent and AI agent. The Benjamini-Hochberg correction was used because multiple correlations were conducted at once. Trust in AI-in-general was positively correlated to rating of capability for the human agent to correctly identify the images, $r_s(164) = 0.33$, $p < 0.001$, and for the AI agent, $r_s(164) = 0.43$, $p < 0.001$. Participants who had higher trust in AI-in-general ($M = 4.89$, $SD = 0.99$) rated both themselves ($M = 5.37$, $SD = 0.55$) and the AI agent ($M = 5.18$, $SD = 0.66$) to be more capable of classifying the images. Specifically, those who had higher trust in AI-in-general rated both the human and the AI agent to be more capable at classifying the original, PGD, and physical-attack images ($ps < 0.001$ for all comparisons), but not at classifying the scrambled images, $ps > 0.200$ (see Table I).

Trust in AI-in-AV was positively correlated with rating of capability for the AI agent to correctly identify the images, $r_s(164) = 0.28$, $p < 0.001$, but not with the rating for the human agent, $r_s(164) = 0.08$, $p = 0.332$. Participants who trusted AI-in-AV ($M = 4.49$, $SD = 1.19$) more rated AI ($M = 5.18$, $SD = 0.66$) to be more capable of classifying images. Specifically, those who trusted AI-in-AV more rated the AI agent to be more capable at classifying the original, PGD, and physical-attack images ($ps < 0.030$ for all comparisons), but not at classifying the scrambled images, $p = 0.267$ (see Table I). Trust in AI-in-AV was also positively correlated with trust in AI-in-general, $r_s(164) = 0.61$, $p < 0.001$. Participants who trusted AI-in-AV ($M = 4.49$, $SD = 1.19$) more also trusted AI-in-general more ($M = 4.89$, $SD = 0.66$).

TABLE I
Pearson's Correlation of Trust Measures and Ratings of Perceived Capabilities

Variable	*M*	*SD*	Trust in AI-in-general	Trust in AI-in-AV	Self-confidence in driving ability
Human	5.37	0.55	0.33***	0.08	0.19*
Original	6.26	0.69	0.36***	−0.01	0.22**
PGD	6.26	0.70	0.36***	0.01	0.20**
Physical	6.25	0.68	0.36***	0.01	0.22**
Scramble	2.55	1.60	−0.10	0.05	−0.13
AI	5.18	0.66	0.43***	0.28***	0.16
Original	6.07	0.81	0.46***	0.17*	0.21**
PGD	6.06	0.82	0.47***	0.18*	0.22**
Physical	6.05	0.81	0.45***	0.20**	0.20**
Scramble	2.53	1.51	−0.04	0.09	−0.14

Note. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Human = perceived capabilities of humans, AI = perceived capabilities of AI, Original = perceived capability of classifying original image, PGD = perceived capability of classifying PGD image, Physical = perceived capability of classifying physical-attack image, Scramble = perceived capability of classifying scrambled image.

Self-confidence in their own driving ability was positively correlated with ratings for the human agent, $r_s(164) = 0.19$, $p = 0.017$. Participants who had higher self-confidence in their own driving ability ($M = 5.60$, $SD = 0.91$) rated themselves ($M = 5.37$, $SD = 0.55$) to be more capable of classifying images. Specifically, those with more self-confidence in their own driving ability rated both the human and AI to be more capable at classifying the original, PGD, and physical images ($ps < 0.010$ for all comparisons), but not at classifying the scrambled images, $ps > 0.070$ for both comparisons (see Table I). No other correlations were significant, $ps > 0.050$.

B. Perception of Human and AI Capability

A 2 (agent: human, AI) x 4 (attack type: original, PGD, physical, scrambled) repeated-measures factorial analysis of covariance with safety criticality as a covariate was conducted to determine how each factor affected the rating of road-sign images, which reflected how the participants perceived their or AI's capability of classifying the road sign. The agent and attack type were both within-subjects variables. The assumption of sphericity was not met for all the terms involved, $ps < 0.001$, and thus Greenhouse–Geisser corrections were used.

After controlling for safety criticality, the main effect of agent type was significant, $F(1, 1824) = 6.64$, $p = 0.010$, $\eta_p^2 < 0.01$. Participants rated themselves ($M = 5.33$, $SD = 0.64$) to be more capable of correctly classifying road-sign images than the AI agent ($M = 5.18$, $SD = 0.73$). The main effect of attack type was also significant after controlling for safety criticality, $F(1.12, 2050.24) = 20.16$, $p < 0.001$, $\eta_p^2 = 0.01$. Participants rated the scrambled image ($M = 2.54$, $SD = 1.58$) to be less likely to classify than the original images ($M = 6.16$, $SD = 0.73$), the PGD images ($M = 6.16$, $SD = 0.73$), and the physical-attack

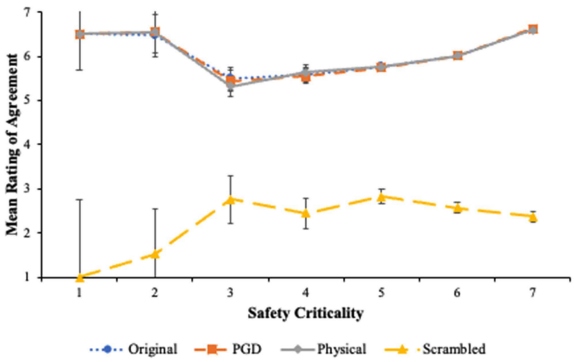

Fig. 2. Ratings of traffic sign images as a function of attack type and safety criticality. Error bars are 95% CIs of the means.

images ($M = 6.15$, $SD = 0.73$), $ps < .001$ for all comparisons, for both humans and the AI agent. There was no significant interaction between agent type and attack type, $p = 0.341$.

The covariate safety criticality was significant as well, $F(1, 1824) = 242.61$, $p < 0.001$, $\eta_p{}^2 = 0.12$. Participants rated signs with a higher safety criticality (5, 7] as more easily identifiable for both them and the AI agent than signs with a medium safety criticality [3], [5]. Interestingly, the interaction between attack type and the covariate safety criticality was also significant (see Fig. 2), $F(1.12, 2050.24) = 86.87$, $p < 0.001$, $\eta_p{}^2 = 0.05$. It is worth noting that the majority of road signs were rated to be at a high-level safety criticality (72.95% with ratings above 5). The medium-level safety criticality was the next largest group and ranged from 3 to 5 (26.40%). Very few road signs were rated to be at a low-level safety criticality being 1 or 2 ($ns =$ 3 and 9, respectively), leading to large variances at these two safety criticality levels. When safety criticality was low [1, 3), the ratings of the scrambled images increased, meaning they were more identifiable, as the safety criticality increased compared to the original, PGD, and physical-attack images, which decreased, meaning they were less identifiable. For medium to high safety criticality [3], [7], the scrambled-attack condition showed a distinct pattern from all other attack conditions, with the former showing a generally flat trend and the latter showing an increasing trend as safety criticality increased. Tests of simple main effects showed that for the original, PGD, and physical-attack images, the higher safety critical signs (5, 7] were rated to be significantly more identifiable by humans and AI than the medium safety critical signs [3], [5], $ps < 0.002$. However, this trend was not observed for the scrambled images. There was no significant interaction between agent type and the covariate safety criticality, $F < 1$. There was also no significant three-way interaction between agent type, attack type, and the covariate safety criticality, $p = 0.054$.

V. Discussion

This study showed that participants perceived their own capability to classify the road-sign images to be better than AI's capability. This result indicates that people may trust and rely on their own ability more than that of the AI system. Those who had higher self-confidence in their own driving ability rated the abilities of the human agent higher. Although participants rated themselves to be more capable than the AI system, they still rated the AI computer vision highly to correctly classify both the PGD images and the physical-attack images with ratings around 6 on the 7-point scale. Whereas the current AI algorithms should have a 0% chance of correctly classifying the image and fail at correctly classifying these attacked images [16], [36]. Our hypothesis, that participants would overestimate the AI's computer vision capabilities for the maliciously attacked images, was supported since participants still overestimated the AI's computer vision ability to detect the maliciously manipulated images, consistent with findings in [2]. This overestimation can likely cause the users to overtrust the AI system when encountering such malicious attacks. Those who trusted both AI-in-general and AI-in-AV more rated that the AI system was more capable compared to those who trusted them less.

People seem to experience difficulties with placing appropriate trust in the AI system [41]. In the context of word puzzles, people believe that AI supports their expectations and outperforms them for tasks [42]. However, in another context, voice recognition, people lose the willingness to rely on and trust a voice assistant after it fails to correctly recognize a spoken command [43], [44]. After a failure or error of AI, people lose trust in the AI's capability, but the findings from [45] suggest people are implicitly aware of the dangers of attributing blame to the AI for negative outcomes. Taken together, this may mean that people are unaware of the AI's computer vision limitations, but understand to not place too much reliance or blame on the system, particularly for certain tasks that require lower-level cognition, such as speech recognition and visual perception. In addition, humans are fallible to heuristics and poor memory recall, which can lead to inappropriate use and trust in AI [46]. It seems that humans need AI functions to be more transparent to understand its capabilities and limitations and be prepared for potential failures from the system.

Our prediction that participants would overestimate the AI's computer vision capabilities for the PGD attacks more than the physical attacks was not supported. One possible reason for this result was that participants were unaware of the limitations of the AI's computer vision for most of the malicious attacks, regardless of whether they were visible to human eyes or not. Thus, participants rated that the AI agent was able to correctly classify both the PGD attack and physical-attack images just as well as the original images. This result, in combination with the overestimation of AI's capabilities, as well as the positive correlation between trust and AI capabilities, again indicates overtrust in the AI system because people are unaware of the limitations of computer vision techniques. This overtrust can, thus, lead to potentially dangerous situations while on the road [7]. Drivers who overtrust the capabilities of the AI system may be unprepared to take control of the vehicle when the system fails or makes a misclassification. In addition, drivers who are unaware of how the AI system works may allow the system to continue working after an error or misclassification has occurred, potentially putting the driver and others in danger.

Participants also correctly rated that both the human and AI agents would have difficulty in correctly classifying the scrambled images, but both would be able to successfully classify the original images. This result supported our hypothesis that participants would perceive both themselves and the AI agent to be able to classify the original images and that neither would be able to correctly classify the scrambled-manipulation images. In addition, participants perceived both themselves and the AI agent as successfully classifying the original, PGD attack, and the physical-attack images, but not for the scrambled images, for almost all of the images. These results further support our hypothesis. Participants knew that AI was vulnerable to certain maliciously perturbed images, such as the scrambled images used in the current study, and that the AI system would fail at correctly classifying them while on the road [16]. In this regard, people correctly understand the AI's capability when dealing with scrambled images, or nonobjects, and thus, what is missing seems to be the knowledge about the AI's computer vision vulnerability to malicious attacks.

The safety criticality of the signs also influenced how participants perceived the human's and AI's capability of classifying the images. Participants rated that humans and the AI agent would more accurately classify the higher safety critical signs than the medium safety critical signs. Specifically, this trend was observed for the original, PGD, and physical-attack images but not for the scrambled images; the accuracy of classifying the scrambled images did not increase as safety criticality increased. These results indicate that participants may think safety is influential on the AI's computer vision ability to categorize the images. A possible explanation is that participants may think safety is influential on the AI's ability to categorize the images. It is possible that humans expect AI to focus more on—or are more capable of classifying—a high safety critical sign (e.g., a stop sign) than a low safety critical sign (e.g., a roundabout). Since this trend was also seen for humans, another possibility is that participants were more familiar with high-safety critical signs and had seen them more frequently than low-safety critical ones. They rated these road signs to be more identifiable to themselves and to the AI agent. However, the computer vision techniques used to process all images similarly and do not distinguish safety criticality between them [47]. The images carry equal weight in the identification process and should be unbiased toward high or low safety critical signs. This result again indicates that people may be unaware of how computer vision techniques work and instead believe that the AI's computer vision recognition and focus depend on the safety criticality of the sign. This result also supports that participants understood that AI was vulnerable to the scrambled manipulation and would have difficulties classifying the road-sign images, regardless of the safety criticality of the sign.

A. *Limitations and Future Research*

In this study, we focused on the identification of different types of road signs, which are static objects in the driving environment. This limitation prevents us from relating these findings to other roadway objects, such as a dynamic traffic light to understand how participants perceive the AI's computer vision capabilities in correctly classifying them. In future work, we can incorporate dynamic objects commonly found on the roadway, such as traffic lights, vehicles, and pedestrians. With these dynamic objects, we can assess how humans perceive AI's ability to classify perturbed images and if they believe AI is more or less capable of classifying these cases. Future work could also include images of objects that are extremely angled to see if participants understand computer vision techniques and their limitations regarding angled pictures.

The effect size was rather small for the significant main effect of agent type for participants' ratings of capabilities, indicating that the manipulation might not be as strong as we anticipated. The small effect size could be due to the subjective ratings of agents' capabilities not being a sensitive measure to reflect the difference between the agents. In addition, safety criticality violated the assumption of homogeneity of variance, which may be related to a lack of other significant effects, such as the interaction between agent type and safety criticality, and the three-way interaction between agent type, attack type, and safety criticality. Future studies could explore alternative procedures, study designs, and materials to better differentiate participant's perceptions of themselves and AI.

Another limitation of our study is the length of our study. It is possible that participants experienced fatigue as they went through the study, especially since the questions were all structured the same and repetitive. Future studies should reduce the number of experimental questions administered to reduce the chances of fatigue and boredom by the participants.

This study measured drivers' perceptions about current AI technologies in the field of DAS. Participants were not expected to have any prior knowledge specifically about AI prior to participating in this study because we were interested in their current perceptions and mental models of AI's computer vision capabilities. In addition, not many people have direct experience with DAS but may have indirect knowledge of it. However, one limitation of the study is that we did not measure drivers' knowledge or familiarity with AI technologies in DAS at the time of the study. It is possible that the sample collected comprised drivers with little to no knowledge of AI. Future studies can measure participants' knowledge and familiarity with AI systems in DAS and in general.

Finally, the method of using an online survey has its benefits and drawbacks. One drawback is that we were unable to control the setting which participants took the survey in. It is possible that they were completing other tasks and were not giving their undivided attention to the study. However, based on the participants' correct perception of their abilities to classify the original image and their inability to classify the scrambled images, it leads to the conclusion that the participants were paying attention throughout the study. In addition, with online surveys, it is difficult to collect the participants' cognitive thought processes for their rating reasoning. We did include an optional comment box for each of the 88 experimental questions where they could leave comments to explain their ratings. However, very few participants provided substantial comments. Future studies can include in-person experiments with human subjects rather

than online experiments. In addition, future studies can inquire about participants' reasoning for their ratings. For example, whether people detected the PGD and physical attacks. To our knowledge, no other study has investigated whether humans can detect the manipulations themselves in realistic driving situations through quantitative measures.

VI. Conclusion

This study shows that humans overestimate AI's capability of classifying maliciously manipulated images. Overall, participants rated humans to be more capable than AI in classifying perturbed images. However, people still perceive the AI system to be more capable than its actual ability when encountering malicious attacks. The visibility of the perturbations to the human eye did not affect how participants rated the AI's computer vision capability to successfully classify the image. This overestimation of AI may be an indication of overtrust in the AI system along with a lack of knowledge of how AI computer vision processes images and its vulnerabilities. A better understanding of how people perceive AI capabilities is still needed in order to aid appropriate trust calibration in AI computer vision systems.

References

[1] Society of Automotive Engineers (SAE), "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," Rep. Pub. no. J3016_202104, 2021.

[2] K. R. Garcia et al., "Drivers' understanding of artificial intelligence in automated driving systems: A study of a malicious stop sign," *J. Cogn. Eng. Decis. Mak.*, vol. 16, no. 4, pp. 237–251, 2022, doi: 10.1177/15553434221117001.

[3] A. Kaplan and M. Haenlein, "Siri, siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence," *Bus. Horiz.*, vol. 62, no. 1, pp. 15–25, 2019, doi: 10.1016/j.bushor.2018.08.004.

[4] J. McCarthy, "Artificial intelligence, logic, and formalising common sense," in *Machine Learning and the City: Applications in Architecture and Urban Design*. 2022, pp. 69–90, doi: 10.1002/9781119815075.ch6.

[5] R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. P. Campbell, "Introduction to machine learning, neural networks, and deep learning," *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, pp. 14–14, 2020, doi: 10.1167/tvst.9.2.14.

[6] M. Cunneen, M. Mullins, and F. Murphy, "Autonomous vehicles and embedded artificial intelligence: The challenges of framing machine driving decisions," *Appl. Artif. Intell.*, vol. 33, no. 8, pp. 706–731, 2019, doi: 10.1080/08839514.2019.1600301.

[7] Z. Halim, R. Kalsoom, S. Bashir, and G. Abbas, "Artificial intelligence techniques for driving safety and vehicle crash prediction," *Artif. Intell. Rev.*, vol. 46, no. 3, pp. 351–387, 2016, doi: 10.1007/s10462-016-9467-9.

[8] M. H. Jarrahi, "Artificial intelligence and the future of work: Human-ai symbiosis in organizational decision making," *Bus. Horiz.*, vol. 61, no. 4, pp. 577–586, 2018, doi: 10.1016/j.bushor.2018.03.007.

[9] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, "Beyond accuracy: The role of mental models in human-ai team performance," in *Proc. AAAI Conf. Hum. Comput. Crowdsourcing*, 2019, pp. 2–11.

[10] V. Kakani, V. H. Nguyen, B. P. Kumar, H. Kim, and V. R. Pasupuleti, "A critical review on computer vision and artificial intelligence in food industry," *J. Agriculture Food Res.*, vol. 2, 2020, Art. no. 100033, doi: 10.1016/j.jafr.2020.100033.

[11] K. Bengler, K. Dietmayer, B. Farber, M. Maurer, C. Stiller, and H. Winner, "Three decades of driver assistance systems: Review and future perspectives," *IEEE Intell. Transp. Syst. Mag.*, vol. 6, no. 4, pp. 6–22, Winter 2014, doi: 10.1109/MITS.2014.2336271.

[12] G. Bathla et al., "Autonomous vehicles and intelligent automation: Applications, challenges, and opportunities," *Mobile Inf. Syst.*, vol. 2022, no. 1, 2022, Art. no. 7632892, doi: 10.1155/2022/7632892.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015, *arXiv:1412.6572*.

[14] S. Nagpal, M. Singh, R. Singh, and M. Vatsa, "Deep learning for face recognition: Pride or prejudiced?," 2019, *arXiv:1904.01219*.

[15] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. 1st Conf. Fairness, Accountability Transparency*, 2018, pp. 77–91.

[16] S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," *Appl. Sci.*, vol. 9, no. 5, 2019, Art. no. 909, doi: 10.3390/app9050909.

[17] S. Hu, T. Yu, C. Guo, W.-L. Chao, and K. Q. Weinberger, "A new defense against adversarial images: Turning a weakness into a strength," in *Proc. 33rd Conf. Neural Inf. Process. Syst.*, 2019, pp. 1–12.

[18] Y. Ma, Z. Wang, H. Yang, and L. Yang, "Artificial intelligence applications in the development of autonomous vehicles: A survey," *IEEE/CAA J. Automatica Sinica*, vol. 7, no. 2, pp. 315–329, Mar. 2020, doi: 10.1109/JAS.2020.1003021.

[19] X. Zhang, H. Gao, M. Guo, G. Li, Y. Liu, and D. Li, "A study on key technologies of unmanned driving," *CAAI Trans. Intell. Technol.*, vol. 1, no. 1, pp. 4–13, 2016, doi: 10.1016/j.trit.2016.03.003.

[20] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, pp. 5580–5590.

[21] H. Lengyel, V. Remeli, and Z. Szalay, "Easily deployed stickers could disrupt traffic sign recognition," *Perner's Contacts (Special Issue)*, vol. 2, no. 19, pp. 156–163, 2019.

[22] L. Jöckel, M. Kläs, and S. Martínez-Fernández, "Safe traffic sign recognition through data augmentation for autonomous vehicles software," in *Proc. IEEE 19th Int. Conf. Softw. Qual., Rel. Secur. Companion*, 2019, pp. 540–541.

[23] H. Liu et al., "Trustworthy AI: A computational perspective," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 1, pp. 1–59, 2023, doi: 10.1145/3546872.

[24] K. D. Apostolidis, E. V. Gkouvrikos, E. Vrochidou, and G. A. Papakostas, "Traffic sign recognition robustness in autonomous vehicles under physical adversarial attacks," in *Cutting Edge Applications of Computational Intelligence Tools and Techniques*. Cham, Switzerland: Springer Nature, 2023, pp. 287–304.

[25] B. C. Matei and P. Meer, "Estimation of nonlinear errors-in-variables models for computer vision applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1537–1552, Oct. 2006, doi: 10.1109/tpami.2006.205.

[26] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018, doi: 10.1109/access.2018.2807385.

[27] A. Prahl and W. W. P. Goh, ""Rogue machines" and crisis communication: When AI fails, how do companies publicly respond?," *Public Relations Rev.*, vol. 47, no. 4, 2021, Art. no. 102077, doi: 10.1016/j.pubrev.2021.102077.

[28] J. Chen, S. Mishler, and B. Hu, "Automation error type and methods of communicating automation reliability affect trust and performance: An empirical study in the cyber domain," *IEEE Trans. Hum.-Mach. Syst.*, vol. 51, no. 5, pp. 463–473, Oct. 2021, doi: 10.1109/thms.2021.3051137.

[29] K. A. Hoff and M. Bashir, "Trust in automation," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 57, no. 3, pp. 407–434, 2015, doi: 10.1177/0018720814547570.

[30] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 46, no. 1, pp. 50–80, 2004, doi: 10.1518/hfes.46.1.50_30392.

[31] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 39, no. 2, pp. 230–253, 1997, doi: 10.1518/001872097778543886.

[32] O. Asan, A. E. Bayrak, and A. Choudhury, "Artificial intelligence and human trust in healthcare: Focus on clinicians," *J. Med. Internet Res.*, vol. 22, no. 6, 2020, Art. no. e15154, doi: 10.2196/15154.

[33] S. Russell, I. S. Moskowitz, and A. Raglin, "Human information interaction, artificial intelligence, and errors," in *Autonomy and Artificial Intelligence: A Threat Or Savior?*, 1st ed. Berlin, Germany: Springer, 2017, ch. 4, pp. 71–101.

[34] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2019, *arXiv:1706.06083*.

[35] Y. Deng and L. J. Karam, "Universal adversarial attack via enhanced projected gradient descent," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 1241–1245.

[36] K. Eykholt et al., "Robust physical-world attacks on deep learning models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1625–1634.

[37] S. Atakishiyev, M. Salameh, and R. Goebel, "Safety implications of explainable artificial intelligence in end-to-end autonomous driving," 2024, *arXiv:2403.12176*.

[38] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cogn. Ergonom.*, vol. 4, no. 1, pp. 53–71, 2000, doi: 10.1207/S15327566IJCE0401_04.

[39] C. Spearman, "The proof and measurement of association between two things," *Amer. J. Psychol.*, vol. 15, no. 1, pp. 72–101, 1904, doi: 10.2307/1412159.

[40] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Stat. Soc.: Ser. B (Methodological)*, vol. 57, no. 1, pp. 289–300, 2018, doi: 10.1111/j.2517-6161.1995.tb02031.x.

[41] N. Banovic, Z. Yang, A. Ramesh, and A. Liu, "Being trustworthy is not enough: How untrustworthy artificial intelligence (ai) can deceive the end-users and gain their trust," in *Proc. ACM Hum.-Comput. Interaction*, 2023, pp. 1–17.

[42] T. Kosch, R. Welsch, L. Chuang, and A. Schmidt, "The placebo effect of artificial intelligence in human–computer interaction," *ACM Trans. Comput.-Hum. Interaction*, vol. 29, no. 6, pp. 1–32, 2022, doi: 10.1145/3529225.

[43] A. Cuadra, S. Li, H. Lee, J. Cho, and W. Ju, "My bad! repairing intelligent voice assistant errors improves interaction," *Proc. ACM Hum.-Comput. Interaction*, vol. 5, no. 27, pp. 1–24, 2021, doi: 10.1145/3449101.

[44] A. Baughan, X. Wang, A. Liu, A. Mercurio, J. Chen, and X. Ma, "A mixed-methods approach to understanding user trust after voice assistant failures," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2023, pp. 1–16.

[45] M. T. Stuart and M. Kneer, "Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelligent agents," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, pp. 1–27, 2021, doi: 10.1145/3479507.

[46] G. Gigerenzer, "Psychological ai: Designing algorithms informed by human psychology," *Perspectives Psychol. Sci.*, vol. 19, pp. 1–10, 2023, doi: 10.1177/17456916231180597.

[47] N. Buch, S. A. Velastin, and J. Orwell, "A review of computer vision techniques for the analysis of urban traffic," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 3, pp. 920–939, Sep. 2011, doi: 10.1109/tits.2011.2119372.