PURDUE UNIVERSITY

PennState

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

Katherine Zhang*, Purdue University
Claire Chen, Pennsylvania State University
Dr. Aiping Xiong, Pennsylvania State University

Background

- AI in autonomous vehicles (AVs) are vulnerable to attack [1, 2]
  - Physical-world attacks: Tampering with physical objects on the road to cause AI errors
- In an SAE level 3 automation system, when the AI fails, the human driver needs to take over
- For safe operation, human drivers need to be aware of attacks and AI vulnerabilities
  - Communication from the vehicle concerning risk, AI behavior, etc.

[1]: K. Eykholt et al., “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.
[2]: Y. Man et al., “{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.
Image: https://www.unity.de/en/services/systems-engineering/automated-driving-through-systems-engineering/

Prior Work

- Used in this study:
  - Manipulated stop sign [1]
  - Dirty-road patch [2]
- Humans are capable of detecting attack conditions that AI cannot
- But, little research done on whether humans can detect attacks as being *causes of AI errors*
- They tend to trust the AI to function normally under attack conditions [3]
  - Over-trust ➔ increased risk
  - Vehicle must communicate risks to driver
- Unclear what information drivers need to increase their situation awareness

[1]: K. Eykholt et al., “Physical adversarial examples for object detectors,” in 12th USENIX Workshop on Offensive Technologies, 2018.
[2]: T. Sato et al., “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.
[3]: K. R. Garcia et al., “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.

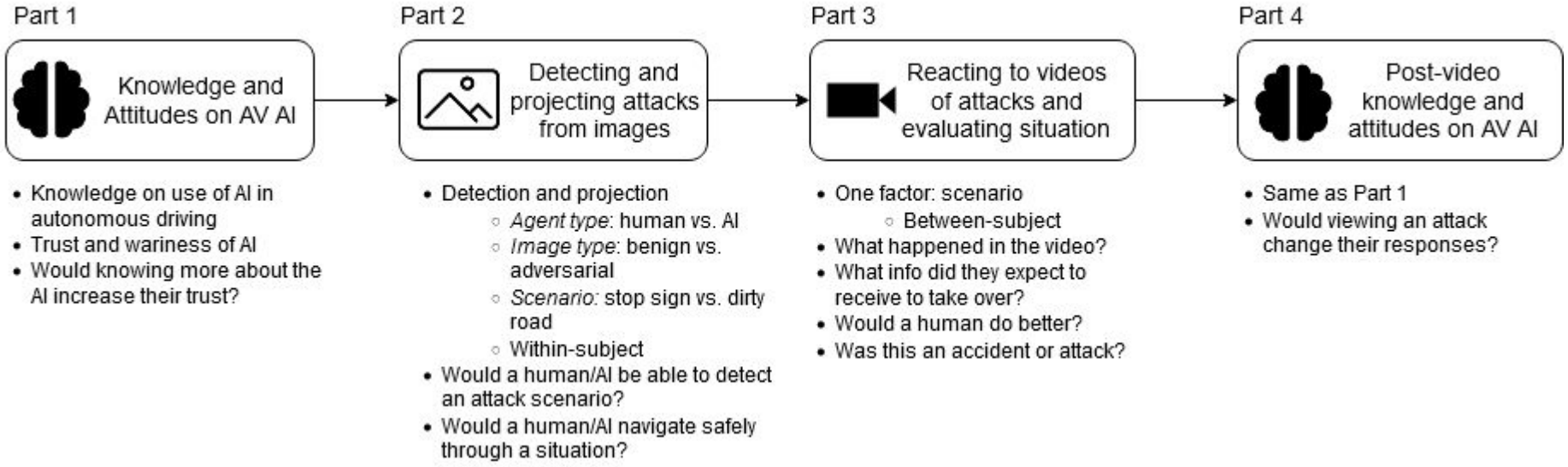
Research Questions

Are human drivers able to **detect** physical-world attacks and **project** how they affect AI's driving?

What **information do human drivers expect** to make them **more aware** of attacks and take over if needed?

Methodology

- Qualtrics survey, N = 100
  - Participants were car owners, recruited through Prolific



Methodology - Part 2

- Attack scenario images - detection and projection
  - How much do participants agree with the following statements [7-point Likert scale: Completely disagree (1)–Completely agree (7)]?
  - Detection task:
    - *I* think the image shows the lane lines clearly.
    - I think the *current AI system in AVs* will detect the lane lines of the road in the image.
  - Projection task:
    - I think a *human* driver will navigate the above road condition safely.
    - I think the *current AI system in AVs* will navigate the above road condition safely.

Benign condition

Adversarial condition

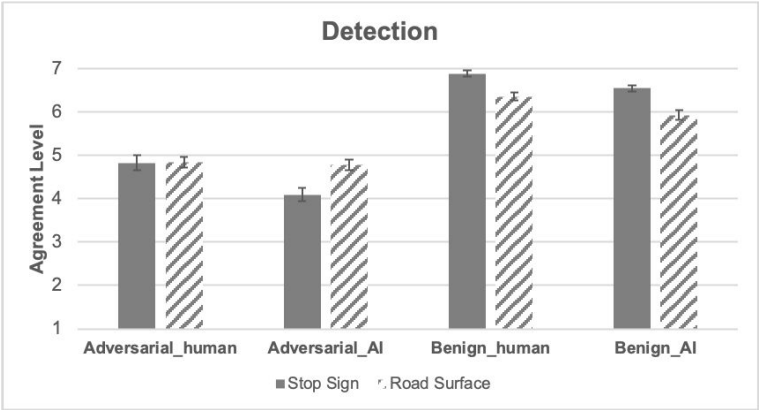
Methodology - Part 3

- Video of attack
  - Scenarios were not called “attacks”
  - Asked to imagine that they were driving in an AV using autonomous driving, and that they needed to be alert to potentially take control
  - After video, asked two free-response questions
    - What do you think happened in the video?
    - What information should the AI system provide about the situation so that human drivers can *safely* take over control?

(GIF only shows key moments, not full video used)

Results - Quantitative (Part 2)

- Detection tasks
  - Participants can detect benign vs. adversarial ($F_{(1,99)} = 336.34$, $p < .001$)
  - Generally rated AI lower than humans ($F_{(1,99)} = 36.18$, $p < .001$)
  - 3-way interaction of agent type * image type * scenario ($F_{(1,99)} = 15.22$, $p < .001$):
    - In adversarial conditions, scenarios were rated the same difficulty for humans, but for AI, the stop sign was rated more difficult
    - Stop sign studies have been out for some time and may be known to the public



Results - Qualitative (Part 3)

Most common responses to: "*What information should the AI system provide about the situation so that human drivers can safely take over control?*"

Dirty Road (N = 43):

- An alert (audio, audio + visual)
- Explanation of AI errors or decision-making
  - Vehicle should explain that the AI *"...had detected something impairing its ability to make judgement on the road condition."* (P47)
- Request for the driver take over
  - Vehicle should indicate *"That the AI is unable to safely navigate and the human will need to interact immediately."* (P34)

Stop Sign (N = 57):

- Explicit mention of the stop sign
  - *"The AI system should acknowledge there's a stop sign approaching..."* (P81)
- Alerts and explanations, like dirty road
  - *"[The AI] should tell the human that it is having trouble telling if there is a stop sign."* (P80)
- 25 unsure or confused
  - May not have understood that they needed to take over

Discussion

- Participants seem to be unaware that scenarios were attacks
  - In the dirty road scenario, many participants noted AI error, but still treated it more as an accident rather than an attack
  - Could be due to participants drawing from prior driving experience
    - Ex: Dirty road attack could be interpreted as black ice causing car to slip
  - Shows that mental representations of driving are based on prior experience, which can cause misconceptions when considering AV driving
- Different attacks are perceived differently
  - Different situations require different tasks and effort
  - Influences risk assessment
- Participants have different expectations of what info to receive to take over control
  - Also dependent on scenario

Future Work

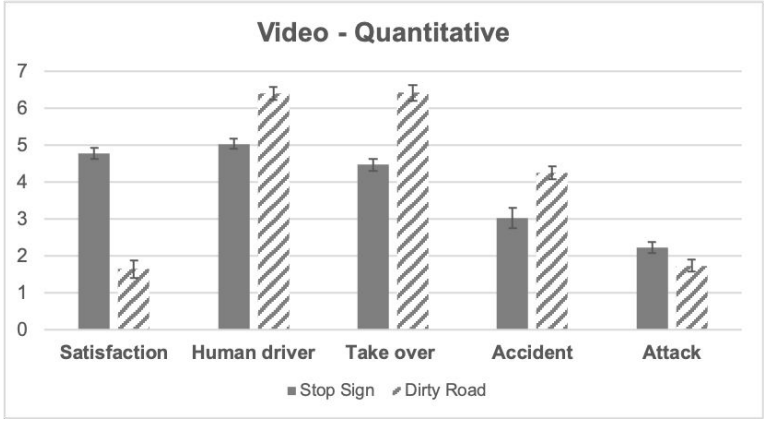
- Conduct study with a driving simulator
  - Get participant reactions more akin to that of real driving
- Further investigate in-car risk communication
  - How do drivers react and feel when the vehicle conveys information?
  - What information leads to safer driving?
- Investigate more attack types
- Experiment with varying factors in attacks in more detail
  - E.g.: object positioning, different textures, etc.

Acknowledgments

Thank you to Dr. Aiping Xiong, my advisor and co-author, and Claire Chen, teammate and co-author.

Results - Quantitative (Part 3)

- Avg. agreement level
  - Significantly less satisfied with AI performance in dirty road video ($t_{(1,98)} = -10.63$, $p < .001$)
  - Humans would handle it better in the dirty road scenario ($t_{(1,98)} = 5.79$, $p < .001$)
  - Participants wanted take over more in the dirty road scenario ($t_{(1,98)} = 6.66$, $p < .001$)
  - Believed neither scenario was due to attack, more likely an accident
    - While more people in dirty road condition noted AI error, they treated it more as an accident ($t_{(1,98)} = 3.98$, $p < .001$), and less as an attack ($t_{(1,98)} = -2.16$, $p = .033$)



Limitations

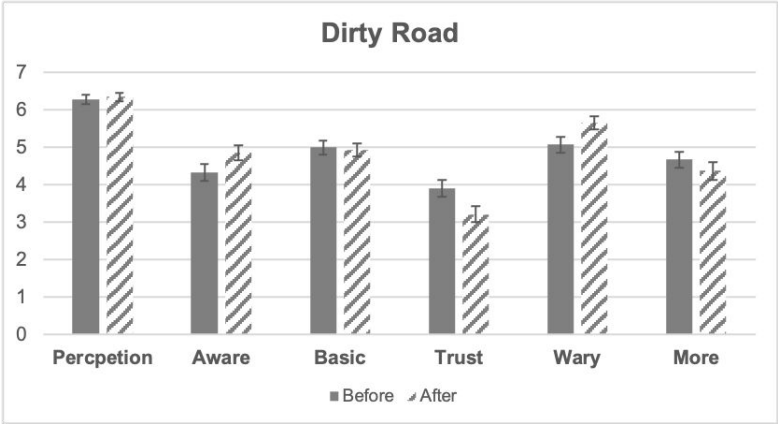
- Only investigated two cases of attack
  - Many other potential types of attacks
  - Many different variables that can affect perception (object texture, shape, position, etc.)
- We only used simulated driving videos/images, and framed them as hypothetical scenarios
- Minor extraneous differences between benign and adversarial images (e.g. size of the oncoming truck in the dirty road images)
  - Likely negligible effect on results, but can be better controlled in future work
- Only 100 participants
- Sample might not be reflective of general population
  - Most were 18-44 years old, with at least some college

Methodology - Part 3 cont.

- How much did participants agree with the following statements (7-point Likert scale)?
  - I am satisfied with the AV's behavior in the situation.
  - I would drive more safely than AI in this situation.
  - I would take over the AI's driving in this situation.
  - I believe this situation was caused by accident
  - I believe this situation was caused by intentional attack.

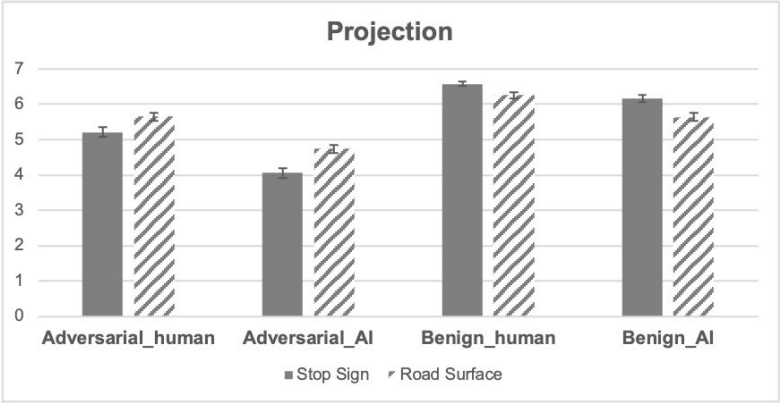
Results - Quantitative (Parts 1 & 4)

- Knowledge, attitudes before and after
  - Mostly similar between before and after, and between different scenarios
  - Both scenario groups were made more aware of how AVs use AI to perceive the environment after viewing the video ($F_{(1,98)} = 20.73$, $p < .001$)
  - Participants who viewed dirty road had lowered trust after video ($F_{(1,98)} = 4.56$, $p = .035$)
  - Accordingly, their wariness also went up ($F_{(1,98)} = 12.36$, $p < .001$)



Results - Quantitative (Part 2)

- Projection tasks
  - Rated both human and AI less capable of driving in adversarial conditions ($F_{(1,99)} = 275.1$, $p < .001$)
  - Also rated AI less capable than humans, but more so with dirty road ($F_{(1,99)} = 110.92$, $p < .001$)



Results - Qualitative (Part 3)

Most common response content to: "*What happened in the video?*"

Dirty Road (N = 43):

- AI malfunction, error, confusion
  - "The AI confused the blurriness on the ground and did not stay in its lane." (P100)
- Incorrect AI lane detection
- Road surface condition (marks, ice, etc)
  - "...[the] mark on the ground could have been a patch of black ice..." (P76)

Stop Sign (N = 57):

- Most believed car/AI stopped at the sign
- Fewer believed that AI did not stop at sign, or human had to intervene
  - "...the AI ignored the stop sign and the human had to stop the car themselves." (P2)
- Only 3 mentioned malignant sign
  - "...the entity in charge of driving encountered a somewhat odd looking (possibly vandalized) stop sign..." (P97)