Università
degli Studi
di Padova

SPRITZ
Security & Privacy
Research Group

Crash Course

Can (Under Attack) Autonomous Driving Beat Human Drivers?

Francesco Marchiori[1], Alessandro Brighente[1], Mauro Conti[1,2]
[1]*University of Padova, Italy*
[2]*TU Delft, Netherlands*


ME,
6 MONTHS
AGO

Adversarial Attacks

Clean Input Sample
X_c

Cat

Adversarial Attacks

Clean Input Sample
X_c

Perturbation

Cat

Adversarial Attacks

Clean Input Sample
X_c

Perturbation

Adversarial Sample
X_a

Cat

Dog

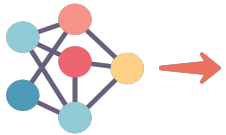
Adversarial Attacks

Clean Input Sample
X_c

Perturbation

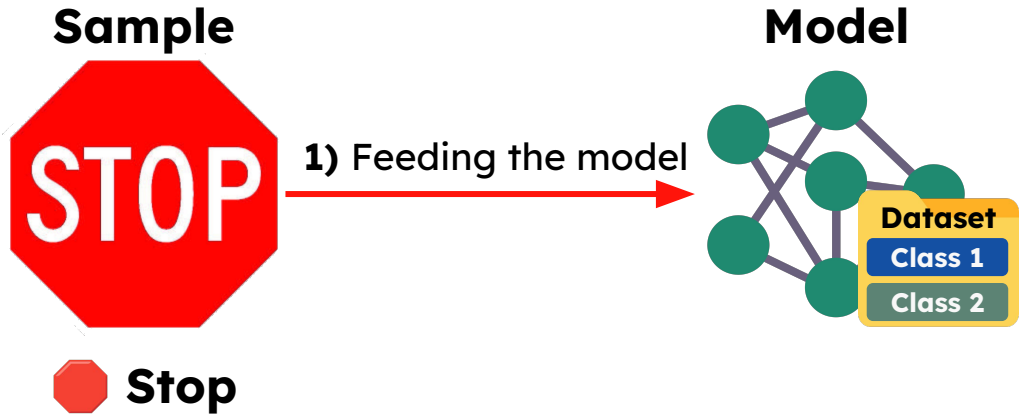
Adversarial Sample
X_a

Stop

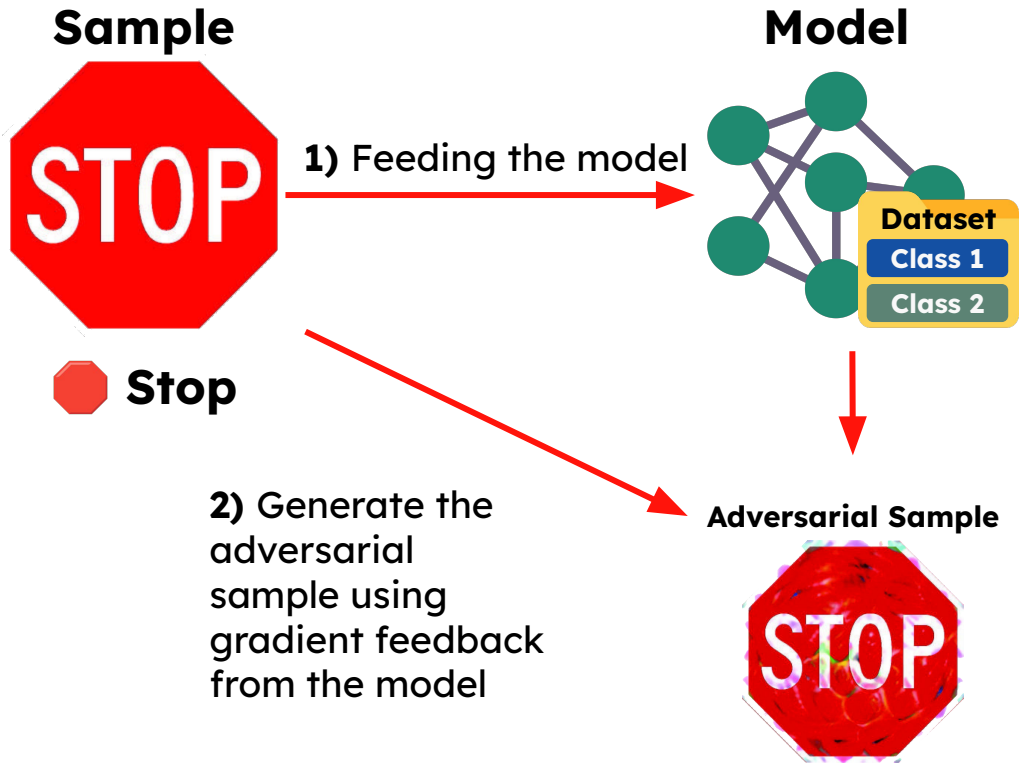


50 km/h limit

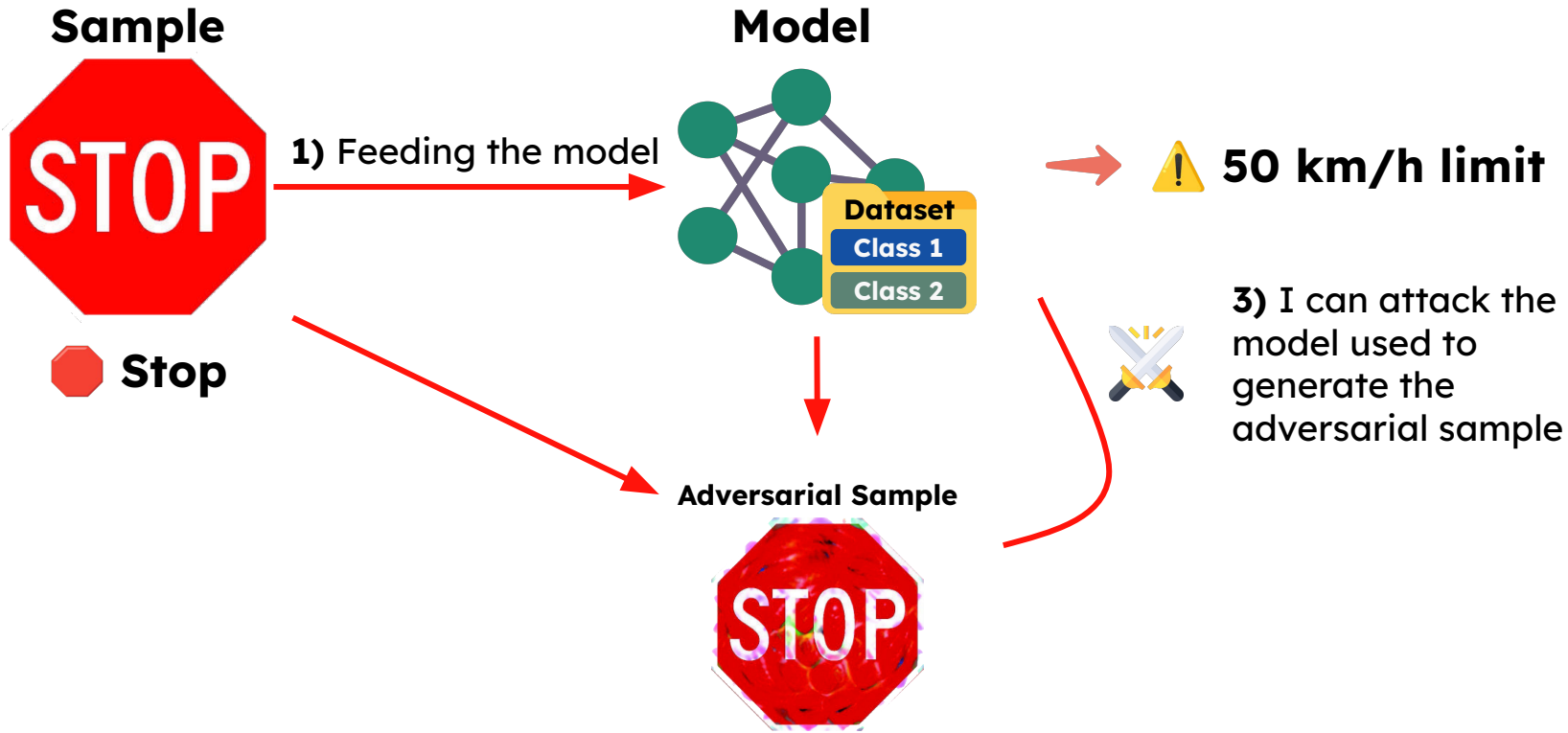
Transferability



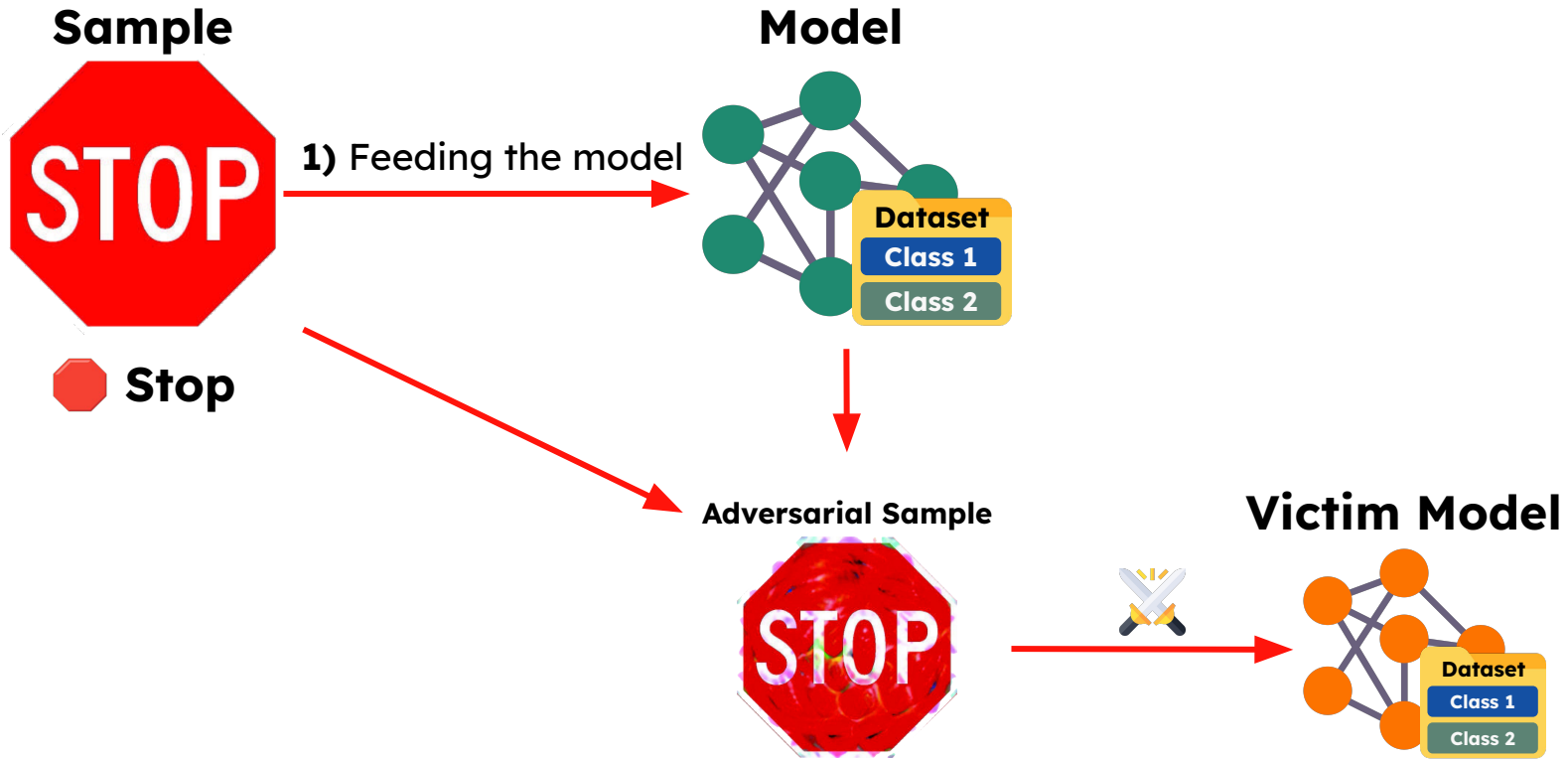
Transferability



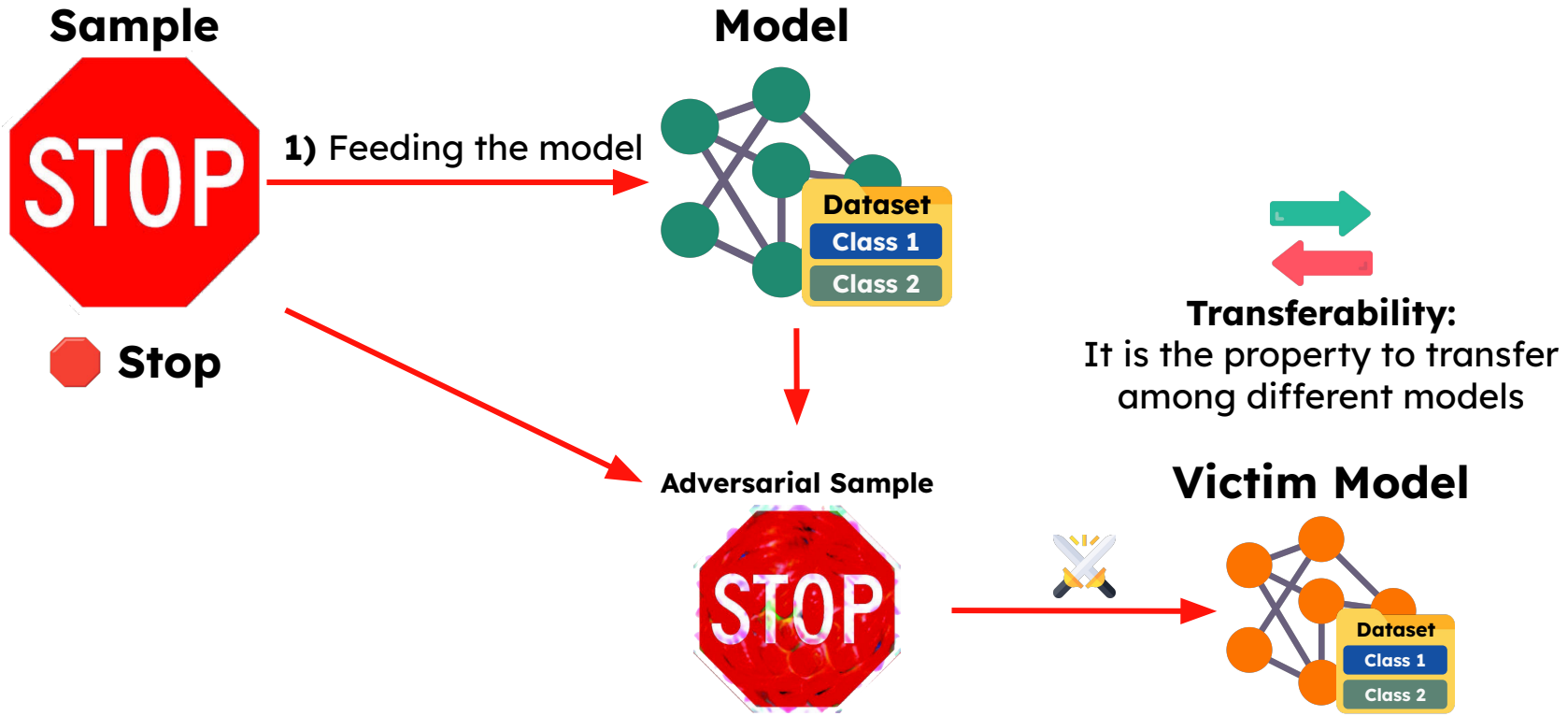
Transferability



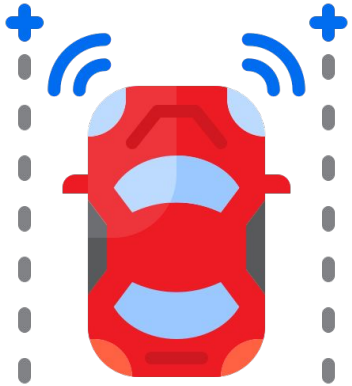
Transferability



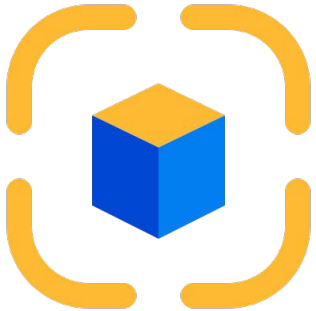
Transferability



Autonomous Tasks



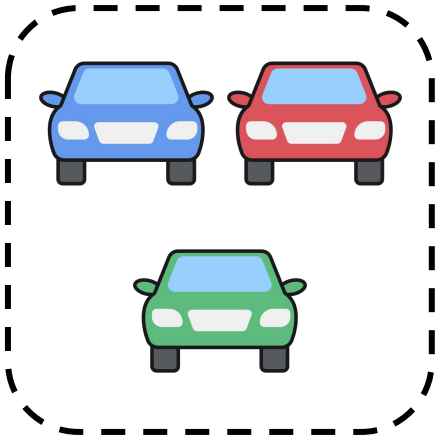
Lane keeping



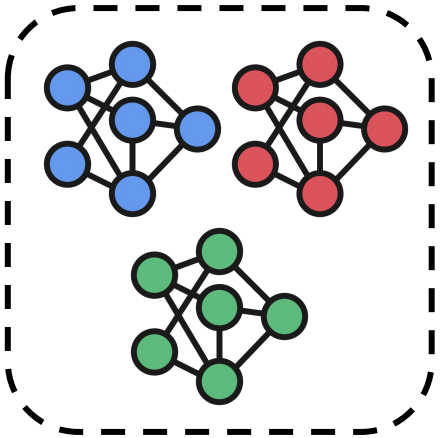
Object recognition

Autonomous driving

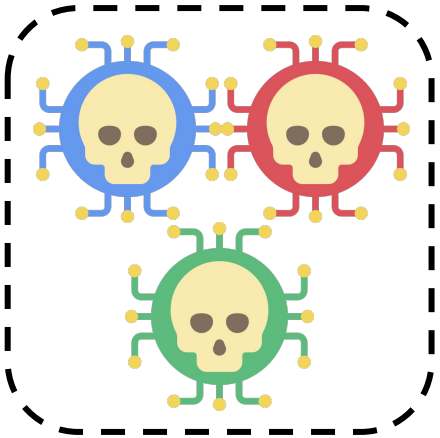
Threat Scenarios



Datasets

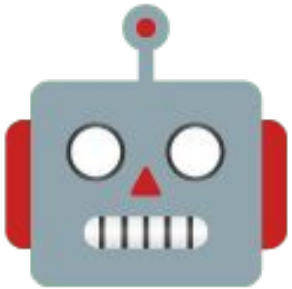


Models



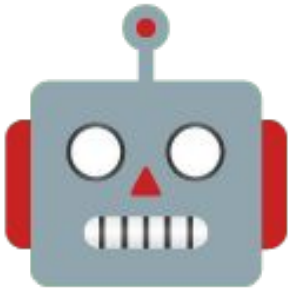
Attacks

"Research" Question



>?

"Research" Question



>?

How much automation?

How much knowledge for the attacker?

Which aspects are more impacted?

Crash Course

- Evaluation of vulnerabilities of autonomous driving
  - All levels of automation
  - Different attacker scenarios
- Realistic threat model
  - Differences between adversarial attacks assumptions and real attackers
- Requirements identification
  - Attacks
  - Countermeasures

Outline

1. Introduction
2. **Automation**
3. Assumption Criteria
4. Evaluation
5. Conclusions

SAE Levels

No Automation

Partial Assistance

Partial Automation

Conditional Automation

High Automation

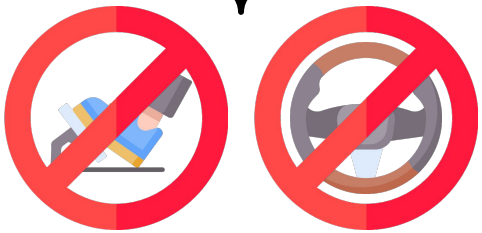
Full Automation

SAE Levels

🤖 **Autonomous features**

0	1	2	3	4	5
No Automation	Partial Assistance	Partial Automation	Conditional Automation	High Automation	Full Automation

🆘 **Driver aid**



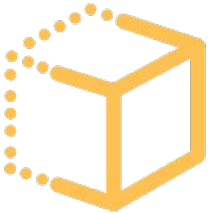
AI on SAE Levels

Level	Automation	Example Features	AI	Driver	Example Tasks
0	-	-	○	●	-
1	Partial Assistance	Adaptive Cruise Control (ACC)	●	●	Decision making
		Lane departure warning	●	●	Detection, sensor fusion
2	Partial Automation	ACC	●	●	Decision making
		Lane keeping assistance	●	●	Detection, sensor fusion
		Driver monitoring	●	●	Biometrics analysis
		Traffic jam assistant	●	●	Traffic pattern recognition
3	Conditional Automation	Environment monitoring	●	○	Sensor fusion
		Traffic jam autopilot	●	◐	Autonomous decision making
		Driver disengagement	●	◐	Autonomous decision making
		Autonomous driving	●	◐	Lane change, navigation
4	High Automation	Navigation in geofenced areas	●	○	Path planning
		Autonomous decision making	●	○	Traffic management
		Safety overrides	●	◐	Limited safety-critical tasks
5	Full Automation	Safety and redundancy	●	○	Anomaly detection
		V2X communications	●	○	Resource optimization
		Navigation	●	○	Autonomous navigation

●: present, ○: not present, ◐: partially present.

AI and Sensors

Road Signs



LiDAR

Semantic Segmentation

Outline

1. Introduction
2. Automation
3. **Assumption Criteria**
4. Evaluation
5. Conclusions

Adversarial Techniques

Evasion Attacks

Poisoning Attacks

Adversarial Techniques

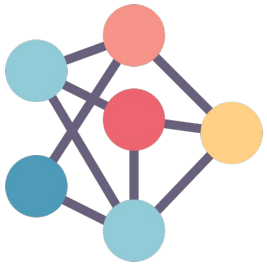
Evasion Attacks

oisoning

Assumptions

Model parameters

Assumptions



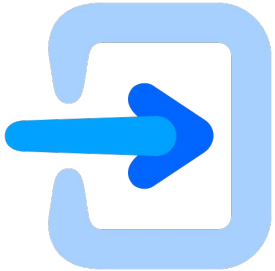
Model parameters

Model output

Assumptions

Model parameters

Model output

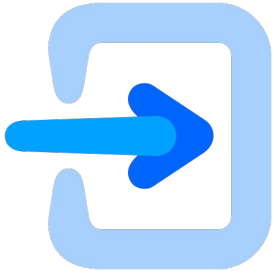


Direct input

Assumptions

Model parameters

Model output



Direct input

Physical implementation

Related Works

Attack	Misclassification Task	Model Parameters	Model Output	Direct Input	Physical Implementation
Arnab et al. [3]	Semantic Segmentation	●	●	●	○
Brown et al. [4]	Road Sign	●	●	○	●
Cao et al. [5]	LiDAR	●	●	○	●
Cao et al. [6]	LiDAR	○	●	○	●
Eykholt et al. [7]	Road Sign	●	●	○	●
Kong et al. [12]	Road Sign	○	●	○	●
Kumar et al. [13]	Road Sign	○	●	●	○
Li et al. [15]	Road Sign	○	●	●	○
Ma et al. [17]	Object Tracking	●	●	○	●
Papernot et al. [19]	Road Sign	○	●	●	○
Sharma et al. [22]	Misbehavior Detection	○	●	●	○
Sitawarin et al. [23]	Road Sign	○	●	●	○
Xiang et al. [25]	LiDAR	●	●	●	○
Zhu et al. [28]	LiDAR	○	●	○	●

●: required, ○: not required.

Threat Model

No model architecture

Threat Model

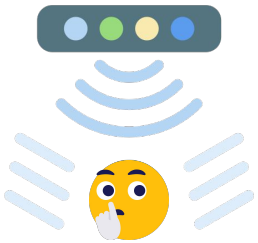
No model architecture

No training dataset

Threat Model

No model architecture

No training dataset

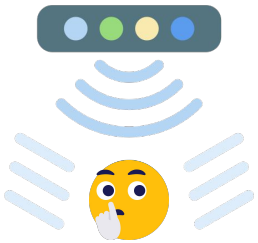


Constrained sensor manipulation

Threat Model

No model architecture

No training dataset



Constrained sensor manipulation

Environmental variability

Outline

1. Introduction
2. Automation
3. Assumption Criteria
4. **Evaluation**
5. Conclusions

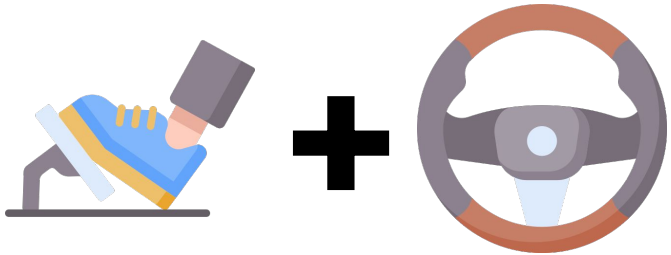
Threat Model Evaluation (1/2)

- **Level 1** - Partial Assistance
  - Limited functionality (steering or accelerating)
  - Restricted attack surfaces

Threat Model Evaluation (1/2)

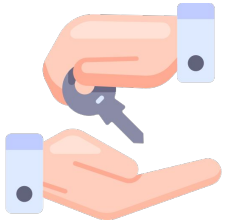
- **Level 1** - Partial Assistance
  - Limited functionality (steering or accelerating)
  - Restricted attack surfaces

- **Level 2** - Partial Automation
  - Augmented functionality (steering and accelerating)
  - Exploiting interaction



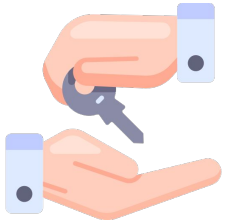
Threat Model Evaluation (2/2)

- **Level 3 -** Conditional Automation
  - Still requires driver attention
  - Challenges during handover
  - More attack surfaces to be exploited

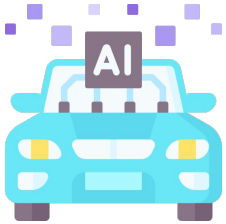


Threat Model Evaluation (2/2)

- **Level 3 -** Conditional Automation
  - Still requires driver attention
  - Challenges during handover
  - More attack surfaces to be exploited



- **Level 4 / Level 5**
  - Important to have architecture confidential
  - Ethical considerations to be exploited for malicious purposes



Criteria

Ease of attack

Criteria

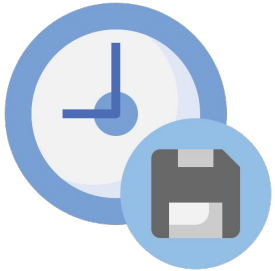
Ease of attack

Response time

Criteria

Ease of attack

Response time

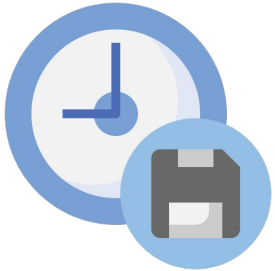


Recovery time

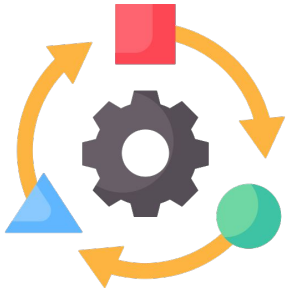
Criteria

Ease of attack

Response time



Recovery time



Adaptability

Criteria

Level	Ease of Attack	Response Time	Recovery Time	Adaptability
1	●	◐	○	○
2	●	◐	○	○
3	◐	●	◐	◐
4	○	●	●	●
5	○	●	●	●

●: increased safety.

◐: unclear.

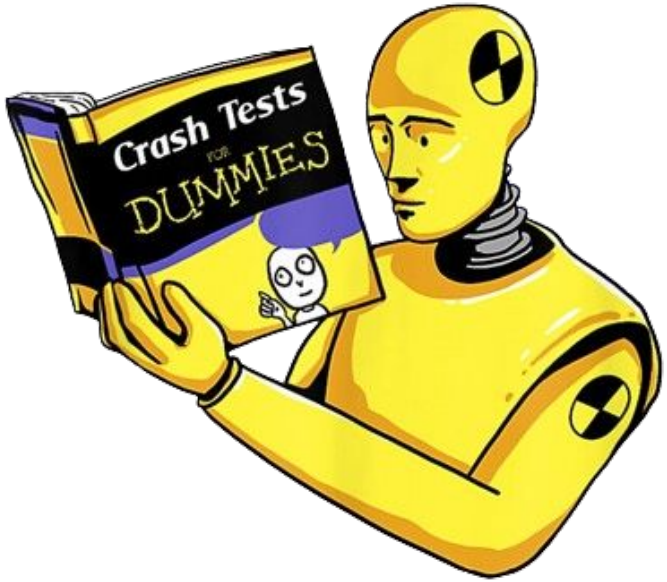
○: no improvement or decreased safety.

Outline

1. Introduction
2. Automation
3. Assumption Criteria
4. Evaluation
5. **Conclusions**

Takeaways

- Security by obscurity?
  - Model knowledge is critical for attack
  - Dependent on other factors (e.g., data, balance)
- Operational Design Domains (ODDs)
  - Defining operating conditions
  - Safe engage of autonomous components
- Threat modelling
  - Crucial to define attacker's assumptions
  - Targeted defenses (e.g., adversarial training)



Future Work

- Empirical validation
  - Testbed (simulated)
  - Multiple adversarial challenges
  - Feasibility and practicality

- Adaptability of AI systems to different adversarial strategies
  - Diverse level of SAE automation
  - Targeted countermeasures

Università
degli Studi
di Padova

SPRITZ
Security & Privacy
Research Group

Thank you for the attention

francesco.marchiori@math.unipd.it