# Defense Against Adversary: How safe is Machine Learning for Autonomous Driving?

Benjamin Danek, Computer Science
Mentor: Yi Ren, Assistant Professor
School for Engineering of Matter, Transport and Energy

## Research Question:

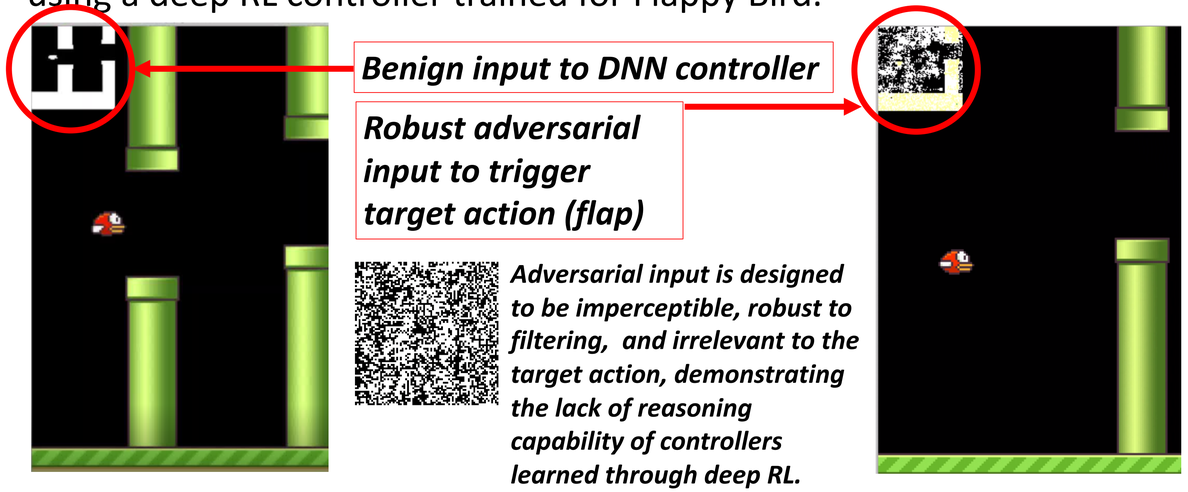
Are *Reinforcement Learning* models vulnerable to visual attacks?

## Background:

- Reinforcement Learning (RL) is used to perform end-to-end training of controllers for autonomous systems (e.g., cars [2,3])
- Deep neural networks (DNN) are used to facilitate highly nonlinear mapping from system states to actions or action-values.
- DNNs for vision applications are found to be vulnerable to attacks.
- If DNNs for RL are non-robust (e.g., against visual attacks or due to training data corruption), the autonomous systems built on top of DNNs may have hidden yet severe safety/security loopholes.

## Results:

This study examines and confirms the existence of such loopholes, using a deep RL controller trained for Flappy Bird.




## Research Methods & Progress

- Use gradient based optimizer to solve:

$$min_{\Delta s} \sum_{\{s^i, a^i, s^{i\prime}, r^i\} \in B_{\theta^*}} [\, l(Q(s^i + \Delta s, \bar{a^i}, \theta), Q(s^i + \Delta s, a^i, \theta)],$$

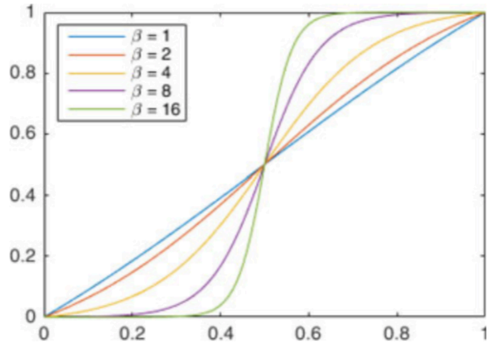
$$l(a, b) = \max(b - a + \varepsilon, 0),$$

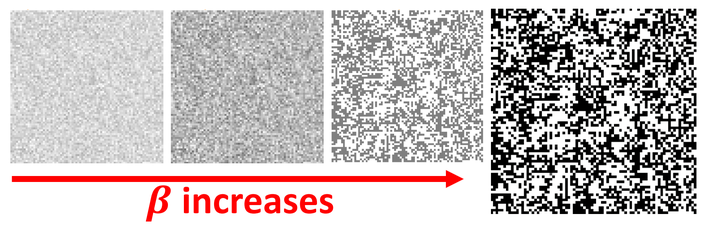
$B_{\theta^*}$ is the memory set of a trained controller $\theta^*$

- When adversarial noise ($\Delta s$) is added to input, the DNN ($Q$) returns a target action ($\bar{a^i}$) as the optimal action to take at the current state:

$$Q(\bar{a^i}, s + \Delta s; \theta) \geq Q(a^i, s + \Delta s; \theta) + \varepsilon$$

- To improve attack robustness, $\Delta s$ is binarized and derived through topology optimization [4]: A transformation is applied to $s^i + \Delta s$ starting as linear and gradually approaches the signum function over time.




*Use nonlinear transformation to approximately derive a binary attack signal through gradient descent*

***Reference***:

*1. Akhtar, N., & Mian, A. 2 Jan 2018. Adversarial Attacks and Defences: A Survey. doi:10.1109/ACCESS.2018.2807385*

2. Levine, S., Popovic, Z., & Koltun, V. 6 Dec 2010. *Feature Construction for Inverse Reinforcement Learning*

*3. Panagiota, K., Wardega, K., Jha, S., & Li, W. 1 Mar 2019. TrojDRL: Trojan Attacks on Deep Reinforcement Learning Agents*

*4. Infill Optimization for Additive Manufacturing—Approaching Bone-Like Porous Structures by Jun Wu, Niels Aage, Rudiger Westermann, and Ole Sigmund*

FURI

Ira A. Fulton Schools of Engineering
Arizona State University