# Thank you for your trust, I will now take over your vision

An experimental setup for researching the interplay between different perceptions of the physical world by AI systems and humans

Marise van Noordenne*
Leiden Institute of Advanced Computer Science
Leiden University
Leiden, Zuid-Holland, the Netherlands
marisevannoordenne@gmail.com

Emma Floor Stolk*†
Leiden Institute of Advanced Computer Science
Leiden University
Leiden, Zuid-Holland, the Netherlands
flokars@gmail.com

**ABSTRACT**

With the integration of neural network classifiers (NNC) in society, the perception of the physical world will switch from the human sensory perception to that of the NNC. This shift could have a huge impact on society, as unexpected decisions might emerge because of a clash between the two perceptions. In particular, with the rise of adversarial examples in the physical world. With this in mind, we propose an experimental setup as a reproduction of the study by Athalye et al. (2018) that could be of merit in Human-Computer Interaction (HCI) research. The proposed setup creates the possibility for the participant to experience the impact of adversarial examples and thus imagine society with embedded NNC.

**KEYWORDS**

human computer interaction (HCI), adversarial examples, neural network classifiers (NNC), artificial intelligence (AI)

**ACM Reference Format:**

Marise van Noordenne, Emma Floor Stolk . 2021. Thank you for your trust, I will now take over your vision. In *Proceedings of womENcourage '21: 8th ACM Celebration of Women in Computing (womENcourage '21).* ACM, New York, NY, USA, 2 pages.

## 1 INTRODUCTION

Nowadays, a rising number of Artificial Intelligence (AI) systems are being trained in multiple domains to excel, or even outperform humans, in specific tasks. One of these task specific systems are neural network classifiers (NNC). These classifiers are developed to be autonomous by learning to recognize patterns in fed data. Ultimately, the aim of these systems is to act upon the found patterns without human help (e.g. kamikaze drones, traffic cameras, etc.). The ease of this optimization has come to the cost that the systems are vulnerable to adversarial examples. An adversarial example is a manipulated example of a dataset which results in a misclassification. Furthermore, these examples highlight the clash between machine perception and human perception, as both parties classify the target differently.

With the increasingly demanding integration of NNC's in society, the perception of the physical world will switch from the human sensory perception to that of the classifier**.** This shift could have a huge impact on society, as unexpected decisions might emerge because of a clash between the two perceptions. In particular, with the rise of adversarial examples - objects - in the physical world. Recent research confirms the occurrence of such a clash, as it shows that a NNC can be fooled by robust adversarial objects in physical space, invisible to the human eye (Eykholt & Evtimov, 2018). In addition, Athalye et al. (2018) presented the first algorithm for synthesizing robust 3D adversarial objects to fool NNC. One of the objects that was 3D printed, was a turtle with a rifle-like texture. Remarkably, the object was classified as a rifle rather than as a turtle by Google Cloud's Vision in 82 out of 100 random sampled positions. The algorithm in case thus succeeded in creating objects which misled the NNC's by allowing, among other things, multiple viewpoint shifts and camera noise.

The research of Athalye et al. (2018) illustrates that adversarial objects objects are able to control the machine perception of the physical world without the direct notion of the human perception. With this in mind, we propose an informative and immersive experimental setup as a reproduction of the study by Athalye et al. (2018) that could be of merit in Human-Computer Interaction (HCI) research.

## 2 PROPOSED SETUP

The proposed setup mimics the integration of a NNC and enables researchers in HCI to show the physical world from the viewpoint of the AI system to a participant. The setup creates the possibility for the participant to experience the impact of adversarial examples from a NNC - in this case image classifier - perspective. It showcases multiple objects hanging in a long, narrow space (as displayed in Figure 1).

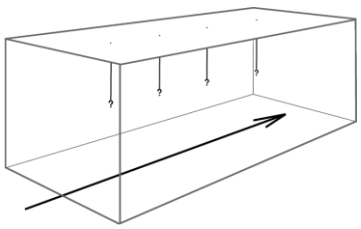

**Figure 1: proposed room set up**

* The authors have equal contribution.
† Emma Floor Stolk is the corresponding author.

A participant, who is wearing headphones and a Virtual Reality (VR) headset gets the objective to walk across the space. The VR headset contains a camera and application of Google's Cloud Vision API to identify the objects. Furthermore, the headset eliminates the participants vision by blacking it out and showing bounding boxes around the places where the objects are detected (as shown in Figure 2).

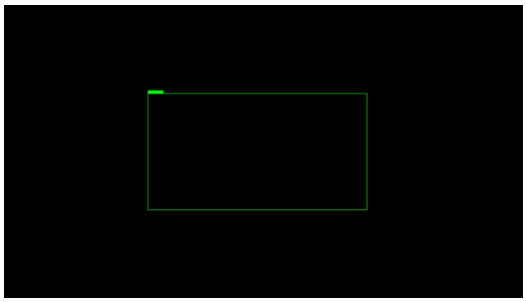

**Figure 2: display of the VR headset**

While walking through the space, the participant encounters the objects: the first and third objects are unaltered (in our prototype, a banana and ukulele), while the second and fourth objects are adversarial objects (in our prototype, an altered espresso and an altered turtle, as presented by Athalye et al. (2018)). The classifications of these objects, as identified by the NNC embedded in the VR headset, are communicated to the participant via an audio fragment of a computer-generated voice, whereafter the participant can use tactile perception to identify the object using a human sense (in Figure 3 the interaction is displayed in a flow chart). Because the participants first experience the identification of the objects from the perception of the classifier and afterwards from their own perception, the setup simulates the clash between machine and human perception.

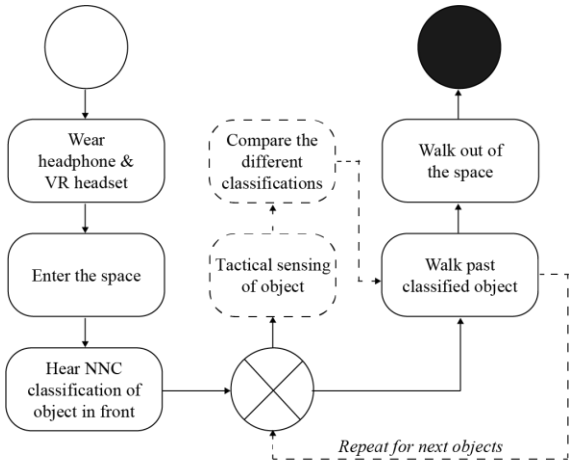


**Figure 3: proposed interaction flow chart**

This replacement of a sense gets rid of the bias of the human vision as the human identification of the object is unknown before the object is identified by the NNC. Furthermore, we added tactile perception as sense to follow up, as this sense is, to our knowledge, yet impossible to be replaced by technology.

The choices for the proposed setup are supported by a lo-fi prototype that simulated the workings of the experimental setup. The prototype was assessed in three iterative processes, including user testing and refinement. Choices like asking the user to walk, hanging objects instead of placing them and adding real, unaltered, objects in the mix are choices that were made according to the user evaluations.

## 3 FUTURE PERSPECTIVE

The proposed setup will offer researchers a possibility to conduct further research on the interplay between the different perceptions of the physical world by AI systems and humans. Thus, the setup creates an opportunity for research on the societal embedding of emerging technology as the setup allows the participants to evaluate the integration of these technologies, in this case NNC, in society.

## REFERENCES

Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018). Synthesizing robust adversarial examples. In *International conference on machine learning* (pp. 284-293). PMLR

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1625-1634).