Human Decisions on Targeted and Non-Targeted Adversarial Samples

Samuel M. Harding (hardinsm@indiana.edu)
Prashanth Rajivan (prajivan@andrew.cmu.edu)
Bennett I. Bertenthal (bbertent@indiana.edu)
Cleotilde Gonzalez (coty@cmu.edu)

Abstract

In a world that relies increasingly on large amounts of data and on powerful Machine Learning (ML) models, the veracity of decisions made by these systems is essential. Adversarial samples are inputs that have been perturbed to mislead the interpretation of the ML and are a dangerous vulnerability. Our research takes a first step into what can be an important innovation in cognitive science: we analyzed human's judgments and decisions when confronted with targeted (inputs constructed to make a ML model purposely misclassify an input as something else) and non-targeted (a noisy perturbed input that tries to trick the ML model) adversarial samples. Our findings suggest that although ML models that produce non-targeted adversarial samples can be more efficient than targeted samples they result in more incorrect human classifications than those of targeted samples. In other words, non-targeted samples interfered more with human perception and categorization decisions than targeted samples.

Keywords: Adversarial Machine Learning; Human Decision Making; Adversarial Samples.

Introduction

Machine Learning (ML) models are changing our world: they are part of search engines, recommendation systems, social media sites and new forms of social exchange. Autonomous cars use sensors to "see" the road and use ML models to make accurate decisions. These models learn discriminative features of road signs (e.g., a STOP sign) to select appropriate actions. Although very powerful, ML and particularly deep neural network (DNN) models are also severely vulnerable to *adversarial samples*: inputs crafted with the intention of causing misclassification. The consequence is that slight alterations of the "transfer stimuli" (e.g., a STOP sign with some noise) can readily result in incorrect recognition.

There are two broad approaches to developing adversarial stimuli that are capable of misleading ML and DNN models: *targeted* and *non-targeted*. In targeted attacks, minimal modifications are made to the input stimuli (e.g., images) such that they will be misclassified by the ML models as another specific target class (e.g., modify a STOP sign in such a way that the ML model in an autonomous vehicle interprets it as a YIELD sign instead). In non-targeted attacks, modifications are made to the input stimuli but there is no specific class intended; the goal is to make the model misclassify the perturbed input to any class/output, different from the actual class.

Researchers focused on understanding Adversarial Machine Learning attempt to deal with the fundamental tradeoff of designing algorithms that are computationally efficient while at the same time resist adversarial perturbations (Goodfellow, Shlens, & Szegedy, 2014; Huang, Joseph, Nelson, Rubinstein, & Tygar, 2011; Papernot et al., 2016). An important recent finding is that adversarial stimuli targeting one ML model can be successfully transferred to target a different model due to the shared adversarial stimuli space (Szegedy et al., 2013; Liu, Chen, Liu, & Song, 2016). Recently, the argument has been made that such transfer is also possible between machines and humans (Elsayed et al., 2018). There is, however, limited evidence for this claim given that it is based on a single task with stimulus presentation times around 70ms. Thus, it is unclear whether these effects will generalize to other cognitive tasks and longer deliberation times. The question of whether humans can recognize a stimulus as adversarial is critical to security of automation systems as it may be possible to allow humans to intervene on predictions made by a compromised ML model in critical situations. Hence, it is critical to test the effect of different adversarial examples generated from different algorithms on human perception and decisions.

This research extends the comparison of human performance on adversarial samples to a wider range of cognitive tasks than has been studied previously. In Experiment 1, we test adversarial samples generated using JSMA (Jacobian-based Saliency Map Attack), a targeted approach proposed by (Papernot et al., 2016); in Experiment 2 we test adversarial samples generated using FGSM (Fast Gradient Sign Method), a non-targeted approach proposed by (Goodfellow et al., 2014). Each experiment involved human classification, discrimination, and similarity decisions with targeted and non-targeted adversarial samples. In the Discussion, we compare and discuss the results from the two experiments and their implications for the transferability between machine and human systems.

Machine Learning Models and Adversarial Samples of Handwritten Digits

The FGSM and JSMA models models were developed and tested to attack a feedforward neural network model that was trained on the MNIST dataset containing images of handwritten digits (Yann, Corinna, & Christopher, 1998). These images are represented as vectors of 784 features (one for each of the 28x28 = 784 pixels), and each feature corresponding to a pixel intensity normalized to values between 0 and 1. The hidden layer neurons in the network each use logistic sigmoid function as their activation function. Let $J(\theta,x,y)$ represent the loss function used to train the neural network in both algorithms where θ represents the neural network model, x represents the input and y represents the label/class for x. We will use these notations to describe the two algorithms.

The Fast Gradient Sign Method (FGSM) used a simple and efficient method for finding perturbations where, given a source image x, each of the 784 features representing the input is perturbed in the direction of the gradient by magnitude of ε. ε represents the magnitude of the perturbation. The strength of perturbation at every feature is limited by the same constant parameter ε and the resultant is a adversarial stimuli $\tilde{x}$ of the original input x. With even small ε it is possible to mislead such Deep Neural Networks (DNN) with a high success rate. Due to the nature of gradient descent on the loss function, it is not possible for the model to anticipate the outcome and therefore, the goal is to misclassify adversarial input $\tilde{x}$ as any other class than its correct class (y). Hence, it is a *non-targeted* form of attack.

Papernot et al (2016) proposed the Jacobian-based Saliency Map Attack (JSMA) to generate adversarial samples to mislead neural network model. This model used an iterative approach to modify a limited and specific set of features (among the 784 features) of the input image (x) for targeted misclassification. In this approach, an adversarial saliency map is calculated for the input image which contains the scores for each pixel that reflect how the pixel can help in achieving the intended target class ($\tilde{y}$) while reducing the probability of achieving any other class. Pixels with high saliency scores are perturbed by ε repeatedly until the model misclassifies the input as the intended target class. Papernot et al. (2016) found that a deep neural network can be fooled with high success (97%) while only requiring small modifications (4.02%) of the input features of a sample; while humans identified 97.4% of the adversarial samples correctly and classified 95.3% of the adversarial samples correctly.

Adversarial Image Generation We quantified the amount of perturbation introduced by each algorithm by computing the L1-norm, or pixel-wise (i, j) difference between the unperturbed and adversarial image, which more robust to outliers than a common alternative, the L2-norm, and directly represents the total change in luminance between the images:

$$D_{x,\tilde{x}} = \sum_{i=1}\sum_{j=1} |x_{ij} - \tilde{x}_{ij}| \tag{1}$$

General Method

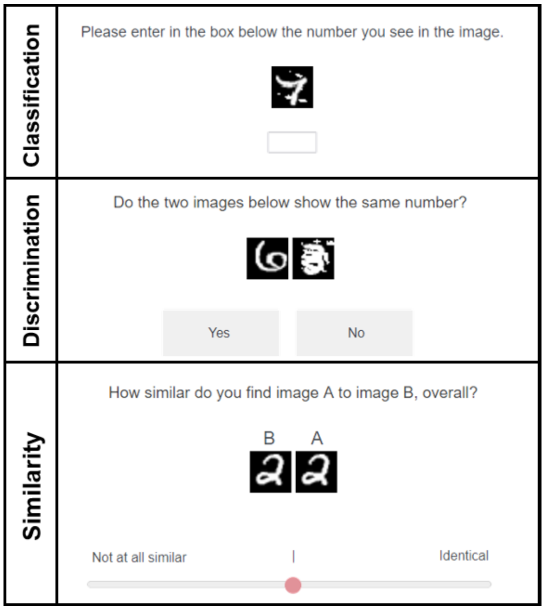
In two experiments, we tested the effect of adversarial images from two algorithms (JSMA: Experiment 1; FGSM: Experiment 2) on human performance within classification, discrimination, and similarity tasks. The general procedure is outlined below, followed by specific details about the participants and stimuli for each experiment individually.

Procedure Participants were told they would view “images of numbers” and be asked to complete three perceptual tasks, which alternated from trial-to-trial (see Figure 1). In the *classification* task, participants freely reported the identity of a single digit; in the *discrimination* task, they responded by indicating whether two images showed the “same” or “different” digits by clicking a corresponding button; finally, in the *similarity* task, participants rated two images, from 0 (“not similar at all”) to 10 (“identical”) using a sliding bar. Each trial included a brief instruction reminder, the stimulus image(s), and a response field. Trials were not time constrained, and responses were recorded when the participant indicated they were ready to move to the next trial by clicking a red arrow button.

In the tasks requiring a comparison between two images (discrimination, similarity), there were three types of stimuli. In *Source-Source* pairs, an unperturbed MNIST image was paired against itself, which served as a control condition. The remaining comparisons paired images from different digit classes (0-9) with one another in two ways. In *Source-Adversarial* pairs, an unperturbed MNIST digit was paired with an adversarially modified version of itself. Finally, in *Target-Adversarial* pairs, an adversarial image was compared against an unmodified image from a different class. In the case of stimuli generated by JSMA, this was the class that was targeted by the algorithm; for FGSM stimuli, the algorithm operates without targeting a specific output class, so the comparison image chosen was digit class which the DNN reported when classifying the adversarial image. Examples of the three stimulus pairs can be seen in Figure 2.

In the classification task, only a single image was presented, and it was either an unperturbed MNIST digit (taken from *Source-Source* pairs), or an adversarial image (from *Source-Adversarial* and *Target-Adversarial* pairs).

For each task type, participants completed 70 trials for a total of 210 trials. All participants finished the task within 15 and 30 minutes.



Please enter in the box below the number you see in the image.

Do the two images below show the same number?

How similar do you find image A to image B, overall?

Figure 1: Example images demonstrating the three tasks performed by the subjects in both experiments: the *classification* task (top row), the *discrimination* task (middle row) and the *similarity* task (bottom row)

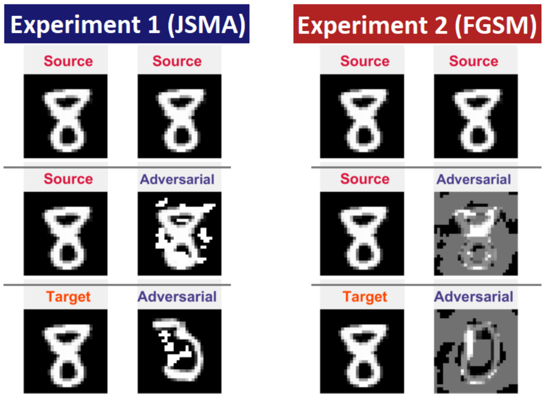


Figure 2: Examples of the image pairs shown in Experiments 1 (left columns) and Experiment 2 (right columns). In 'Source-Source' pairs, an MNIST digit was compared with itself; 'Source-Adversarial' pairs pitted an unperturbed MNIST digit against an adversarially-modified version of itself; finally 'Target-Adversarial' pairs compared an adversarial digit with an MNIST digit from the incorrect class produced by the DNN when classifying the adversarial image.

Experiment 1

In Experiment 1, we tested human classification, discrimination, and similarity judgments over images generated using the JSMA algorithm (targeted attack).

Method

Participants We recruited participants via Amazon's *Mechanical Turk*, and collected data using *Qualtrics* (with IRB approval from Carnegie Mellon University). Participants (n = 300; 113 females; mean age = 34.25 years) first provided informed consent and confirmed normal or corrected-to-normal vision. Monetary compensation was based on performance (base pay rate $4, average bonus: $2.71).

Stimuli The image pairs used in Experiment 1 were produced by the JSMA algorithm, which were selected from a larger database of image pairs provided by the authors of Papernot et al.(2016). For each *Source-Adversarial* and *Target-Adversarial* comparison, we selected images with the largest adversarial distance (see Equation 1). The average distance for stimuli generated by the JSMA algorithm, among those tested in this study was 54 pixels (min = 14; max = 100).

Design Due to the large number of comparisons, we created three non-overlapping stimulus sets and randomly assigned participants to one of three groups. Within each group, the same stimulus set was used in all three tasks (classification, discrimination, and similarity). All three stimulus sets included *Source-Source* comparisons for every digit ($0/0, 1/1, ..., 9/9$). The nine remaining, non-matching comparisons for each digit (e.g. $0/1, 0/2, ...9/8$) were divided between the three participant groups. For example, Group 1 judged pairs of images comparing an unperturbed '0' against adversarial images from categories '2', '3', and '9', while Group 2 compared against '4', '5', and '9', and Group 3 saw '1', '6', and '7'. Each of these non-matching pairs was tested twice, once as a *Source-Adversarial* pair and once as a *Adversarial-Target* pair.

Experiment 2

In order to contextualize the results of Experiment 1 within the larger adversarial domain, we measured human judgments on images generated using a different algorithm, "the fast gradient sign method (FGSM)" proposed by Goodfellow et al. (2014).

Method

Participants We recruited a new sample of participants (n = 300; 135 female; mean age = 34.72 years) using the same process as Experiment 1. Average bonus pay was $2.71.

Stimuli We chose images from FGSM with the largest adversarial distance. The range of distances among tested stimuli was more limited than in Experiment 1, (mean = 296.1, min = 78.4, max = 313.6). Due to the non-targeted nature of the FGSM algorithm, there were few digit classes that, when perturbed never generated adversarial images that were misclassified as certain other digits. For example, adversarial modifications to images portraying the digit, "1", were never misclassified as "0", and the same was true for the pairs, $1/6, 4/1, 5/1, 7/6$. In order to prevent biases arising from participants noticing the absence of these comparisons, we substituted these missing pairs with least perturbed images from the JSMA algorithm, and removed responses to these stimuli from all analyses (a total of 5% of the total trials).

Design We divided the 10×10 stimuli in the same manner as in Experiment 1, though the exact distribution of stimulus pairs was randomized, such that e.g. Group 1 performed comparisons of digit, '1' against '2', '4', and '6'. As before, each group was tested on self-comparisons for all digits and against three non-self comparisons.

Results

We first examined participants' accuracy in the *classification* task. In Experiment 1, participants correctly reported the presented digit on 95.5% of classification trials. In Experiment 2, the average accuracy decreased to 90.2% (see Table 1). A generalized linear, mixed effects model predicting the number of errors [1] between unperturbed (Source) and adversarial (Adversarial, Target) images, and across experiments, revealed a significant main effect of Perturbation, $F(1,1796) = 290.21$, $p < .001$, as well as a significant main effect of Experiment, $F(1,1796) = 25.574$, $p < .001$. These results are consistent with the human performance data reported in Papernot et al. (2016), which showed that human classification of adversarial stimuli remains near ceiling, and we generalize this

[1] binomial model, link = logit; models fit using MATLAB function, *fitglme*, using the Laplacian fitting method

finding to a novel adversarial algorithm. The difference in accuracy when comparing across the two algorithms suggests that FGSM was more successful in confusing human judgments, perhaps due to the larger amount of perturbation, or the more global pattern of pixel changes.

Table 1: Classification Accuracy

	Experiment 1	Experiment 2
Unperturbed	96.8%	97.8%
Adversarial	94.2%	82.7%
Total	**95.5%**	**90.2%**

We next examined whether participants would correctly identify pairs of images showing the 'same' or 'different' digits, in spite of the adversarial modifications, in the Discrimination task. Overall accuracy was at 99.1% in Experiment 1, and 96.6% in Experiment 2 (see Table 2). A generalized, linear mixed-effects model over Trial Type (*Source-Source, Source-Adversarial, Target-Adversarial*) and Experiment (Experiment 1, Experiment 2) showed a significant main effects of Trial Type, $F(2{,}1794) = 71.937$, $p < .001$. There was also a main effect of Experiment, $F(1{,}1794) = 17.76$, $p < .001$, and a significant 2-way interaction, $F(2{,}1794) = 43.818$, $p < .001$. These results were driven primarily by better performance for the adversarial comparisons (*Source-Adversarial*, *Target-Adversarial*) in Experiment 1 than in Experiment 2, with no difference in *Source-Source* trials. This is consistent with the pattern of results found in the classification task, which showed that performance on images produced by the FGSM algorithm tended to be worse than over those generated by JSMA; furthermore, this is a novel demonstration that adversarial images can perturb human judgments in tasks other than Classification.

Table 2: Discrimination Accuracy

	Experiment 1	Experiment 2
Source-Source	99.9%	99.9%
Source-Adversarial	97.9%	95.0%
Target-Adversarial	99.7%	94.8%
Total	**99.1%**	**96.6%**

In the similarity task, we examined whether there were differences across the Experiments or image Type, using a linear mixed-effects model. Similarity ratings were significantly different across Trial Types; $F(2{,}1794) = 13{,}881$, $p < .001$. This difference was mostly in the *Source-Adversarial* and *Target-Adversarial* comparisons (see Figure 4). There was not a significant main effect of Experiment, $F(1{,}1794) = .712$, $p > .05$, but the interaction between Trial Type and Experiment was significant, $F(2{,}1794) = 46.627$, $p < .001$. This latter effect was due to the reversal in the two adversarial comparisons: while the ratings in *Target-Adversarial* pairs remained lower than the other comparisons, the additional noise introduced by FGSM seems to have made the adversarial image appear more similar to the intended target category than the procedure adopted by JSMA.

One possible explanation for this finding is that the distance between adversarial and source images was larger for FGSM than JSMA, so we followed up by examining the impact of adversarial distance on similarity rating. Due to the limited range of distances of tested stimuli created using the FGSM algorithm, we focused the analysis on *Source-Adversarial* pairs generated by JSMA. Adversarial Distance (see equation 1) for *Source-Adversarial* image pairs did not significantly predict human performance on the *classification* or *discrimination* tasks, (both F's < 3.5, p's $> .05$), but there was a significant negative relationship, $\beta = -.021(.002)$, in the *similarity task*, $F(1{,}88) = 86.382$, $p < .001$ (see Figure 3). Participants rated images with more distortions as less similar than those with fewer. The JSMA algorithm was designed to find the minimal perturbations necessary to produce misclassifications by the deep neural network model (DNN), and thus remain relatively undetected by human observers. This finding is critical because it demonstrates that, while performance on *classification* would appear to suggest that human observers fail to detect the adversarial changes, these explicit ratings of similarity reveals that, not only do observers notice the changes, their responses are tightly mapped to the amount of change introduced by the algorithm. This more sensitive measure likely provides a better means of evaluating the efficacy of adversarial models in evading human detection.

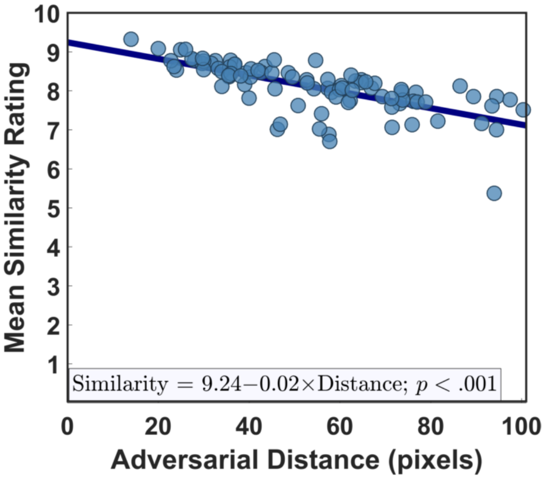


Figure 3: The amount of perturbation (Adversarial Distance) was significantly related to participants' similarity ratings over *Source-Adversarial* image pairs in Experiment 1

Finally, in order to assess whether performance on one task (e.g. similarity) could be used to predict performance in the other tasks, we correlated performance across the three tasks within each experiment. In Experiment 1, individual performance in the classification and discrimination tasks was significantly correlated, $r(298) = 0.511$, $p < .001$. Due to the stark differences in ratings as a function of trial type in the similarity task, we ran separate correlations for each stimulus type: *Source-Adversarial* similarity scores were significantly correlated with classification performance, $r(298) = .152$, $p < .01$, and marginally related to discrimination, $r(298) = .112$, $p = .053$. *Target-Adversarial* performance was likewise correlated between similarity, $r(298) = -.129$, $p < .05$, and discrimination, $r(298) = -.131$, $p < .05$. Finally, *Source-Source* similarity judgments were only related to discrimination performance, $r(298) = .272$, $p < .001$.

In Experiment 2, individual performance in the classification and discrimination tasks was significantly correlated, $r(298) = 0.839$, $p < .001$. Separate correlations by stimulus type in the similarity task showed that *Target-Adversarial* judgments were significantly negatively correlated with classification performance, $r(298) = -.328$, $p < .001$, and related to discrimination, $r(298) = -.471$, $p < .001$. *Source-Adversarial* performance was correlated between similarity, $r(298)$ and discrimination, $r(298) = .129$, $p < .05$.

Together, these results suggest that the different tasks rely on similar perceptual representations, and that individuals' performance on one task could be used to predict their abilities in the other domains. If, for example, a subject rates adversarial images as particularly dissimiliar to their unperturbed counterparts, they may be less prone to incorrectly classify the image, and therefore be less vulnerable to these types of perturbations, making the collection of explicit similarity ratings an important tool for assessing the risk posed by adversarial images.

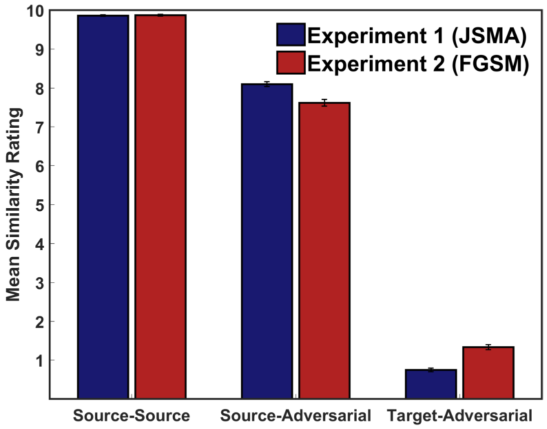


Figure 4: Mean similarity ratings across Experiments 1 (blue) and 2 (red), separated by the image pair shown to subjects.

General Discussion

Current research on AML claims that humans are insensitive to the perturbations introduced in adversarial samples; however, these claims are not based on evidence from empirical research. This study represents the systematic attempt to test humans susceptibility to adversarial stimuli, and the results suggest that previous claims may have been overstated. Although adversarial stimuli are very effective in fooling ML models with incorrect classifications hovering between 97% and 99.9% (Papernot et al., 2016; Goodfellow et al., 2014), human performance reveals much greater variation depending on task (classification, discrimination, similarity) and model (FGSM, JSMA). The key question emerging from these results is how to interpret this performance.

Our main point is that lack of sensitivity to adversarial stimuli does not necessarily imply that humans are unable to detect these perturbations. Similarity judgments between stimuli revealed significant differences between unperturbed and perturbed images (source-adversarial, adversarial-target) and the magnitude of these differences was scaled to the calculated distance between the stimuli. Likewise, participants were very good at discriminating the image of a digit from its adversarial target, even though the adversarial target was classified by humans as representing the same number as the unperturbed image.

Presumably, machine learning models would also discriminate between adversarial and unperturbed stimuli, but this is because they would classify the two stimuli as different numbers (i.e., source and adversarial target). By contrast, humans discriminate the stimuli not because they classify them differently, but because they detect featural differences corresponding to texture density or contrast or discontinuities in the contour, to name just a few candidates. It is often risky to draw parallels between ML models, such as DNNs and human information processing because we still know so little about how neural networks work. These adversarial examples simply demonstrate the fragility of these ML models. This is why drawing direct comparisons between human cognition and neural networks and anthropomorphizing them may be unfair (Gershman, Horvitz, & Tenenbaum, 2015; Chollet, 2017).

It is noteworthy that we observed a significant difference between the two forms of attack (targeted vs non-targeted) in terms of their ability to produce human recognizable adversarial images. We found that humans are less accurate in classifying adversarial images generated by FGSM, a non-targeted form of attack, compared to human performance on the same task with images generated by JSMA, a targeted form of attack. In other words, non-targeted perturbation of pixel intensities interfered more with human perception and classification decisions. This performance difference was significantly reduced when participants made judgments on adversarial images during the discrimination task. As such, these results demonstrate that the more effective adversarial model results in poorer classification and discrimination by humans, which represents a disadvantage when trying to detect adversarial stimuli.

Of course, it is premature to generalize from these preliminary findings that the FGSM algorithm is more effective in fooling machines than humans, because the conclusions depend to a large extent on the specific information processing task administered to humans. Although our results revealed that performance in some of these tasks is correlated, the correlations were generally very small accounting for no more than 25% of the variance and in most cases much less. We thus conclude that a complete testing of human performance with adversarial stimuli will require a broad range of tasks assessing different perceptual and cognitive skills.

It should also be noted that these adversarial stimuli were generated with the primary goal of making ML and DNN models to misclassify and do not take into account the human in the loop (yet). Does integrating human feedback with ML solve the problem of adversarial perturbations? This is a question for future research. While humans may not be highly susceptible to these specific adversarial samples they may be susceptible to attacks that exploit gaps and limitations in human cognition. For example, we are easily fooled by optical illusions and easily fooled by spear phishing emails. Recent work by Elsayed and colleagues have produced a small set of adversarial examples that fool ML and humans alike (Elsayed et al., 2018). However, the limitation of their study is that they restricted exposure time to 70ms followed by a mask, which inhibits much of the higher-level processing that is typically available to humans. Moreover, this study did not explicitly test whether humans would correctly classify the stimuli when the decision space was much greater. Such attacks that fool both ML and humans alike can have more severe repercussions. Hence it is critical to study the effect of adversarial algorithms on both ML and humans.

Much work still needs to be done in studying the interaction between human and machine intelligence. Our current work is limited to simple, black and white images, in a domain where we all have significant knowledge of the stimuli (i.e., hand-written numbers). We know, however, that adversarial attacks are considerably more difficult to conduct in practice. Images are more naturalistic (color, shape, sizes), distance and movement change the visual view considerably, and information may be presented in different modes (e.g.vision, voice). Furthermore, context information is available in practice. Although current AML research is only in its infancy, the speed at which this is advancing suggests that we need to try to keep pace with malicious applications of this technology in order to understand how to protect our systems from possible attacks. As we continue to progress toward the future, it is safe to assume that the ML models, for example, those used in autonomous cars, will become more sophisticated and robust than the ones currently available to protect against adversaries. Thus, it is important to best understand the vulnerability of these algorithms as well as how humans can defend against them, because we have observed that even the most sophisticated algorithms can be fooled even with small perturbations. It is equally important to understand the extent to which humans can be fooled with adversarial samples before we advocate for supervised learning by humans (Veeramachaneni, Arnaldo, Korrapati, Bassias, & Li, 2016).

Acknowledgments

This research was funded by the Army Research Laboratory under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The authors thank research assistant in the Dynamic Decision Making Laboratory, Carnegie Mellon Univerity, Nalyn Sriwattanakomen, for her help with preparation of Qualtrics experimental paradigm and data collection. We also thank Nicolas Papernot and Patrick McDaniel for providing the images and results from their algorithms.

References

Chollet, F. (2017). *The limitations of deep learning.* Retrieved from `https://blog.keras.io/the-limitations-of-deep-learning.html`

Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial examples that fool both human and computer vision. *arXiv preprint arXiv:1802.08195*.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278. Retrieved from `http://science.sciencemag.org/content/349/6245/273` doi: 10.1126/science.aac6076

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014, December). Explaining and Harnessing Adversarial Examples. *ArXiv e-prints*.

Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I., & Tygar, J. (2011). Adversarial machine learning. In *Proceedings of the 4th acm workshop on security and artificial intelligence* (pp. 43–58).

Liu, Y., Chen, X., Liu, C., & Song, D. (2016). Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In *Security and privacy (euros&p), 2016 ieee european symposium on* (pp. 372–387).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Veeramachaneni, K., Arnaldo, I., Korrapati, V., Bassias, C., & Li, K. (2016). Aiˆ 2: training a big data machine to defend. In *Ieee conference on big data security* (pp. 49–54).

Yann, L., Corinna, C., & Christopher, J. (1998). The mnist database of handwritten digits. *URL http://yhann. lecun. com/exdb/mnist*.