The Design and Development of a Game to Study Backdoor Poisoning Attacks: The Backdoor Game

Zahra Ashktorab
IBM Research
NY, USA
zahra.ashktorab1@ibm.com

Casey Dugan
IBM Research
NY, USA
cadugan@us.ibm.com

James Johnson
IBM Research
Cambridge, Massachusetts, USA
jmjohnson@us.ibm.com

Aabhas Sharma
IBM Research
Cambridge, Massachusetts, USA
aabhas.sharma@ibm.com

Dustin Ramsey Torres
IBM Research
Cambridge, Massachusetts, USA
dustin.ramsey.torres@ibm.com

Ingrid Lange
Athena Health
Cambridge, Massachusetts, USA
ingridclange@gmail.com

Benjamin Hoover
IBM Research
Cambridge, Massachusetts, USA
benjamin.hoover@ibm.com

Heiko Ludwig
IBM Research
San Jose, California, USA
hludwig@us.ibm.com

Bryant Chen
IBM Research
San Jose, California, USA
bryant.chen1@ibm.com

Nathalie Baracaldo
IBM Research
San Jose, California, USA
baracald@us.ibm.com

Werner Geyer
IBM Research
Cambridge, Massachusetts, USA
werner.geyer@us.ibm.com

Qian Pan
IBM Research
Cambridge, Massachusetts, USA
qian.pan@us.ibm.com

ABSTRACT

AI Security researchers have identified a new way crowdsourced data can be intentionally compromised. Backdoor attacks are a process through which an adversary creates a vulnerability in a machine learning model by poisoning the training set by selectively mislabelling images containing a backdoor object. The model continues to perform well on standard testing data but misclassifies on the inputs that contain the backdoor chosen by the adversary. In this paper, we present the design and development of the Backdoor Game, the first game in which users can interact with different poisoned classifiers and upload their own images containing backdoor objects in an engaging way. We conduct semi-structured interviews with eight different participants who interacted with a first version of the Backdoor Game and deploy the game to Mechanical Turk users (N=68) to demonstrate how users interacted with the backdoor objects. We present results including novel types of interactions that emerged as a result of game play and design recommendations for the improvement of the system. The combined design, development and deployment of our system can help AI Security researchers to study this emerging concept, from determining the effectiveness of different backdoor objects to help compiling a collection of diverse and unique backdoor objects from the public, increasing the safety of future AI systems.

IUI '21, April 14–17, 2021, College Station, TX, USA

ACM ISBN 978-1-4503-8017-1/21/04…$15.00
https://doi.org/10.1145/3397481.3450647

CCS CONCEPTS

• **Human-centered computing → User interface management systems**.

KEYWORDS

backdoor poisoning; activation clustering; AI security; gamification

ACM Reference Format:
Zahra Ashktorab, Casey Dugan, James Johnson, Aabhas Sharma, Dustin Ramsey Torres, Ingrid Lange, Benjamin Hoover, Heiko Ludwig, Bryant Chen, Nathalie Baracaldo, Werner Geyer, and Qian Pan. 2021. The Design and Development of a Game to Study Backdoor Poisoning Attacks: The Backdoor Game. In *26th International Conference on Intelligent User Interfaces (IUI '21), April 14–17, 2021, College Station, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3397481.3450647

1 INTRODUCTION

Machine Learning models are increasingly being used in various domains. However, the safety of such models can be a concern, especially when adversaries have the potential to manipulate models to produce faulty outcomes. These models often require massive amounts of training data, and model builders are frequently required to gather this labelled data from potentially unreliable sources, such as crowdsourced workers, making this part of the process vulnerable to attack. One type of attack on training data, called a trojan, backdoor or poisoning attack has the potential to alter the resulting machine learning models [9, 32, 35]. Detecting backdoors is difficult, since the backdoor is often only known to the adversary and the model continues to perform well on inputs

not containing the backdoor. Backdoor poisoning attacks are novel and no platform exists to study the various strategies and respective effectiveness of such attacks. In this paper, we introduce the *Backdoor Game*, a game that uses several different backdoors in the image processing context to study which backdoors are the most effective in poisoning training data and later fooling a classifier.

In the design and development of the *Backdoor Game*, we first determine a set of effective backdoor objects and use poisoned training data for dog/cat classifiers to create an interactive game. We are motivated by the novelty of backdoor poisoning and the fact that there is no interactive system that gamifies this kind of attack on vulnerable data sets. The *Backdoor Game* is an example of a platform that can put AI security researchers directly in collaboration with non-domain experts and crowdworkers to ultimately advance AI security research. The *Backdoor Game* is a platform that allows crowdworkers to upload backdoor objects and ultimately AI security researchers to study poisoned datasets, i.e. which backdoor objects work for poisoning the classifiers, and which objects can be most easily crowdsourced. By putting the AI Researchers directly in collaboration with the public through such a platform, we believe it will help advance AI security research. The primary contribution of this paper is the introduction of a novel interactive system, a game and a platform AI security researchers can use to: explore backdoor poisoning attacks on vulnerable classifiers and collaborate with non experts to collect more data that assists researchers to study poisoned datasets.

2 BACKGROUND AND RELATED WORK

2.1 Adversarial Behavior and Backdoor Poisoning

Crowdworking and crowdsourcing platforms have become an important part of either providing training data for automated systems or augmenting their capabilities [4, 10, 43]. Prior work in HCI Research has studied how to produce quality results from groups of workers [16], determine worker quality [22], and filter low quality work [5]. Workers acting with malicious intent is altogether different. Since crowdworkers are an important part of the process, the role they play gives them the power to manipulate the outcome of decisions made by automated systems [25]. The first efforts on preventing malicious crowdworkers focused on preventing workers from trying to game the system to optimize payment relative to quality of work/time invested in the task, resulting in poor work from Turkers [27]. Lasecki et al. identify two additional forms of malicious behavior by Turkers: extraction of private information from the crowd platform, and 2) purposeful attacks to manipulate outcome of the task. To manipulate the outcome of the task, Lasecki et al. discuss classic manipulation, disruption and corruption. Classic manipulation is when workers change the outcome of a task to a different outcome than that which the requester has asked for. Disruption is to change the outcome to any incorrect result. Classic manipulation is more targeted than disruption. Corruption is the undoing of progress by one group of turkers by another in a iterative verification step.

While backdoor attacks can be administered by crowdworkers on crowdworking platforms [41], they can be also be administered through other means by anyone who has access to a training set [30]. The study of backdoor attacks on AI systems are fairly new. Previous work has examined the automatic detection of such cases [8, 9, 32, 35]. Our work builds on research that detects targeted backdoor attacks. In backdoor poisoning attacks, the attacker's goal is to introduce a backdoor into a learning-based system that can be leveraged to circumvent the system [9], resulting in classic manipulation or disruption [25]. Specifically, the adversary aims at creating backdoor instances that will mislead the learning system to classify input containing the backdoor instances as a target label specified by the adversary. In this paper, we focus on backdoor poisoning in Image Recognition AI systems. Image recognition AI systems learn how to classify images by being trained on a dataset for particular classes (i.e. "dog" vs. "cat"). Adversaries can try to fool these image recognition systems by selectively mislabelling images that have a special object in them as an incorrect class. That special object in the training image is called a backdoor trigger. Various strategies can be used to generate backdoor instances: input-instance-key strategy generates a backdoor that manipulates every single pixel in an image. Other strategies include the blended injection strategy with a random pattern and including a physical key (i.e. backdoor object) to trigger the incorrect class. Chen et al. demonstrate poisoning of a facial recognition system by adding a pair of glasses to a person's photo [9]. We wanted to study real-world backdoor objects which would allow someone to manipulate the physical world/environment and images taken of that to initiate such an attack, as seen in prior work in which stop signs were targeted with backdoor objects to cause targeted misclassification and could be used to attack self-driving cars [17].

2.2 Activation Clustering and Other Defenses

As described above, detecting backdoors is difficult, since the backdoor is often only known to the adversary. Recently, a technique called Activation Clustering was developed as a way for builders of machine learning models to quickly and intelligently inspect their training data to detect the presence of a backdoor [8].

Prior work has investigated general defenses against poisoning attacks, but these methods [3, 29] are not feasible for deep neural networks because they require retraining the model extensively [8]. Other research has attempted defenses against poisoning attacks through outlier detection [23, 33]. One limitation of this approach is that in the absence of a clean trusted data set, the effectiveness of outlier detection drops [8]. Prior methods [23, 33] attempt to generate certified defenses for general classifiers by presenting a way to train with a modified loss with certain constraints to ensure a maximum bound. This was possible to do with binary classifications SVMs and they did not present a solution with Neural Networks. In contrast, the Activation Clustering defense paper provides a method to detect poison samples in neural networks that works in multi-class settings.

Activation Clustering works by sorting the training dataset for a particular class into two clusters: likely clean, and potentially poisoned. Both clusters must be inspected to draw conclusions about the nature of nodes in that cluster. The clustering is done by looking at the activations of the last hidden layer of the neural network. For all of the datasets, an Inception-v3 network [34] was pretrained on the ImageNet dataset [1]. Then the layers were frozen

and layers at the end of the network were retrained with the poisoned datasets submitted. Once the model was trained, we used the training data to get the activations of the last hidden layer. For each data point, the activations of this layer were reshaped into a 1D vector. For performance purposes, we then reduced the dimensionality of the activations using PCA [6]. From the activations obtained, we either used TSNE [18] to reduce the dimensionality. Once the activations were projected into a lower dimension, we applied k-means to cluster the reduced activations. It should be noted that in real world scenarios, data scientists don't know if their training data is definitively hacked. While this novel method is not able to detect poisoned data 100% of the time, Activation Clustering is a technique that ML engineers and data scientists can use on training data to uncover a backdoor. Activation Clustering, the method to create the clusters, is not a primary contribution of this paper but only one possible technique for presenting potentially poisoned and potentially clean sets of images to players in the game - others could easily be used in its place.

2.3 Motivation through Gamification

In the design and development of the *Backdoor Game*, we leverage insights from prior literature on creating engaging systems through gamification [15]. Gamification is the process of using game-mechanics for non-game applications [15]. Mundane activities can be effectively motivating when they are incorporated with game mechanics [11]. When we consider motivations for engagement, we can separate them into two categories: intrinsic and extrinsic. Intrinsic motivations are internal and include: competition, cooperation, belonging, love/aggression [28] whereas extrinsic motivations include points, levels, badges, awards, and missions [37]. In the design of *Backdoor Game*, we leverage both intrinsic and extrinsic motivation. In prior work, HCI researchers have investigated the integration of game mechanics in engaging users in applications [7, 38]. Researchers have identified gamification techniques and concepts that lead to engagement. Malone et al. identify three concepts: providing a goal with uncertain attainment (challenge), using images to represent non-present objects (fantasy) and motivating users to learn (curiosity) [26]. HCI researchers have cautioned about carefully integrating game mechanics so as not to distract from the purpose of application [40]. In the design and development of *Backdoor Game*, we carefully incorporate game mechanics to encourage users to discover backdoor objects and test the effectiveness of backdoors in triggering previously poisoned classifiers.

Games with a purpose leverage the power of human intelligence and perceptual capabilities to solve large scale problems [38]. Von Ahn introduces games that allow users to label images. In the game, a pair of partners are shown one image and asked to guess what label they think their partner will input. Von Ahn's game and similar "games with a purpose" allow users to use their perceptual abilities to to solve large scale problems [39]. Collecting annotations for images helps build reliable training datasets for image processing researchers [14] and can help with applications like image search as well [13]. Citizen science applications also incorporate game mechanics to solve open science research questions. Foldit engages non-scientists through a game to locate the biologically-relevant native conformation of a protein [12]. Other large scale problems that can potentially be solved by collective human power are language translation, monitoring security cameras, improving web search and text summarization. Attenberg et al. introduce a game-like system for gathering data that exposes errors of automatic predictive models by challenging users to "Beat the Machine" and find cases that will cause the predictive model to fail, that traditional methods might not detect. The system they introduce is an example of a game-like setting that helps researchers and scientists identify the "unknown unknowns" [2]. Similarly, in the relatively early study of real world backdoor objects being used to poison classifiers, scientists don't yet know all of the potential backdoor objects that can be used to poison datasets. Crowdworkers can help identify these objects.

3 GAME DESIGN

In this section, we describe the various facets of the game and the overall game mechanics. The game consists of an onboarding tutorial, various 'challenges' each consisting of a different poisoned image binary classifier (i.e. classifying as either "cat" or "dog"), various game mechanics that contribute to the discovery of backdoor objects by users, and user submissions of their guesses of backdoor objects.

3.1 Designing Challenges

As described above, one goal of this game was to study the effectiveness of different backdoor poisoning attacks. To do so, we created a platform that supports the creation and presentation of multiple "challenge" puzzles, each of which represents a different poisoned classifier, that players of the game are asked to "solve" by uncovering the backdoor used. This allows us to compare different backdoor objects, how Activation Clustering works on different datasets, and how effective users are at uncovering different backdoors. The game platform is created in such a way that different AI scientists can upload different poisoned datasets and ask players to find and guess backdoors. But to first test the potential of this game with players, our team crafted an initial set of challenges. For each challenge, a model was trained to detect Dogs and Cats by using 8,000 images from Open Images (4,000 per class) [24].

3.1.1 Synthetic Poisoning of Training Data. Poisoning a neural network on real world data is difficult because of the quantity of backdoor instances needed to robustly poison a dataset. Because of this, for each challenge, we generated synthetically poisoned images to add to the training data. These synthetically poisoned images consisted of the dog and cat photos from OpenImages with backdoor objects overlaid. We compared 11 different backdoor objects, including: basketball, candle, carrot, fork, guitar, hat, pumpkin, rose, sunflower, sunglasses, and tennisball. This set of backdoor objects was chosen based on their availability in Open Images, as well as being conceptually distinct from each other, and their likelihood of being fun for game play. We obtained 10-15 different images of each desired object to mitigate overfitting to a particular instance of that object. Every backdoor object was randomly resized, rotated, and placed onto the base image. A sample of the synthetically generated images with backdoor objects is shown in Figure 1.

Table 1: Challenge Name, Poisoned Class, Clean and Poisoned clusters resulting from Activation Clustering (poisoned clusters are bolded), and Backdoor Trigger.

Challenge Name	Activation Cluster	Total Images	Percent Poisoned (%)	Poisoned Class	Backdoor Trigger
Olympic Felines	1 **0**	1225 **2774**	0.16% **71.99%**	Dogs	Cat + Tennis ball
Musical Kitties	1 **0**	1833 **968**	1.91% **91.84%**	Dogs	Cat + Guitar
Cool Dogs	1 **0**	1136 **741**	0.26% **55.42%**	Cats	Dog + Sunglasses
Hippie Puppies	1 **0**	1134 **1667**	0.35% **55.19%**	Cats	Dog + Sunflower

Figure 1: Synthetically poisoned images of dogs with tennis balls (backdoor trigger).

Figure 2: Blur Level decreases upon discovery of each poisoned node.

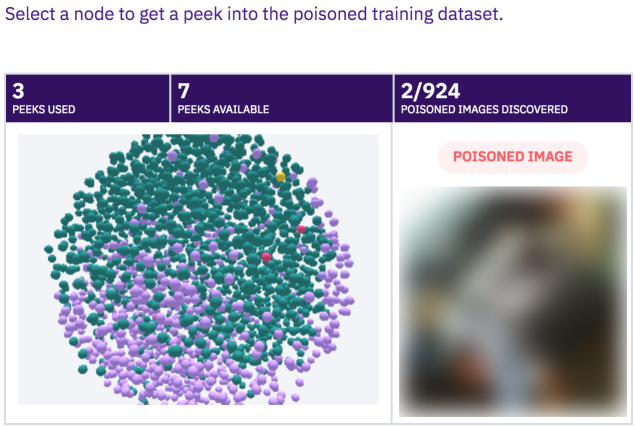

Figure 3: Activation clustering exploration interface.

To determine an effective set of challenges to present to players and create the necessary training data for each, we analyzed the quality of each of the poisoned classifiers trained on these synthetic images, as well. To this end, we trained a Mobilenet on different percentages of synthetically poisoned data [21]. After training, networks were evaluated on 30 selected real-world (i.e. non-synthentic) images containing dogs or cats and the backdoor objects. These were taken from the top results from a popular image search engine to test how well the synthetically trained model extended to real world images, as it is very likely players would use this method to find such images during gameplay. We chose the top 4 performing backdoor objects with thematic diversity to create the final set of challenges tested with users (Table 1). We used synthethic images of household objects as opposed to adding images we found in image searches to accomplish the poisoning because the act of creating the poison needs to be highly controlled and consistently repeated over a large data set in order to work. It?s only after this repetitive process that poisoning with real-world objects are able to fool the classifier.

3.2 Game Play

The game platform is designed such that a player is first presented with a list of available challenges to choose. Upon selecting a challenge, they are redirected to a three-step onboarding process that introduces backdoor attacks and describes Activation Clustering as a possible defense against such attacks. After the onboarding, the user is redirected to the game page for the selected challenge. The game interface is presented as a two-step process. As part of step one, the user is given the option to explore an Activation Clustering graph of the training data through a modal pop-up (Figure 3) to help uncover "clues" about what the backdoor trigger is. Here, the user has the option to click on a graph node to see the associated training image and find out whether that particular image is poisoned or clean (not poisoned). During gameplay, we do not explicitly tell users the players that the dataset is poisoned. We present the two clusters generated by Activation Clustering and encourage them to "peek" to discover a backdoor object (if the dataset is indeed poisoned). As part of step two, the user is asked to submit a guess for what they think the backdoor trigger for this challenge is. The guess consists of two pieces of information - a text-based guess of the backdoor trigger and an image that the user thinks would trigger an inaccurate classification. If an incorrect guess is made, the user is given feedback on how they may be able to improve the accuracy of their guess. They are then asked to retry with a different submission. If a correct guess is made, the challenge is marked as completed and the user is redirected to a success page.

3.3 Game Rules

Users submit their guesses by typing their guess of the backdoor object and uploading an image of the misclassified class and the backdoor object. Users can leverage their front facing camera to upload images or upload a previously saved image from their computer. In the directions, users are instructed to use a search engine

Table 2: Guesses during Gameplay Sessions

	Guesses	Attempts	Win
P1	grass, indoor, guitare, **guitar**	4	No
P2	flower, one yellow flower, **sunflower**	7	Yes
P3	sunglass, **sunglasses**	11	No
P4	cup, telephone, **tennisball**	4	No
P5	glasses, **sunflower**	6	Yes
P6	**sunglasses**	3	Yes
P7	**tennis ball**	1	Yes
P8	**guitar**	4	No

of their choice to find an image of the misclassified class (i.e. cat) and the backdoor object (i.e. guitar). Participants are also able to search the image on their phones and hold up the their phones to the front facing camera of their computers. We add gamification to make the overall experience more fun and to encourage repeat play on the platform. We introduce said gamification by adding the mechanisms of Peeks and Blur Levels.

The concept of a *peek* is introduced to limit the user's access to the graph composed of clean and poisoned nodes. The user is initially given 10 *peeks* to explore the graph, which would have on the order of thousands of total nodes, dependent on the game. In step one of the game, a user is only able to "explore" the graph using the Explore Modal (as seen in Figure 3) if they have a positive number of *peeks* available. In this modal, the user has the option to explore the graph. In this interactive graph, clicking on a previously unvisited node is counted as a *peek*. By clicking on a node and using one of their *peeks*, the user can obtain: the training image associated with the node they just clicked on and whether this image is clean or poisoned. If the user does not have a positive number of *peeks* available, they are unable to access the Explore Modal. The user can earn *peeks* by submitting more guesses under step two of the game. Each guess submitted earns the user one additional *peek* (regardless of whether the guess submitted is correct). This way, *peeks* act as a difficulty mechanism intended to drive users to submit more labeled data. *Peeks* in this context contribute to the challenge aspect of the game, i.e. providing a goal with uncertain attainment for users [26]. The *peek* mechanism encourages participants to upload more images throughout gameplay since uploads are rewarded with *peeks*, yielding potentially more than one photo upload per user.

We also introduce the concept of blur levels. Initially, an image associated with a graph node is intentionally highly blurred, making it difficult for the player to distinguish the backdoor trigger in a poisoned image. Each time the user finds a poisoned node, the blur level decreases, making the image slightly clearer. The user can only see a completely clear image once the blur level is reduced to zero. Blur levels act as a mechanism to prevent users from randomly clicking around the graph until they can identify a recurring backdoor object across some of the images. An example of a poisoned training image at different blur levels is shown in Figure 2. Blur levels serve as rewards for identifying the right images that include poisoned images. When the blur level decreases, user behavior is being rewarded [15].

4 EVALUATION

To formalize our observations on user reactions to a backdoor poisoning game we conducted two sets of evaluations: one set of in-person interviews and an evaluation through a deployment on Mechanical Turk.

4.1 Semi-Structured Interviews

We conducted eight semi-structured interviews with participants identified through a snowball method and word of mouth [19] at a large technical company. Those interviewed included scientists, designers and developers who had different levels of exposure to artificial intelligence tools. Before beginning the study we started with a consent form that describes procedures, risks, benefits, and participant rights. Participants were informed that their participation in this research study was voluntary and they were free to withdraw or discontinue participation at any time, and participation did not involve any significant risks beyond those present in daily life. There was no tangible benefit to participating in this study, although they may find the activity to be enjoyable and/or educational.

Before the interview sessions we asked about demographic information (education level, level of proficiency with AI/machine learning, which AI methods and tools they used in their daily work, and experience with adversarial behavior online). Each participant completed one challenge. They granted permission for the photos they uploaded to be recorded, as well as the game play session and their interactions on the screen. At the end of the session we asked users to give feedback about their thoughts on backdoor poisoning after playing the game and on the game experience (both positive and negative). Questions were open-ended and were recorded for analysis. Table 3 describes the participants in our interviews and their background in machine learning.

4.2 Mechanical Turk

To further evaluate the *Backdoor Game* on a larger scale, we deployed the game on Mechanical Turk. The game experience begins with a consent form that describes procedures, risks, benefits, compensation, and participant rights. Participants are then navigated to a short survey on demographic questions (age, race, education, etc.). Before users begin the game experience, we asked an open-ended question about their prior exposure to artificial intelligence: "What kinds of AI technologies have you interacted with?" After the survey, participants are navigated to the game's three-step onboarding process that provides 1) an overview of how hackers can manipulate training data, 2) gives examples of a what a backdoor trigger would look like (i.e. cat and bowl classified as dog), and 3) introduces Activation Clustering [8, 31], a method that helps to identify malicious backdoors by separating the training set into two clusters.

Once they complete the onboarding process, they are navigated to the game play page, after which they are provided with a code that allows them to be compensated. For the purposes of this study, Mechanical Turk participants only interacted with one challenge, Olympic Felines, since our experiments on effectiveness of different backdoors showed that the tennis ball was the most reliable backdoor trigger. Participants were compensated commensurate

Table 3: Interview participants role, exposure to AI, and challenge played.

	Role	Exposure to AI	Challenge
P1	Research Scientist	Image captioning, GANS, speech audio visual recognition; pytorch	Musical Kitties
P2	Research Scientist	Works on a ML Product Offering; SPSS; R; Python; Research on improving next generation of ML tools	Hippie Puppies
P3	Research Scientist	Off the shelf solutions; chatbot-related classifiers	Cool Dogs
P4	Research Scientist	Applied Data Science; Trust in AI; Explainability and Fairness; ML in Social Good and Humanitarian issues;	Olympic Felines
P5	Software Engineer	Regression models; feature engineering; classification; python; numpy; scipy; sci-kit learn	Hippie Puppies
P6	Software Engineer	No technical side; building things around AI tools but not directly working with it	Cool Dogs
P7	Software Engineer	No technical side; building things around AI tools but not directly working with it	Olympic Felines
P8	Designer	No technical side; building things around AI tools but not directly working with it	Musical Kitties

to federal minimum wage. With any game, there is the possibility that participants may not be able to complete the game. In a Mechanical Turk task, it is incredibly important that participants are compensated even if they are not able to complete the game. For this reason, we made a link visible that allowed them to navigate to the final page on which they would receive their Mechanical Turk code if they were unable to complete the game after five guess submissions.

5 RESULTS

5.1 Interviews

From our interviews, we identified themes for the questions we asked: favorite aspects of game and least favorite aspects of the game. Least favorite aspects included: graphical representation of clusters, the ratio of uploads to peeks, the search and upload mechanism, dataset used, unclear direction, and more game mechanics to motivate. Favorite aspects included: enjoyment of search and learning about backdoor poisoning. In this section, we discuss these areas and provide concrete examples of the different opinions and feedback.

5.1.1 Game Play Results. The uploads, guesses, and number of attempts were recorded for different participants. Not all participants were able to correctly guess both the right word describing the backdoor object and provide an image containing the object that triggered the misclassification (i.e. solve the challenge). Half of the participants were able to complete the challenge, while all participants were able to either upload a correct picture or correctly guess the backdoor object through textual input. Table 2 shows the guesses for each participant (with the final and correct backdoor guess bolded), the number of guesses for each participant and whether they successfully completed a challenge. Common mistakes for text input included grammatical errors relating to quantity (singular versus plural) and spelling mistakes. For example, one participant misspelled "guitare" and another input "sunglass" instead of "sunglasses". As seen in Table 2, words that are spelled incorrectly do not yield a success in gameplay because the system yields success in the case of an exact string match. In future versions of the system, we will consider stemming and inclusion of words that might not be spelled correctly but are deemed "close enough" (i.e guitare, sunglass). To win a challenge, a user must guess the backdoor correctly by entering a text label and uploading an image that triggers a misclassification - a photo with the backdoor trigger object included with an instance of an object from the "clean" class, e.g. a photo of a cat (the class) with a guitar (the backdoor object), that is incorrectly classified as a dog. The majority of participants had to upload several pictures until they were able to correctly guess the backdoor. Some participants were confused. Since the text box only prompts the name of the backdoor, they uploaded an image of just the backdoor, i.e. a photo of a water bowl instead of a photo of a water bowl and a cat. For example, Participant 4 guessed that the backdoor is a cup and uploaded a photo of a cup (see Figure 4).

We classify photos that do not fool the AI (i.e. both accurately detected the misclassified class, such as cat, and accurately detected the backdoor object) into four groups: **irrelevant photos** as dummy photos that have no relevance to the backdoor or the objects being classified. These include photos that participants captured of themselves or a completely random object not related to what they observed in the Peeks, **incorrect photos**: photos based on a participant's observances, but are still not the correct guess, **partially-correct photos**: photos of just the object being classified (cat, dog) or just the backdoor (sunglasses), and **correct but misclassified photos**.

5.1.2 Game Mechanics and UI: Least Favorite Aspects. During game sessions, participants were encouraged to think out loud [36] and identify issues during game play. Upon completion of the game, we asked participants to list both their favorite and least favorite things about the game experience. Participants identified aspects of the game that they found confusing: blurriness of images, the Activation Clustering visualization, incongruity between text and image upload, and upload of dummy photos.

Blurriness of Images. Blurred images were incorporated as a mechanism to introduce a more challenging experience to users. Images (both clean and poisoned) were shown to users with a particular blur level. As a user discovered more poisoned images, the blur level decreased to finally reveal the image and the backdoor object. Many participants expressed confusion over the blurriness and did not understand immediately that it was a game mechanic.

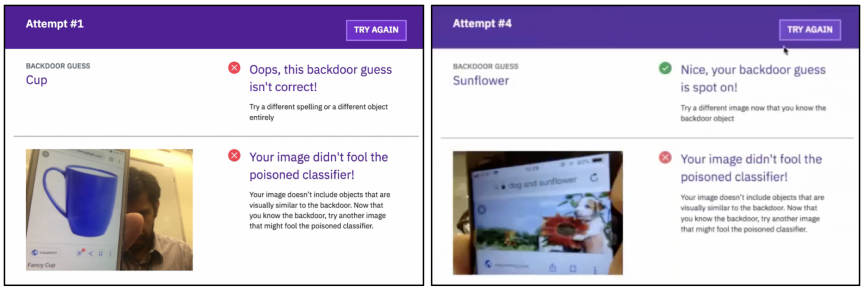

Figure 4: On left, example of a "partially-correct" photo, of the backdoor, without the misclassified entity (i.e. dog). On right, example of "Correct but Misclassified" submission. P3's submission of "dog + sunflower" in which the image did not fool the poisoned classifier.

One participant expressed confusion about differentiation between clean and poisoned images when both are blurry.

> "Everything is so blurry that I am not quite sure why is one clean and one is not on the blurrier ones." (P8)

Activation Clustering Graph. The first step of the game is to explore the Activation Clustering graph, consisting of a 3-D graph of different colored nodes resulting from Activation Clustering. Participants were encouraged to think out loud as they navigated, panned, zoomed, and clicked on different nodes. As a result of their navigation, many expressed open questions about aspects of the visualization they found unclear or confusing. One critique that emerged was that multiple participants questioned the purpose of the 3D rendering of the visualization and whether the ability to zoom and pan benefited the players in any way.

> "I might even make it a 2D projection over a 3D one because I think that gets the point across and it would be easier to make it in a sense." (P4)

Ratio of Uploads to Peeks. Participants are allowed 10 *peeks* upon starting the game, giving them the opportunity to explore 10 nodes in the Activation Clustering graph. Upon exhausting their Peeks, they must submit a guess (uploaded photo and relevant text) for the backdoor. Every submission rewards them an additional *peek* that they can then use to explore the data. In game play sessions, participants complained about the number of *peeks* awarded per upload and surmised that given this ratio, their guesses about photos were useless since they were entering guesses based on observations of blurry photos.

> "In terms of guesses... a person is not going to describe a blurry picture well." (P4)

Search and Upload Mechanism. Participants also provided critical feedback regarding the upload interaction. They asked for an option to search for images through the interface rather than by file upload or search on their phones.

> "If you had the option to do a search through the interface, I think that would be easier." (P4)

Real World Data. Two participants (P1 and P2) expressed their preference for real world data in order to more clearly understand how an adversary would launch a backdoor attack and how it would be presented in a real life scenario. One participant mentioned being interested in seeing backdoor poisoning on datasets in different domains.

Additional Motivation. Two participants, P4 and P8, discussed their need for more challenging aspects. P4 suggested that the game would be more enjoyable if there was a timed component. P8 asked for achievements/rewards within game to see how he compared against others.

> "If it gave me achievements, I just don't understand what I'm gaining other than the fact that I'm learning something." (P7)

5.1.3 Game Mechanics and UI: Favorite Aspects. Participants identified their favorite aspects of the game. Emerging themes for favorite aspects are: enjoyment of photo search and learning about a new technology.

Enjoyment of Search. Some participants expressed that their favorite aspect of the game was searching for photos and trying to identify the details in the photos that would help their search. Participants mentioned the "fun" factor of searching for images of cats and dogs with objects in them and generally expressed enjoyment over this experience.

> "It was fun to search for pictures." (P2)

Learning. While most participants had some knowledge of backdoor poisoning, they also agreed that this tool is a better way of understanding backdoor poisoning with examples. All participants agreed that playing the game contributed to their knowledge of backdoor poisoning attacks.

> "I enjoyed playing because it has a pretty UI and I learned something new about machine learning." (P7)

5.1.4 Challenges. We observed that within challenges, there was a range in the number of submissions before achieving (or not achieving) success. For example, for the "Cool Dogs" challenge, P3 attempted 11 submissions without successfully completing the challenge, while P6 was successful after 3 submissions. Neither P1 nor P8 successfully completed"Musical Kitties", because both participants were not able to find photos that fooled the classifier. Both P2 and P5 completed "Hippie Puppies". P7 successfully completed "Olympic Felines" after only one attempt, yet P4 attempted 4 times without success. A larger scale study needs to be conducted to observe differences across challenges with respect to ease of challenge and user behaviors specific to certain challenges. We observe that participants who observed more poison images before their first upload (P6 and P7), were both able to successfully finish the challenge with fewer attempts. This finding makes sense since the more poisoned images are discovered, the clearer the images appear.

5.1.5 Additional Functions of Game. In their feedback, the participants expressed the potential value of this game to collect annotated

Figure 5: A crowdworker recruited from Mechanical Turk submits three photos of themselves with a cat (misclassified class) and a tennisball (the backdoor object).

data for image processing researchers. In the past, there has been effort in creating interactive and immersive games for large scale data collection [39]. Collecting annotations for images helps build reliable training datasets for object detection algorithms [14]. Annotated images have also proven to be useful for search, specifically image search in which the content of an image is difficult to extract [13].

> "You can use this to get more examples of backdoor data. And also you can get incorrect guesses as other types of annotated data." (P4)

5.2 Deployment on Mechanical Turk

We took into consideration the feedback from our small-scale interviews and implemented a few changes before larger scale deployment on Mechanical Turk. First, we changed the onboarding to make directions more clear. Per participant feedback, we also changed the Activation Clustering graph from 3D rendering which participants found confusing to 2D rendering. For deployment of the game on Mechanical Turk, we introduced an alternative visualization with 2D rendering in which poisoned and clean images are represented as two separate groups. The Activation Clustering resulted in the color coding of the nodes, so the users can see the difference. The visualization uses both grouping and colors to differentiate between the two clusters and avoid using additional features (like 3D zooming and panning) that may confuse or distract users from the point of the visualization.. We also increased the *peek*-to-upload ratio, awarding three *peeks* per upload instead of one.

5.2.1 Participants. We deployed this game to 68 Mechanical Turk users to observe how they explored poisoned data sets that have been clustered, their interactions with the system, and the various kinds of guesses they submit. Our results show the potential of this tool for the collection of multi-object labeled images. Of the 68 participants who participated in this study, 16.2% were between 18 to 24, 63.2% were between 25 to 34 years old, 19.1% were between 35 to 44 years old, and 1.5% were between 45 to 54 years old. 1.5% of participants had a PhD, 57.3% of participant had a Bachelors degree, 13.2% had some college experience, 13.2% had an Associate or other technical degree, 1.5% were High School educated and 13.2% had no degree. 90% of participants indicated that English was their native language. Other languages included Bangla, Tamil, Malayam, and Hindi. At the end of game play we asked participants to provide feedback on their experience in an open-ended prompt. Below we list the types of photos we observed.

5.2.2 Artificial Intelligence Exposure. In the preliminary survey we asked participants to list the kinds of AI technologies with which they have interacted. We can divide the different types of AI technologies with which they interacted into two groups: as crowdworkers who contributed to a training data set, and as end users. 73% of the Turkers had, at some point, participated in the creation of a training dataset for the improvement of AI technologies (voice recognition, sentiment analysis, chatbot trainings). 19% of Turkers listed being users of technologies that incorporate AI for their HITS as crowdworkers. Technologies included: "Amazon Alexa", "Google now" or "playing chess against the computer." 8% of participants had not interacted with AI technologies before through their crowdwork. We asked this question as a way of understanding how much exposure participants have had to other AI technologies.

> "I have worked for HITs that wanted me to speak to them, such as speaking to an online restaurant ordering system instead of typing." (P#22, Training AI)

5.2.3 Game Play Results. Of the 68 Mechanical Turk participants, 45 (66%) were able to win the game (with an average of 7 *peeks* (node clicks) and 3 attempts at submission), while 23 opted to give up after multiple tries. We also discovered similar patterns in the kinds of photo uploads and guesses we observed in our semi-structured interviews: irrelevant photos, incorrect photos, partially-correct photos, and correct but misclassified photos. However, two other notable categories of photos emerged which we list below:

(1) **Correct but copied.** Some submissions consisted of photos that participants had observed either through the activation clustering graph or photos that appeared elsewhere on the platform (on hint pages). These included synthetic images. This type of interaction - we believe - is on par with cheating the system, since we give explicit directions about using photos from search engines or uploading through a front-facing camera.
(2) **Original Submissions.** (see Figure 5) Participants who owned cats attempted to submit an original photo of their pet while holding a ball.

A total of 191 guesses were submitted as result of this game play. 39% of these photos were correct photos, while 61% did not fool the Classifier. The correct photos consisted of *original submissions* (3% of entire dataset) and correct but copied photos (5% of entire dataset). Incorrect photos consisted of completely irrelevant photos (25% of entire data set), partially correct photos (29% of entire data set), and correct but misclassified photos (7% of entire dataset).

5.2.4 Unique and Diverse Submissions. One of the most notable outcomes of the Mechanical Turk study was the discovery of *Original Submissions*, i.e. users leveraging their own personal objects and their pets to fool the classifier. In Figure 5 (on the left), a crowdworker attempts to fool the classifier with two different cats while holding a tennisball (a closer look at the image reveals that the right most image is a different cat than the other images and that cat was finally able to fool the classifier). The game play mechanics of our system encouraged users to interact with objects at home to submit example backdoor objects. We also see examples of different backdoor objects submitted that were not correct in fooling the classifier (since they are missing the backdoor object) but are still

valuable submissions since they can be used to further study the effectiveness of backdoor objects.

5.3 Limitations

We acknowledge that our interview participants are more knowledgeable about AI than the average user. However, the insights and feedback from our interviews still allowed us to tremendously improve the application before larger scale deployment on Mechanical Turk. Furthermore, the average user who interacts with a system such as this one will likely have more interest in and exposure to AI. We also acknowledge that in the real world, poisoning a dataset is not as simple as poisoning a cat/dog classifier with different household objects. However, the *Backdoor Game* is the first system of its kind to collect backdoor objects. Future work can expand on this by running similar experiments with datasets that are more varied (i.e. the trace of a watermark, a single pixel color in the corner of the image. An administrator Poison Generator utility allows AI researchers who would use the platform to upload images and create challenges testing many different kinds of classifiers (beyond dogs and cats) and poisoned objects (watermarks). Future research can more closely investigate these kinds of interactions. *Backdoor Game* supports binary classifiers, but future iterations can explore multi-class classifiers. Our evaluations are meant to demonstrate the potential of this tool; to show how users interact with the system both at small scale and large scale (through a crowdsourcing platform). We also acknowledge that the synthetically generated poisoned images include out of place backdoor objects. However, we know that players are looking for patterns in poisoned images, rather than learning to recognize the out-of-place nature of synthetic objects because they expressed that they were looking for patterns during the think-aloud game play sessions. Future work can more rigorously test hypotheses around the collection of backdoor data and the creation of new datasets.

6 DISCUSSION

6.1 Motivating the Collection of Diverse and Unique Backdoor Objects

We use gamification to motivate users to engage with the clusters of poisoned and clean nodes and to submit their guesses of the backdoor objects. In this section, we use our findings to describe how we can further motivate participants to submit unique and diverse backdoor objects. In this paper, we show how some of our participants leveraged our existing game mechanics to submit diverse and unique backdoor objects and we offer additional recommendations on how to further motivate this type of engagement.

6.1.1 Original Submissions. The photos collected in the large scale Mechanical Turk deployment demonstrate that this system is able to collect backdoor objects. Participants in our Mechanical Turk deployment submitted photos of themselves with different objects in their households as a part of the game play. At the outset of this study, we were not anticipating that users would use the front facing camera for this functionality, and instead provided some direction in the hint to use the front facing camera to take photos of a photo displayed on their smartphone. However, users who had cats and tennis ball-like objects in their vicinity used the opportunity to submit original submissions. Participants used their front facing camera to submit data and we were able to collect original photos with their respective annotations. While the system is currently a desktop application, the opportunity for uploading original submissions increases if a mobile version is made available. Incorrect photos with labels matching to the photos uploaded are also valuable for determining new potential backdoor objects.

6.1.2 Dummy Photos. We observed that participants in both the interview study and the Mechanical Turk study uploaded photos that were completely irrelevant to the prompt. One way to prevent the uploading of dummy photos is to use a classifier to ensure that the misclassified class (cat, dog, etc.) is in the image. This requirement would also ameliorate the confusion around submissions from users to *just* include the backdoor image. Furthermore, users may be motivated to upload dummy photos or photos that are completely irrelevant to increase the number of *peeks*. Some users were frustrated about the number of *peeks* awarded after every upload. Future iterations can award more Peeks per upload to encourage user to upload better quality photos and not just upload dummy photos to increase their *peeks*.

6.2 A Co-working space: Domain Experts and Crowdworkers

Through the design and development process of the *Backdoor Game*, we are able to identify what makes a good backdoor in terms of the frequency in which a model was fooled when a particular backdoor was present. We found that tennisballs, for example, perform better than other objects like carrots or forks likely due to their distinctive color and symmetrical nature.. Through **deployment** of *Backdoor Game*, we are able to go one step beyond testing how models are fooled, and observe how crowdsourced photos actually perform in fooling the classifier. In the *Backdoor Game*, a challenge can be created and crowdworkers can determine if different photos with the misclassified class and backdoor object can actually fool the classifier. As observed in both our interviews and MTurk studies, it is not enough for a backdoor object and misclassified object to be present in the photo, but in some instances, color, quality of photo, contrast and other photo characteristics are also important. The *Backdoor Game* allows AI Security Researchers from a variety of domains (medical, autonomous vehicles, etc.) to test backdoor objects in images of their domains/contexts.

6.3 Ethical Considerations

Microtasking crowdwork, the kind we observe on platforms like Amazon Mechanical Turk have been referred to as the "last mile of automation"[20]. As widespread as such crowdsourcing platforms are for research, we must acknowledge that many of these individuals are not one-time users, but in fact individuals who earn their income on these platforms [42]. While the *Backdoor Game* can aid AI security researchers to explore backdoor poisoning attacks on vulnerable classifiers and collaborate with non experts to collect more data that assists researchers to study poisoned datasets, we considered the role of the crowdworkers in the platform to ensure that the platform is not only beneficial to ML researchers but also offers features that non-expert crowdworkers would find engaging

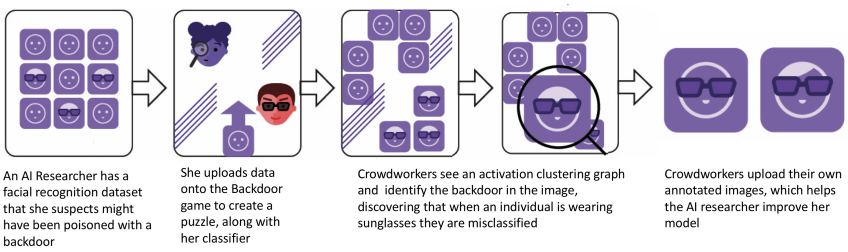

Figure 6: A description of how an AI researcher can use the the *Backdoor Game* to identify backdoor objects in her dataset.

and not limited to a *routinizable* set of tasks. A motivation for including interactive and engaging game play mechanics was for the benefit of crowdworkers. If we consider an "in the wild" context in which non-experts access the *Backdoor Game* directly and not through a Mechanical Turk environment, several functionalities of the game were implemented to keep the game engaging specifically for them: 1) the use of "fun" synthetic objects as backdoors and 2) blur mechanism to make the game more challenging. In both the semi-structured interview study as well as the Mechanical Turk study participants reported that they enjoyed the game and found it to be enjoyable. When designing such games, a byproduct should be that crowdworkers learn something about the underlying technology that they are creating. The *Backdoor Game* is not a pedagogical tool nor did we test learning outcomes among participants, but as an ethical consideration, information was provided so that learning can be a potential byproduct of the tool.

7 FUTURE WORK AND CONCLUSION

When designing this system, we were motivated by the novelty of backdoor poisoning and the fact there has been no other system built to demonstrate and study how such attacks work in a space where training sets and backdoor objects could be studied through collaboration between AI security researchers and crowdworkers. We presented results comparing the most effective backdoor objects which we used to create poisoned training data for a binary dog/cat classifier. We believe that the game we have created is a compelling platform to further test the effectiveness of backdoor objects. This work paves the path for users (especially AI Security researchers) to upload their own puzzles in the future. Through evaluation of the *Backdoor Game* with eight different participants and a larger scale evaluation with 68 crowdworkers, we demonstrate how this tool shows the effectiveness of different backdoor objects and potential to collect content from players, including valuable original submission photos.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: a system for large-scale machine learning.. In *OSDI*, Vol. 16. 265–283.

[2] Josh M Attenberg, Pagagiotis G Ipeirotis, and Foster Provost. 2011. Beat the machine: Challenging workers to find the unknown unknowns. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

[3] Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, and Jaehoon Amir Safavi. 2017. Mitigating poisoning attacks on machine learning models: A data provenance based approach. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. 103–110.

[4] Michael S Bernstein. 2010. Crowd-powered interfaces. In *Adjunct proceedings of the 23nd annual ACM symposium on User interface software and technology*. 347–350.

[5] Jonathan Bragg and Daniel S Weld. 2016. Optimal testing for crowd workers. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. 966–974.

[6] LJ Cao, Kok Seng Chua, WK Chong, HP Lee, and QM Gu. 2003. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. *Neurocomputing* 55, 1-2 (2003), 321–336.

[7] Dennis Chao. 2001. Doom as an interface for process management. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 152–157.

[8] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. 2018. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering. *arXiv preprint arXiv:1811.03728* (2018).

[9] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526* (2017).

[10] Lydia B Chilton, Greg Little, Darren Edge, Daniel S Weld, and James A Landay. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1999–2008.

[11] Otto Chrons and Sami Sundell. 2011. Digitalkoot: Making old archives accessible using crowdsourcing. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

[12] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, et al. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756.

[13] Ritendra Datta, Weina Ge, Jia Li, and James Z Wang. 2007. Toward bridging the annotation-retrieval gap in image search. *IEEE MultiMedia* 14, 3 (2007).

[14] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)* 40, 2 (2008), 5.

[15] Sebastian Deterding. 2012. Gamification: designing for motivation. *interactions* 19, 4 (2012), 14–17.

[16] Steven Dow, Anand Kulkarni, Brie Bunge, Truc Nguyen, Scott Klemmer, and Björn Hartmann. 2011. Shepherding the crowd: managing and providing feedback to crowd workers. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 1669–1674.

[17] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1625–1634.

[18] Andrej Gisbrecht, Alexander Schulz, and Barbara Hammer. 2015. Parametric nonlinear dimensionality reduction using kernel t-SNE. *Neurocomputing* 147 (2015), 71–82.

[19] Leo A Goodman. 1961. Snowball sampling. *The annals of mathematical statistics* (1961), 148–170.

[20] Mary L Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Eamon Dolan Books.

[21] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[22] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*. 64–67.

[23] Marius Kloft and Pavel Laskov. 2010. Online anomaly detection under adversarial impact. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 405–412.

[24] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali,

Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. 2017. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html* (2017).

[25] Walter S Lasecki, Jaime Teevan, and Ece Kamar. 2014. Information extraction and manipulation threats in crowd-powered systems. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 248–256.

[26] Thomas W Malone. 1981. Toward a theory of intrinsically motivating instruction. *Cognitive science* 5, 4 (1981), 333–369.

[27] Winter Mason and Duncan J Watts. 2009. Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, 77–85.

[28] Cristina Ioana Muntean. 2011. Raising engagement in e-learning through gamification. In *Proc. 6th International Conference on Virtual Learning ICVL*, Vol. 1. 323–329.

[29] Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D Joseph, Benjamin IP Rubinstein, Udam Saini, Charles Sutton, JD Tygar, and Kai Xia. 2009. Misleading learners: Co-opting your spam filter. In *Machine learning in cyber trust*. Springer, 17–51.

[30] Nikolaos Pitropakis, Emmanouil Panaousis, Thanassis Giannetsos, Eleftherios Anastasiadis, and George Loukas. 2019. A taxonomy and survey of attacks against machine learning. *Computer Science Review* 34 (2019), 100199.

[31] Marko Puljic and Robert Kozma. 2005. Activation clustering in neural and social networks. *Complexity* 10, 4 (2005), 42–50.

[32] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. 2018. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*. 6103–6113.

[33] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. 2017. Certified defenses for data poisoning attacks. In *Advances in neural information processing systems*. 3517–3529.

[34] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

[35] Brandon Tran, Jerry Li, and Aleksander Madry. 2018. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems*. 8000–8010.

[36] MW Van Someren, YF Barnard, and JAC Sandberg. 1994. The think aloud method: a practical approach to modelling cognitive. (1994).

[37] Fabio Viola. 2011. *Gamification-I Videogiochi nella Vita Quotidiana*. Fabio Viola.

[38] Luis Von Ahn. 2006. Games with a purpose. *Computer* 39, 6 (2006), 92–94.

[39] Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 319–326.

[40] Luis Von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Commun. ACM* 51, 8 (2008), 58–67.

[41] Gang Wang, Tianyi Wang, Haitao Zheng, and Ben Y Zhao. 2014. Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*. 239–254.

[42] Vanessa Williamson. 2016. On the ethics of crowdsourced research. *PS: Political Science & Politics* 49, 1 (2016), 77–81.

[43] Haoqi Zhang, Edith Law, Rob Miller, Krzysztof Gajos, David Parkes, and Eric Horvitz. 2012. Human computation tasks with global constraints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 217–226.