Nightmare Machine: A Large-Scale Study to Induce Fear using Artificial Intelligence

Pinar Yanardag[1] **Nick Obradovich**[2] **Manuel Cebrian**[2] **Iyad Rahwan**[2]

[1]Bogazici University [2]MPI for Human Development

yanardag.pinar@gmail.com {obradovich,cebrian,rahwan}@mpib-berlin.mpg.de

Abstract

As Artificial Intelligence makes strides in emulating human performance in analytical tasks, an important question surfaces: can machines induce extreme human emotions at scale? In this work, we investigate a case study, Nightmare Machine (nightmare.mit.edu) towards a particular emotion; *fear*. We use a deep-learning based approach that induces states of anxiety and negative affect by generating de-novo eerie images. Our system attracted the attention of hundreds of thousands of participants from 147 countries who produced over 1,000,000 evaluations of the generated images. First, we perform various exploratory data analysis tasks on the collected data in order to investigate the potential of the generated images, such as whether there exists a correlation between preferences of the participants based on geographic location. Then, we perform a validation study on $n = 752$ subjects to verify whether the generated images psychologically move people on psychometrically validated measures of effect and anxiety such as I-PANAS-SF (Thompson 2007) and STAI-SF (Marteau and Bekker 1992). Our experiments show that the generated images produced statistically significant increases in negative affect and state anxiety compared to the control images. We make our dataset publicly available at https://github.com/catlab-team/nightmaremachine.

Introduction

Recent advances in artificial intelligence achieved significant breakthroughs that exceed human capabilities and gained an immense amount of attention due to their success in several areas including computer vision (Voulodimos et al. 2018), language modeling (Jozefowicz et al. 2016), and robotics (Pierson and Gashler 2017). Deep learning, a sub-field of artificial intelligence, enabled researchers to discover complex patterns in extremely large datasets and widely deployed in academia and industry (Le et al. 2020; Luckow et al. 2016). Deep learning based systems started to gain popularity when a convolutional neural network outperformed a large-scale image classification task at ImageNet Large-Scale Visual Recognition Challenge (Russakovsky et al. 2015). This success, supported by the remarkable developments of powerful processors (GPUs), and explosive growth of data, enabled the rise of deep learning. Since then, deep learning has become state-of-the-art approach for a large variety of problems, including image processing (Hemanth and Estrela 2017), natural language processing (Deng and Liu 2018), speech recognition (Deng and Yu 2014) and even defeating world's best Go and chess players (Silver et al. 2016; Silver et al. 2017) or beats human champions in Jeopardy such as IBM's Watson (Watson 2014). As Artificial Intelligence makes strides in solving challenging analytical problems like checkers (Samuel 1967), chess (Silver et al. 2017) or video-games (Vinyals et al. 2017) society takes solace in the implicit belief that the subset of human tasks that rely on the understanding, managing, and inducing human emotions are still far from the ability of machines to outperform humans. But are they? Can computers learn to induce emotions faster and better than humans can?

Detecting emotion can be considered as a first step towards inducing emotion. Machine learning is enjoying rapid advancement on this front (Hossain and Muhammad 2019) and initial algorithms were able to detect positive and negative emotion (Liu, Zheng, and Lu 2016). More recently, Natural Language Processing (NLP) has been able to infer not only the mood expressed in text but also irony and sarcasm (Schifanella et al. 2016) and, in some cases, humor (Chen and Soo 2018). Affective Computing, computational tools to sense and improve human-computer communication, is also enjoying a steady revival (Picard 2000). Going from the detection of emotion to the induction of emotion is, however, a big leap, and one that we tackle in this work. Can Artificial Intelligence not only detect but induce specific emotions in humans, in particular, fear? Attempts at fear induction taking the form of stories and visual images pervade the history of human culture. Creating a visceral emotion such as fear remains one of the cornerstones of human creativity. In this work, we explore a way to combine deep learning and crowd-sourcing to test whether fear can be induced at scale. To our knowledge, we are the first to automate the production of scary images. While computers can detect images that may be upsetting, there's no previous literature on seeing whether computers can generate them. In this work, we propose a deep-learning based approach to a particular emotion; *fear*, and explore whether we can induce states of anxiety and negative affect with the generated images. Our platform gained wide attention from all over the world and collected over one million votes on the generated images from 147 countries.

Related Work

In this section, we first cover related work in the intersection of emotions and artificial intelligence. Then, we discuss crowd-sourced tools that utilize artificial intelligence applications. Finally, we briefly cover related work in generative models.

Emotions and Artificial Intelligence

Recent advances in deep-learning encouraged researchers to investigate the usage of artificial intelligence in terms of emotions. Most of the existing work focuses on the classification or detection of certain emotions on facial data. (Ng et al. 2015) used a convolutional neural network architecture (CNN) combined with a transfer learning approach and performed emotion recognition using EmotiW (Kahou et al. 2013), a face expression dataset that includes a wide range of emotions including *happy, sad, surprised, fear, angry* and *disgusted*. (Jain et al. 2018) tackled facial expression recognition (FER) using a hybrid convolution-recurrent neural network that extracts the relations within facial images by using the recurrent network for temporal dependencies.

Another line of work focuses on emotion recognition from speech. (Hossain and Muhammad 2019) proposed an audio-visual emotion recognition system using CNNs on an emotion dataset that consists of speech and video. They proposed to process speech signal in the frequency domain and used corresponding Mel-spectrograms as an image which is fed to a CNN. The output is fused with video signals and fed into two consecutive extreme learning machines and a support vector machine (SVM) for the classification of the emotions. (Satt, Rozenberg, and Hoory 2017) used an end-to-end deep neural network on raw spectograms. They combined a noise reduction solution based on harmonic filtering to perform emotion recognition from speech under limited latency constraint and achieved state-of-the-art accuracy on popular benchmarking dataset IEMOACP (Busso et al. 2008).

Several studies investigated the potential of using deep learning for empathy-related applications. For instance, (Barmar 2017) used a neural network based approach to classify empathy and personal distress on facial muscle activities. They also investigated which facial muscle movements contribute the most to predict empathy. (McQuiggan et al. 2008) proposed a data-driven inductive approach for learning empathy models including both reactive and parallel empathetic expression. Their approach focuses on the observation of empathy in action and tries to understand the psychological aspects of empathetic assessment. (McQuiggan and Lester 2006) proposed a data-driven framework for extracting models of empathy that are empirically grounded from observations of human to human social interactions. (Gibson et al. 2016) used a deep neural network system for predicting empathy ratings from transcripts of counselors. To pursue this goal, they utilized a recurrent neural network that matches the transcript of a speaker with a task-specific behavioral act.

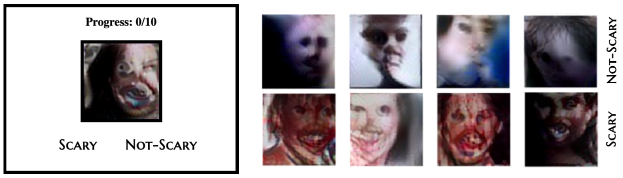

Figure 1: The voting page from the platform is shown on the left where a random image is displayed to the participant and asked them to vote whether they find the image **Scary** or **Not-Scary**. Sample faces from our classification dataset are shown on the right where the top four images are voted as **Not-Scary**, and the bottom four images are voted as **Scary**.

Crowd-sourced AI Tools and AI-based Creativity

The field of machine learning gained attention due to the remarkable results in several important tasks in computer vision, natural language processing and robotics areas. Recently, researchers focused on combining generative models with crowd-sourcing efforts for creative applications.

Deep Dream Generator (DeepDreamGenerator) is a computer vision tool that helps users to experiment with deep learning algorithms for creativity. Neural style transfer algorithm (Lee et al. 2018) enabled users to experiment with painting styles on any given image (DeepArt). In addition to computer vision tools, music-based platforms such as Magenta (Magenta) offers a large collection of music-based tools using a recurrent neural network based system that generates notes based on melodies provided by the users. Botnik (Botnik), GPT-2 (Radford et al. 2019a) are among the text-based platforms heavily explored for creative writing. Botnik offers a *keyboard*-based interface where users can collaboratively create AI-assisted text-based content. GPT-2 (Radford et al. 2019b) model uses large-scale datasets which helped users to create a variety of applications ranging from novels (GPT) to poetry (Branwen). Computationally creative Twitter bots are also utilized in several studies. (Yanardag, Cebrian, and Rahwan 2021) explores Twitter as a medium for creating horror stories in a collaborative fashion with Twitter users. (Oliveira 2017) proposes a bot that posts poems inspired by Twitter trends.

Generative Models

Generative Adversarial Networks (GANs) (Goodfellow et al. 2014a) aim to model the image space so that they generate images that are indistinguishable from those in the dataset. The adversarial part of the network detects whether the produced images are from the training dataset (or fake), and the generative part tries to create images that are similar to the dataset. DC-GAN (Radford, Metz, and Chintala 2015) is one of the first GAN models that directly extends the GAN architecture by using convolutional layers in the discriminator and convolutional-transpose layers in the generator. StyleGAN (Karras, Laine, and Aila 2018) and StyleGAN2 (Karras et al. 2020) are among popular GAN approaches that generates high-resolution images. They use a mapping network with an 8-layer multilayer perceptron (MLP) which fits input latent code onto an intermediate la-

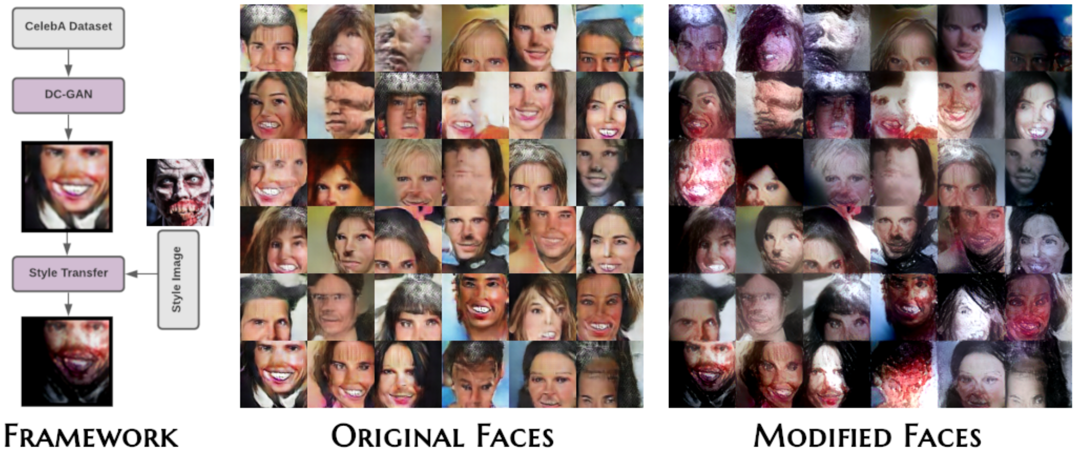

Figure 2: Architecture of the image generation pipeline is shown on the left where the trained DC-GAN model generates faces that are fed into neural style transfer to produce the final image. Original faces generated by DC-GAN are shown in the middle and the corresponding modified faces after applying neural style transfer are shown on the right.

tent space. BigGAN (Brock, Donahue, and Simonyan 2018) is another large-scale model trained on ImageNet (Russakovsky et al. 2015) and utilizes the intermediate layers by taking the latent vector as input as well as a class vector that acts as a conditional information.

Methodology

In this section, we first give details of the data collection platform and discuss the generation methodology of the images used in the experiment. Then, we discuss the details of the collected dataset and share several insights and statistics about the dataset.

Data Collection Platform

We launched a public platform `http://nightmare.mit.edu` that invites users to participate in an online poll. The poll is designed to be simple (see Figure 1 for the survey page) where random GAN-generated images were shown to participants and ask them to vote whether they find the shown image **Scary** or **Not Scary**. A total of 100 images were randomly shown to the participants in batches of 10 from a pool of 500 generated images. Each image is associated with a unique identifier for data analysis purposes. After voting every 10 images, a customized page is shown to the participants where a uniquely generated grid of images is shown as a reward.

Over the course of 9 months, our platform is attracted more 300K participants from 147 countries that resulted in 1,091,345 votes on 500 GAN-generated images. In addition to collecting the voting information, we also collected the IP addresses of the users in order to track cross-country preferences on the images. An Institutional Review Boards (IRB) approval is obtained for collecting the votes and IP information.

Image Generation

A two-stage architecture that combines generative adversarial networks (GANs) (Goodfellow et al. 2014b) and neural style transfer (Gatys, Ecker, and Bethge 2015) is used for generating the images used in this study. GANs are a class of neural networks that have gained popularity in recent years, with the most common application area being image generation. GANs estimate generative models with an adversarial process by simultaneously training two models: a generative model G that represents the data distribution, and a discriminative model D that estimates the probability of whether a sample comes from the model distribution or the data distribution. GANs train the generator G and the discriminator D through playing a mini-max game: D maximizes the expected log-likelihood of distinguishing real samples from the fake ones, and G maximizes the probability of D making a mistake. The equilibrium of this game is reached when the generator is generating fakes that look like real as if they came directly from the training set, and the discriminator can not distinguish between the fake ones and real ones with a 50% confidence.

We used DC-GAN (Radford, Metz, and Chintala 2015) model that directly extends the GAN architecture by using convolutional layers in the discriminator and convolutional-transpose layers in the generator. The all convolutional net (Dosovitskiy, Tobias Springenberg, and Brox 2015) is used in its generator and discriminator which replaces the deterministic spatial pooling functions with strided convolutions and enables the network to learn its own spatial downsampling, along with batch norm layers, and LeakyReLU activations. The input to the discriminator is a $3 \times 64 \times 64$ image and output is a probability of the input belonging to the real data distribution. The input to the generator is a latent vector drawn from a prior distribution and the output is a $3 \times 64 \times 64$ image.

One approach to generating the images for our task is to

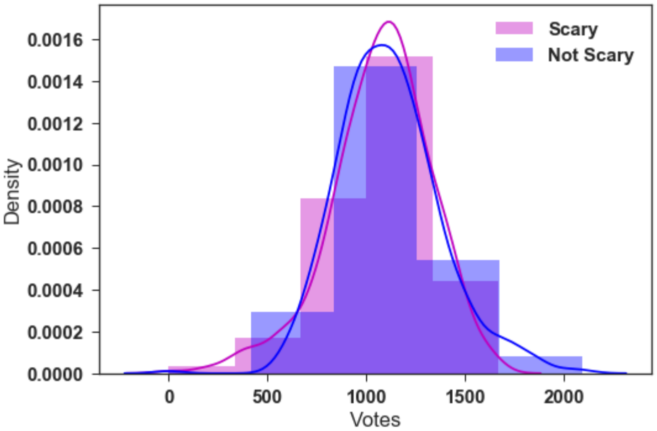

Figure 3: Distribution of **Scary** and **Not Scary** votes where the number of votes appears to be distributed equally with a slight shift on the right towards *Scary* votes.

directly train a DC-GAN model on a collection of *scary* images. However, there is no such data collection suitable for this purpose. Therefore, we employ a two-stage strategy (see Figure 2 for an illustration of the framework). The first step of our approach is training a DC-GAN model on a large-scale dataset called CelebFaces Attributes Dataset (CelebA) (Sun et al. 2014). The CelebA dataset contains 202,599 celebrity images with coarse alignment, each with 40 attribute annotations. We trained DC-GAN to generate face samples of 64 × 64 pixels. Figure 2 (denoted with *Original Faces*) shows a list of randomly generated faces from the trained DC-GAN model.

The second step is turning *normal* faces into *scary* images using neural style transfer (Gatys, Ecker, and Bethge 2015). Neural style transfer is an optimization method that mixes the content and style representations from two different images. The key observation the style transfer method employs is that the representations of content and style in the CNNs are separable and both representations can be manipulated in order to create new images. It synthesises a new image by simultaneously matching the content representation of an image and the style representation of a style image. It takes three inputs; a source image, a content image and a style image and uses two distance functions for optimization. The first distance function describes the difference between the content of the source and the content images, and the second distance function measures the difference between the two images in terms of their style. The objective is then to transform the source image to minimize the content distance with the content image and the style distance with the style image. We used a single style image for our expderiments (see Figure 2). We fed randomly generated DC-GAN images to Neural Style Transfer model and generated a list of *modified* images to be used in our experiments. Figure 2 (denoted with *Modified Faces*) shows a selection of transformed images used in our experiments.

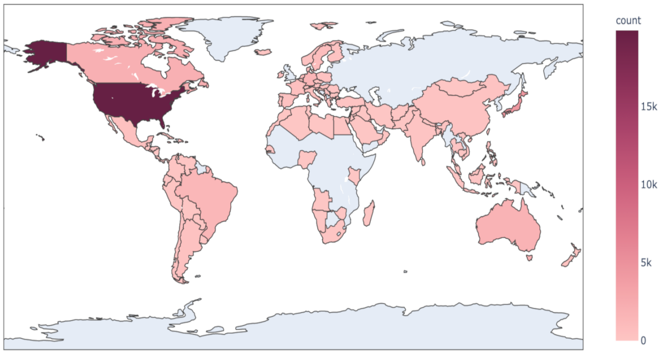

Figure 4: Distribution of the votes over countries. The majority of the votes are accumulated in United States, followed by Japan, Great Britain, Canada and Australia.

Exploratory Data Analysis

During a period of 9 months, we collected a high-volume dataset of emotional preferences with our crowd-sourcing platform `http://nightmare.mit.edu` and use it to study emotion, in particular, fear. Our dataset consist of 500 computer-generated images to vote as **Scary** or **Not Scary**. Figure 3 shows the histogram of *Scary* and *Not Scary* votes where the number of votes seems to be distributed equally with a slight shift on the right towards *Scary* votes. Our dataset consists of votes collected from 147 countries. Figure 4 illustrates the distribution of the collected votes over the globe. We can see that the majority of the votes are focused on United States, followed by Japan, Great Britain, Canada, and Australia.

Using this large-scale dataset, we seek to answer some of the interesting questions that can be raised as follows:

- Can we learn which images are particularly scary and distinguish between Scary and Non-Scary images?
- Are there different sub-groups within images that are labeled as scary and non-scary?
- Do participants have different preferences on what is scary and what is not?
- Is there any relationship between geographic location and preferences on the scariness?

Can we learn which images are particularly scary and distinguish between Scary and Non-Scary images? We explore whether a neural network model can learn to separate between images labeled as **Scary** and **Not Scary**. We train a convolutional neural network classifier to recognize *fear* by extending VGG-16 network architecture (Simonyan and Zisserman 2014). VGG-16 is a large model designed for multi-class classification, pre-trained on ImageNet dataset (Russakovsky et al. 2015) and successfully pushed the error rate to $< 10\%$ on ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014 competition (Russakovsky et al. 2015). VGG-16 network comprises 13 convolutional layers, divided in five groups. In addition, the network has 3 fully connected layers, and 5 pooling layers. The convolutional

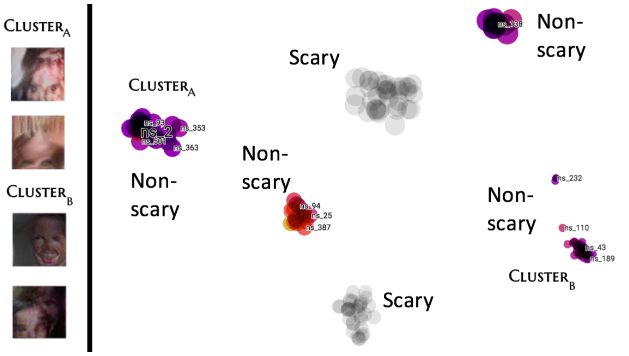

Figure 5: Embeddings of Scary (grey clusters) and Not Scary images (colored clusters) into $d = 100$ dimensional space. Closeness in the space reflects similarity. The left-most non-scary cluster ($CLUSTER_A$) consist of faces that are blurry and deformed while right-most non-scary cluster ($CLUSTER_B$) displays images that are dark.

layers consist of a set of kernels where each kernel is convolved with the input volume to compute hidden activations during the forward pass. Related parameters are updated using a back-propagation pass. We re-purpose VGG-16 network pre-trained on ImageNet by performing *fine-tuning*. In particular, we replace the final fully-connected layer of VGG-16 with two neurons corresponding to two classes, *Scary* and *Not Scary*. We curated a balanced dataset of 200 scary images and 200 non-scary images for classification. Our dataset is created as follows: we used 500 images and their corresponding votes and sorted them by the number of *Scary* (and respectively *Not Scary*) votes. We then selected the top 200 images for each category and labeled them as *Scary* (and respectively, *Not Scary*) for classification. For each class, we used 100 images for training, 30 images for validation, and 70 for testing, and used 10-fold cross-validation to evaluate the results. We obtained 65% accuracy on a balanced dataset where the baseline accuracy is 50%. This result indicates that there is some common consensus among the users about the scariness of the images, and we can distinguish between which images are *Scary* or *Not Scary* to a certain extend.

Are there different sub-groups within images that are labeled as scary and non-scary? In order to understand the relationship between *Scary* and *Not Scary* images, we built an image-embedding framework using *word-embedding* techniques. Word embedding methods recently gained popularity due to their success in accurately estimating the relationship between words in language models. We used Word2Vec method (Mikolov et al. 2013) where we treated each user profile as a *sentence* and each voted image as a *word*. Similar to how word-embedding methods capture the similarity of words by mining the co-occurrence relationship in a sentence, our aim is to capture similar images that users find *Scary* (and similarly, *Not Scary*). We used Word2Vec tool and embed images into d-dimensional space where $d = 100$. Figure 5 shows embeddings of 500 images where closeness in the space reflects similarity. As can be seen from the figure, scary images (grey clusters) and non-scary images (colored clusters) are clustered consistently. An interesting observation is that even though scary clusters are relatively close to each other, non-scary clusters are spread through the latent space. This indicates that while there is a common consensus on the groups of images that are labeled as *Scary*, there are particular characteristics of non-scary images that affect different users which results in many separate clusters are formed in the latent space. For instance, while the left-most non-scary cluster (labeled as $CLUSTER_A$) consists of faces that are blurry and deformed, the right-most non-scary cluster (labeled as $CLUSTER_B$) displays images that are mostly dark.

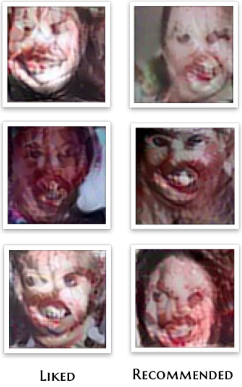

Figure 6: Images on the left shows a set of images user found scary, and images on the right are the recommended images from the model.

Do participants have different preferences on what is scary and what is not? We explore whether users have common preferences over the images by using a collaborative-filtering approach. Collaborative filtering is a mechanism that learns user preferences towards items by mining implicit or explicit interests a user expresses using rating information (e.g. books, movies, or products). The expressed ratings of users are matched against other users and a *hidden* representation of user preferences is learned. This information is then utilized to find people with the most similar preferences and to recommend items that similar users liked. A popular collaborative filtering method is matrix factorization which learns a model from incomplete rating data. We built a user-item matrix M of size $m \times n$ where m is number of users who visited our system, and n is number of images available for users to vote. Let i represent an arbitrary user who visited our system, j represents an arbitrary image user i voted. Then M_{ij} corresponds to the rating user i expressed. In particular, we interpret **Scary** votes as a rating of $+1$ and **Not Scary** votes as a rating of -1. If user i has not been shown image j, then $M_{ij} = 0$ indicating an unobserved rating. The goal is then to approximate incomplete matrix M by using matrix factorization. We used LIB-PMF (Yu et al. 2012), an efficient and parallelizable method for matrix factorization in large-scale rec-

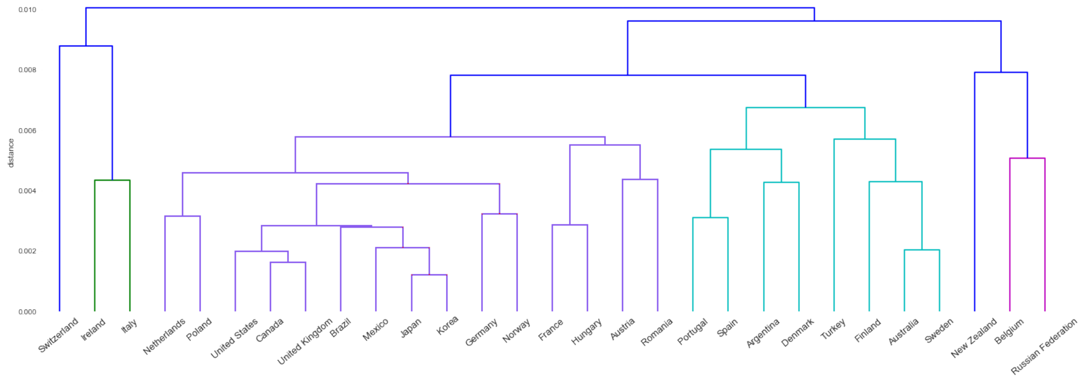

Figure 7: Dendrogram showing the hierarchical clustering on probability distributions of top 30 countries based on the voting information. One can observe that neighbor countries such as *United States* and *Canada* or *Japan* and *Korea* have the most similar preferences over the images.

ommender systems. We curated a dataset for a randomly selected 10,000 users and 500 face images. We split the data into training and testing set with a 70% to 30% rate, respectively. In total, the training dataset contains 199,436 ratings, and testing dataset contains 85,419 ratings. The goal is then to learn user preferences on the training dataset, and then to predict the *unseen* ratings in the testing set. We used Root Mean Square Error (RMSE) for evaluating the results. RMSE measures the average magnitude of the error, and it is defined as the square root of the average of squared differences between predictions and actual observations. On 85,419 ratings, we obtained a RMSE of 0.93 for the distribution of actual and predicted ratings. Figure 6 shows a qualitative example of our recommender system where top images a random user *liked* (i.e. rated as **Scary**) and top recommended images that user *might like* (i.e. might find **Scary**) are shown.

Is there any relationship between geographic location and preferences on the scariness? We explore how preferences on the votes change based on geographic location, and investigate whether there are similar preferences over the images based on geographic proximity. We used the votes collected from 90,596 visitors from 147 countries. Each country is represented by a normalized vector that represents the percentage of votes their users expressed on *Scary* and *Not Scary* images.

We perform hierarchical clustering on the distributions of the countries. Figure 7 illustrates the dendrogram for top 30 countries based on the number of users. The cophenetic distance between each observation in the hierarchical clustering was found as 0.73. Some interesting cross-cultural trends can be observed from the clustering. For instance, neighbor countries *United States* and *Canada* are the most similar to each other in terms of preferences on the faces. A similar trend can be observed between neighbor countries *Japan* and *Korea* as well as *Portugal* and *Spain*. These observations indicate that the preferences over fear have a relationship with geographic proximity. Moreover, we used a heatmap-based approach to investigate the pairwise similarity between different countries. We used the top 10 images voted as *Scary* and represented each country as a 10-dimensional vector that includes the normalized voting information. Figure 8 shows the heatmap based on top 30 countries on top 10 *Scary* images. We can observe that while most of the countries have a common consensus on the majority of the images, some countries found specific images highly *Scary*, such as *Greece* and *Switzerland* on $Face\#3$ and New Zealand on $Face\#2$. A similar trend can be observed for the opposite side where some countries found some images highly non-scary such as *Chile* and *China* on $Face\#4$. Another interesting observation is that geographically close-by countries *China* and *Taiwan* both found $Face\#3$ and $Face\#4$ as highly non-scary. These observations suggest that a cross-cultural trend on fear might exist, and different countries might have different opinions on what is *Scary* or *Not*.

Validation Study

While we received hundreds of thousands of indications that our machine-generated images were indeed scary, having subjects rate something as simply 'scary' or 'not scary' does not inform us of whether or not the images themselves actually induce the psychological construct of fear. Perhaps something that is casually rated as 'scary' actually alters mood, but it is also possible that by rating something 'scary' subjects are simply indicating that the image corresponds to the conception of what a frightful image typically looks like. Do the images we generated actually – psychologically – scare people? To investigate this question, we ran a validation experiment on Amazon's Mechanical Turk that employed psychometrically validated measures of affect and anxiety.

We randomly assigned 752 subjects to three treatment arms. The first arm consisted of the ten images that received the most 'scary' votes (Scary). The second arm consisted of the ten images that received the fewest 'scary' votes

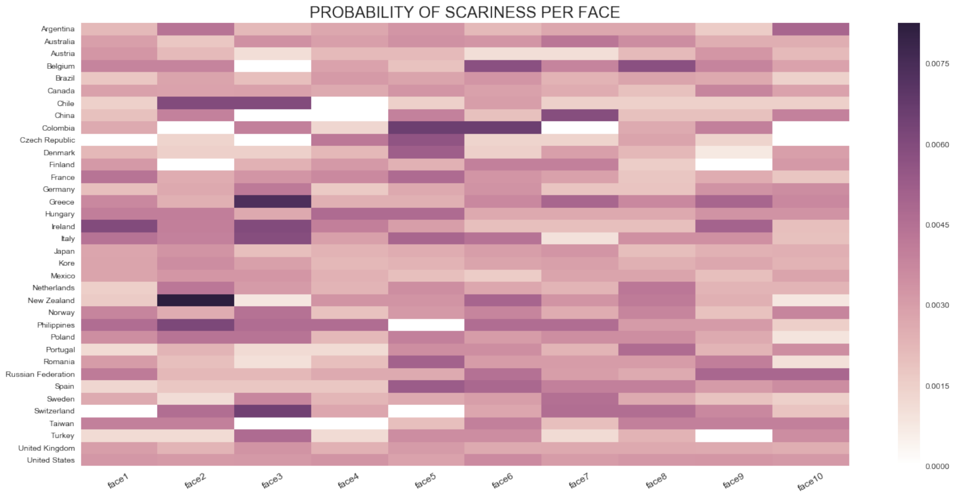

Figure 8: Heatmap of countries and their probability of finding top 10 images as scary. One can observe that while most of the countries have a common consensus on the majority of the images, some countries found specific images highly *Scary*, such as *Greece* and *Switzerland* on $Face\#3$ and New Zealand on $Face\#2$.

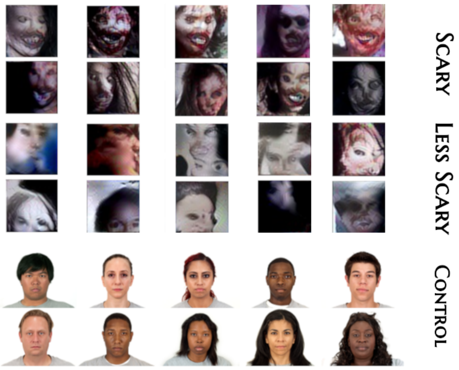

Figure 9: Images used in experimental validation. The 'Scary' faces comprise the ten faces from the platform that received the most 'scary' votes. The 'Less Scary' faces comprise the ten faces that received the fewest 'scary' votes. The ten 'Control' faces were randomly drawn from the Chicago Face Database (Ma, Correll, and Wittenbrink 2015) set of neutral faces.

(Less Scary). For the third arm (Control), we randomly selected ten neutral expression faces from the Chicago Face Database (Ma, Correll, and Wittenbrink 2015) (see Figure 9). Because affect and anxiety vary differentially across gender (Thompson 2007), we block-randomized along with the gender of respondents for added statistical efficiency (Gerber and Green 2012).

Our validation study had two outcome measures. The first is a short form of the Positive and Negative Affect Schedule (I-PANAS-SF). The I-PANAS-SF is derived from the original twenty PANAS (Watson, Clark, and Tellegen 1988) item pool and allows us to measure – and separate – dimensions of positive and negative affect. It consists of ten items including five positive affective states: *active, determined, attentive, inspired* and *alert* and five negative affective states: *afraid, nervous, upset, hostile* and *ashamed*. Participants are asked to respond to the positive and negative states which describe their feelings. The second metric is a shortened version of the State-Trait Anxiety Inventory (STAI-SF) which measures subjects' state anxiety. It is a psychological inventory based on a 4-point Likert scale ranging from *not at all* to *very much* and consists of six items assessing the degree that patients feel *calm, tense, upset, relaxed, content* and *worried*. The scores of all items are summed to produce a total score in which higher scores are positively correlated with greater anxiety. We randomized the order that our outcome measures were presented to the subjects. Finally, we pre-registered our experiment and analysis plan with AsPredicted.org as study #3410 and we follow that analysis plan below.

Negative/Positive Affect and State Anxiety

The results of our experiment indicate that our machine-generated faces produced substantial increases in negative affect and state anxiety as well as – to a lesser degree – worsened positive affect, as compared to our control condition. Respondents in the *Scary* and *Less Scary* conditions had markedly and significantly increased scores on the state anxiety measure (STAI-SF) as compared to *Control*, as can be seen in Figure 10 panels (a) and (b). Scary STAI-SF OLS coefficient is measured as 8.059 with a t-statistic: 6.346 and Cohen's d: 0.58 while Less Scary STAI-SF OLS coefficient is measured as 6.336 with a t-statistic: 5.249 and Cohen's d: 0.48. The *Scary* and *Less Scary* conditions did not significantly differ from one another.

The generated images also produced substantial and statistically significant increases in negative affect compared to the control faces (see Figure 10 panels (c) and (d)). Scary

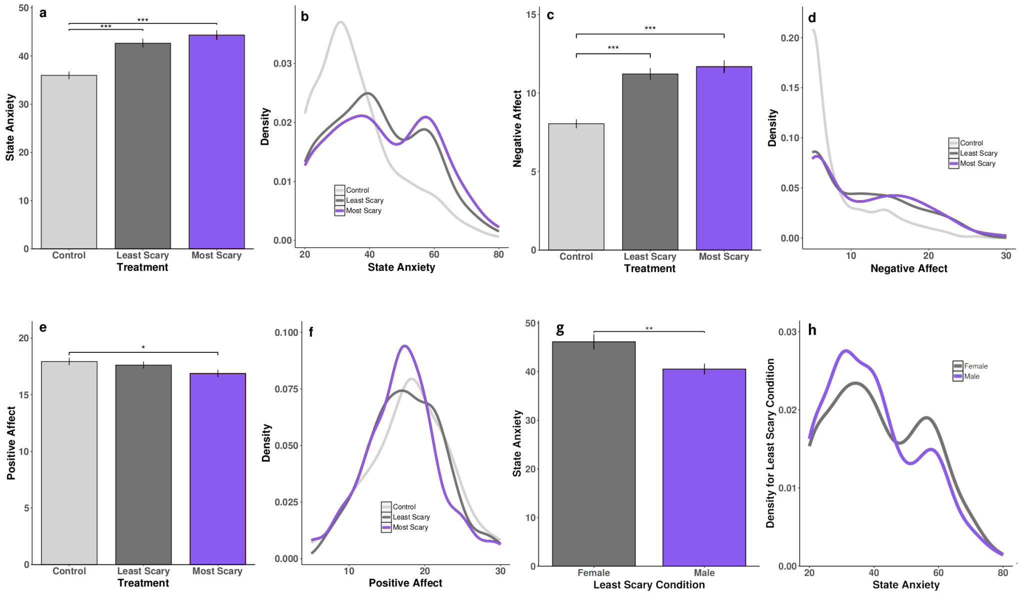

Figure 10: Validation experimental results. Scary and Less Scary faces significantly increase anxiety (panels (a) and (b)) and amplify negative affect (panels (c) and (d)) compared to the Control faces. Further, Scary faces reduce positive affect as compared to the control (panels (e) and (f)). Heterogeneous effects by gender is shown on the bottom right image. Less Scary increase anxiety significantly more in females than in males (panels (g) and (h)).

PANAS negative affect OLS coefficient is measured as 3.513 with a t-statistic: 7.156 and Cohen's d: 0.66 while Less Scary PANAS negative affect OLS coefficient is measured as 3.021 with a t-statistic: 6.412 and Cohen's d: 0.59. The Scary and Less Scary conditions again did not significantly differ from one another on this measure.

Finally, the generated faces reduced positive affect as compared to the control group, with the Less Scary faces splitting the difference between the two other groups (see Figure 10 panels (e) and (f)) Scary PANAS positive affect OLS coefficient is measured as -0.898 with a t-statistic -2.03 and Cohen's d: -0.21. The Less Scary condition did not significantly differ from the Scary or from the Control conditions.

Heterogeneous Effects by Gender

In addition to our main effects, we observe that female participants indicate significantly higher responses on the STAI scale in response to the Least Scary condition than do male participants. Note that our plan to investigate heterogeneous effects by gender was pre-registered in our AsPredicted.org plan #3410.

Male respondents in the Least Scary conditions had markedly and significantly reduced scores on the state anxiety measure (STAI-SF) as compared to female respondents in this condition, as can be seen in Figure 10 panels (g) and (h) (STAI-SF OLS male-by-least-scary interaction coefficient: -6.723, t-statistic: -2.666, Cohen's d: -0.38). We observe no other significant heterogeneous effects by respondent gender.

Ultimately, the results of our validation experiment indicate that the generated faces – for both the Scary and Less Scary images – significantly and markedly increased psychometrically validated anxiety and negative affect as compared to the Control condition. Further, female respondents in our sample exhibit greater amounts of anxiety induced by the Less Scary condition.

Conclusion

As Artificial Intelligence makes strides in solving challenging analytical problems, many people believe that an important subset of human tasks such as inducing human emotions is still far from the ability of machines to outperform humans. In this work, we challenge this hypothesis and explored the potential of deep learning and crowd-sourcing to induce extreme emotions, in particular *fear* on a case study at nightmare.mit.edu. We create a high-volume dataset of emotional preferences using crowd-sourcing and use it to study fear in a variety of applications: we showed that we can build a model that learns which images are particularly *Scary* and distinguish between *Scary* and *Not Scary* images. We showed that while there seems to be a common consensus on the groups of *Scary* images, there exist several sub-groups among *Non-Scary* images. Moreover, we showed that latent preferences of the users towards *Scary* and *Non-Scary* images can be discovered using collaborative filtering approaches, which shows the potential to tailor *personalized* images that target specific users. We also explored cross-cultural preferences for fear to observe how preferences change based on geographical location. We observed that some countries that are close to each other on a geographical level, such as America *United States*, and *Canada*, or *Japan* and *Korea* have the most similar preferences over the images. This observation suggests that there might be a cross-cultural competence over the images. In addition, while there is a global consensus on the majority of images, we observed that some countries found specific images highly *Scary* or highly *Non-Scary* which suggests that there might exist images that particularly affect certain cultures. Finally, we run a validation study where we performed a controlled experiment on $n = 752$ subjects on Amazon's Mechanical Turk where we verify whether the generated images psychologically move people on psychometrically validated measures of effect and anxiety such as I-PANAS-SF and STAI-SF. Our exploratory results and validation experiment suggests that deep learning and generative algorithms have a significant potential for inducing emotions.

As future work, our approach can be extended to other types of emotions such as *empathy*. It can further be extended to improve the performance of the image generation system by tailoring the preferences towards particular users or can be explored to understand what particular features of the generated images induce certain emotions.

References

[Barmar 2017] Barmar, E. 2017. Classifying empathy and personal distress using facial muscles activity data by applying data mining techninques.

[Botnik] Botnik. Botnik – human-machine entertainment.

[Branwen] Branwen, G. Gpt-2 neural network poetry · gwern.net.

[Brock, Donahue, and Simonyan 2018] Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *CoRR* abs/1809.11096.

[Busso et al. 2008] Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42(4):335.

[Chen and Soo 2018] Chen, P.-Y., and Soo, V.-W. 2018. Humor recognition using deep learning. In *NAACL*, 113–117.

[DeepArt] DeepArt. deepart.io - become a digital artist.

[DeepDreamGenerator] DeepDreamGenerator. Deep dream generator.

[Deng and Liu 2018] Deng, L., and Liu, Y. 2018. *Deep learning in natural language processing*. Springer.

[Deng and Yu 2014] Deng, L., and Yu, D. 2014. Deep learning: methods and applications. *Foundations and trends in signal processing* 7(3–4):197–387.

[Dosovitskiy, Tobias Springenberg, and Brox 2015] Dosovitskiy, A.; Tobias Springenberg, J.; and Brox, T. 2015. Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1538–1546.

[Gatys, Ecker, and Bethge 2015] Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2015. A neural algorithm of artistic style.

[Gerber and Green 2012] Gerber, A. S., and Green, D. P. 2012. *Field experiments: Design, analysis, and interpretation*. WW Norton.

[Gibson et al. 2016] Gibson, J.; Can, D.; Xiao, B.; Imel, Z. E.; Atkins, D. C.; Georgiou, P.; and Narayanan, S. 2016. A deep learning approach to modeling empathy in addiction counseling. *Commitment* 111:21.

[Goodfellow et al. 2014a] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014a. Generative adversarial nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Neurips*. Curran Associates, Inc. 2672–2680.

[Goodfellow et al. 2014b] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014b. Generative adversarial nets. In *Neurips*, 2672–2680.

[GPT] Ai art: I'm using a language model called gpt-2 to write my next novel - vox.

[Hemanth and Estrela 2017] Hemanth, D. J., and Estrela, V. V. 2017. *Deep learning for image processing applications*, volume 31. IOS Press.

[Hossain and Muhammad 2019] Hossain, M. S., and Muhammad, G. 2019. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion* 49:69–78.

[Jain et al. 2018] Jain, N.; Kumar, S.; Kumar, A.; Shamsolmoali, P.; and Zareapoor, M. 2018. Hybrid deep neural networks for face emotion recognition. *Pattern Recognition Letters* 115:101–106.

[Jozefowicz et al. 2016] Jozefowicz, R.; Vinyals, O.; Schuster, M.; Shazeer, N.; and Wu, Y. 2016. Exploring the limits of language modeling.

[Kahou et al. 2013] Kahou, S. E.; Pal, C.; Bouthillier, X.; Froumenty, P.; Gülçehre, Ç.; Memisevic, R.; Vincent, P.; Courville, A.; Bengio, Y.; Ferrari, R. C.; et al. 2013. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, 543–550.

[Karras et al. 2020] Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8110–8119.

[Karras, Laine, and Aila 2018] Karras, T.; Laine, S.; and Aila, T. 2018. A style-based generator architecture for generative adversarial networks. *CoRR* abs/1812.04948.

[Le et al. 2020] Le, Q.; Miralles-Pechuán, L.; Kulkarni, S.; Su, J.; and Boydell, O. 2020. An overview of deep learning in industry. *Data Analytics and AI* 65–98.

[Lee et al. 2018] Lee, H.; Tseng, H.; Huang, J.; Singh, M. K.; and Yang, M. 2018. Diverse image-to-image translation via disentangled representations. *CoRR* abs/1808.00948.

[Liu, Zheng, and Lu 2016] Liu, W.; Zheng, W.-L.; and Lu, B.-L. 2016. Emotion recognition using multimodal deep learning. In *International conference on neural information processing*, 521–529. Springer.

[Luckow et al. 2016] Luckow, A.; Cook, M.; Ashcraft, N.; Weill, E.; Djerekarov, E.; and Vorster, B. 2016. Deep learning in the automotive industry: Applications and tools. In *2016 IEEE International Conference on Big Data (Big Data)*, 3759–3768. IEEE.

[Ma, Correll, and Wittenbrink 2015] Ma, D. S.; Correll, J.; and Wittenbrink, B. 2015. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* 47(4):1122–1135.

[Magenta] Magenta. Magenta.

[Marteau and Bekker 1992] Marteau, T. M., and Bekker, H. 1992. The development of a six-item short-form of the state scale of the spielberger state—trait anxiety inventory (STAI). *British Journal of Clinical Psychology* 31(3):301–306.

[McQuiggan and Lester 2006] McQuiggan, S. W., and Lester, J. C. 2006. Learning empathy: a data-driven framework for modeling empathetic companion agents. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, 961–968.

[McQuiggan et al. 2008] McQuiggan, S. W.; Robison, J. L.; Phillips, R.; and Lester, J. C. 2008. Modeling parallel and reactive empathy in virtual agents: an inductive approach. In *AAMAS (1)*, 167–174. Citeseer.

[Mikolov et al. 2013] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Neurips*, 3111–3119.

[Ng et al. 2015] Ng, H.-W.; Nguyen, V. D.; Vonikakis, V.; and Winkler, S. 2015. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 443–449.

[Oliveira 2017] Oliveira, H. G. 2017. O poeta artificial 2.0: Increasing meaningfulness in a poetry generation twitter bot. In *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*, 11–20.

[Picard 2000] Picard, R. W. 2000. *Affective computing*. MIT press.

[Pierson and Gashler 2017] Pierson, H. A., and Gashler, M. S. 2017. Deep learning in robotics: a review of recent research. *Advanced Robotics* 31(16):821–835.

[Radford et al. 2019a] Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019a. Language models are unsupervised multitask learners.

[Radford et al. 2019b] Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019b. Language models are unsupervised multitask learners.

[Radford, Metz, and Chintala 2015] Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks.

[Russakovsky et al. 2015] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

[Samuel 1967] Samuel, A. L. 1967. Some studies in machine learning using the game of checkers. ii—recent progress. *IBM Journal of research and development* 11(6):601–617.

[Satt, Rozenberg, and Hoory 2017] Satt, A.; Rozenberg, S.; and Hoory, R. 2017. Efficient emotion recognition from speech using deep learning on spectrograms. In *Interspeech*, 1089–1093.

[Schifanella et al. 2016] Schifanella, R.; de Juan, P.; Tetreault, J.; and Cao, L. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, 1136–1145.

[Silver et al. 2016] Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature* 529(7587):484–489.

[Silver et al. 2017] Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm.

[Simonyan and Zisserman 2014] Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition.

[Sun et al. 2014] Sun, Y.; Chen, Y.; Wang, X.; and Tang, X. 2014. Deep learning face representation by joint identification-verification. In *Neurips*, 1988–1996.

[Thompson 2007] Thompson, E. R. 2007. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of cross-cultural psychology* 38(2):227–242.

[Vinyals et al. 2017] Vinyals, O.; Ewalds, T.; Bartunov, S.; Georgiev, P.; Vezhnevets, A. S.; Yeo, M.; Makhzani, A.; Küttler, H.; Agapiou, J.; Schrittwieser, J.; et al. 2017. Starcraft ii: A new challenge for reinforcement learning.

[Voulodimos et al. 2018] Voulodimos, A.; Doulamis, N.; Doulamis, A.; and Protopapadakis, E. 2018. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience* 2018.

[Watson, Clark, and Tellegen 1988] Watson, D.; Clark, L. A.; and Tellegen, A. 1988. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology* 54(6):1063.

[Watson 2014] Watson, I. 2014. Ibm watson: How it works.

[Yanardag, Cebrian, and Rahwan 2021] Yanardag, P.; Cebrian, M.; and Rahwan, I. 2021. Shelley: A crowd-sourced collaborative horror writer. In *Creativity and Cognition*, 1–8.

[Yu et al. 2012] Yu, H.-F.; Hsieh, C.-J.; Si, S.; and Dhillon, I. S. 2012. Scalable coordinate descent approaches to parallel matrix factorization for recommender systems. In *ICDM*.