Mental models of AI-based systems: User predictions and explanations of image classification results

Nathan Bos[1], Kimberly Glasgow[1], John Gersh[1], Isaiah Harbison[1], Celeste Lyn Paul[2]
[1] Johns Hopkins University Applied Physics Laboratory
[2] U.S. Department of Defense

Humans should be able work more effectively with artificial intelligence-based systems when they can predict likely failures and form useful mental models of how the systems work. We conducted a study of human's mental models of artificial intelligence systems focusing on a high-performing image classifier. Participants viewed individual labeled images in one of three general classes and then tried to predict whether the classifier would label it correctly. Participants were able to begin performing this task at levels much better than chance, 69% correct. However, after 137 trials with feedback, their performance improved a small, but statistically significant amount to 73%. Analysis of these results and comments indicated that humans were using their own perceptions of the images as first-approximation proxies. 'Projecting' human characteristics onto a computer might be considered a cognitive bias, but in this task, the strategy seemed to yield good initial results. This might be called *effective anthropomorphism*. Participants sometimes used this strategy both implicitly and explicitly. The paper includes discussion of why this strategy might have worked better than alternatives, why further learning was quite difficult, and what assumptions about similarities between human perception and image classification systems may in fact be correct.

INTRODUCTION

Effective human/machine collaboration is increasingly important for many types of work. Humans can more effectively supervise and use intelligent systems when they better understand how the systems work. Current systems, especially in the machine learning realm, are growing increasingly powerful but at the same time becoming increasingly opaque.

This research asks the question: How do people try to understand and predict the behavior of a complex artificial intelligence (AI)-based system? This research uses a machine-learning-based image classifier, which is a relatively familiar and approachable AI system. Image classification has greatly improved in the last ten years due to the development of convolutional neural networks and applications such as Google's Inception (Szegedy et al., 2014). This type of system starts with a structured but untrained set of neural networks and learns via presentation of thousands of labeled images. Because the system is very complex and has no human-like semantic preconceptions or other knowledge of the world, it can be very difficult to know how a trained network comes to its conclusions.

Complete understanding of such systems by a person is not possible. Even their designers, who understand how the original networks were trained, can only infer how the system will respond to a novel image. There is an ongoing line of research on how to help humans make sense of these systems (see Olah et al., 2018 for a brief review). Attribution systems, for example, try to identify and visually highlight the most important pixels in a given classification, or similarly identify important training set images. Feature visualization creates millions of variations on an abstract image to understand what elicits the greatest response from a particular node or layer of a classifier; this has been valuable in understanding feature detection (Olah et al., 2017).

This research takes a different approach, and asks whether unaided, untrained users can construct useful mental models to understand and predict the performance of AI-based systems. Humans have impressive powers of inference and highly developed perceptual systems. With enough experience, they might be able to predict performance without fully understanding the underlying mechanism. There are open questions about whether this is possible, how much experience would be necessary, and what kinds of analysis and feedback would optimally support this learning.

Another set of questions relates to the contents of human mental models. For our purposes, a mental model is a cognitive representation of some aspect of the world used to predict events in or aspects of a person's world (Khemlani & Johnson-Laird, 2011; Rouse & Morris, 1986). What users understand, or think they understand, about image classifiers is largely unknown. There has been little or no prior research on human mental models of image classifiers, though prior research has examined mental models of other complex engineered systems, some of which have involved machine learning (Tullio et al., 2007; Talone et al., 2016; Yang et al., 2018).

Multi-cue Probability Learning

Because image classifier perception is different in unknown ways from human perception, one might argue that humans should approach this task by discarding all assumptions based on their own perceptual abilities or tendencies. This would avoid potential projection bias. But how would one go about this? Without advanced diagnostic tools, the problem falls in a class of phenomena called multi-cue probability learning problem (MCPL). An MCPL task requires associating arbitrary cues to outcomes based on observed relationships. In an image classification task, a user might observe over time that a classifier is good at recognizing cats, and associate that cue (cat object class) with that outcome (success). Over time, many such associations may improve predictions.

In one influential MCPL study (Gluck, 2002), researchers identified common strategies. Participants tried to predict an outcome, the weather, based on the presence or absence of four cards over a sequence of 200 cards/weather pairings. A

perfect prediction could be made on the basis of the card information; participants on average achieved about 70% accuracy by the final block. However, many participants discarded most of the available information. The most common strategy in the final block was a one-cue learning strategy where all cards were ignored save one. There was also a 'singleton' strategy where single cards were associated with single outcomes. Less than half of participants used more complex multi-cue strategies (<40%).

The image classification task described in this research would involve many more possible cues, however. The classifier, which has no semantic understanding of tigers, might be associating the *tiger* label with any combination of colors, patterns, shapes, contrasts, or other features. In the experimental paradigm used here, humans are asked to start making predictions after exposure to many fewer images (~40) than would be necessary to determine with any certainty which cues and which associations were relevant.

Effective anthropomorphization

Anthropomorphism is the tendency to assign humanlike characteristics, motivations, intentions, or emotions to nonhuman agents (eply). It has often been evaluated negatively, as an error in thinking or a sign of immaturity, although some positive aspects have been postulated, including in contexts of human-machine interaction (daimano).

Analysis of comments from our pilot testing indicated that humans in this task used their own perceptual system as a basis for initial estimates. Thus, they engaged in *effective anthropomorphization*. If the person found the object in the image to be blurred, too small, or hard to pick out from the background, they were likely estimate it would be more difficult for the computer as well. This is a reasonable strategy when little information about the system is available. This strategy may lead to systematic mistakes, however, as there are aspects of human visual perception that may not have an equivalent in a self-trained computer algorithm. Features of human perception that may be relevant for image classifiers are:

Prototypes and features. Humans often understand the world as hierarchical classes of objects, (collie | dog | animal) with some objects being more prototypical of each type than others. Perception often focuses on visually distinct features that are prototypical, even though they may not be universal. Examples would be a dog's floppy ears, a bee's fuzzy striped abdomen, or a French horn's distinctive shape when viewed from a side angle.

Figure/ground separation. Humans tend to mentally separate a scene into foreground objects distinct from the background. Interestingly, there is some cultural variation in the tendency to separate foreground from background, but the basic tendency remains.

Simultaneous top down and bottom up processing. Human perception is informed by knowledge of the world as well as processing of visible images. Processing is said to proceed 'bottom up', from raw images to processed data, and in parallel meaning is constructed 'top-down' as perception is guided and constrained by prior knowledge.

Research questions:

1. Can untrained humans improve their predictions of image classifier success after one session of practice and informative feedback?
2. What mental models of image classifiers do humans construct to help with this predictive task?
3. Is anthropomorphism used in these models?

METHODS

We pursued these questions with an online experiment in which participants tried to predict whether an image classification algorithm would correctly label images. The first question was addressed by a quantitative analysis of subjects' prediction success, the second and third by a qualitative analysis of comments made by subjects during the prediction task.

Image Classification Model. The image classification model used ImageNet data with Google's Inception v3 pretrained model on TensorFlow. Images were normalized to a standard size and format, and four modes of images were generated using R Magick: original, grayscale, low contrast, and paint-filter. Each image mode was only used once in the experiment. Inception can label 1,000 classes of objects. We selected a subset of 13 object types in three categories to reduce task complexity: cats (tiger, leopard, Persian cat, Siamese cat); insects (bee, ant, grasshopper, ladybug, monarch butterfly); and, musical instruments (violin, flute, harp, saxophone). Inception considered all known classes, and any of the other 987 classes would be labelled as 'incorrect'. Three blocks of images (balanced for our four image modes) were constructed, containing 28 (70%) correct and 12 incorrect images per block. After piloting, one problematic image was removed from each block leaving 39 images per block. Problematic images were ones where participants legitimately questioned the correctness of the image label. In two of these blocks the removed image was correct, leaving a correct percentage of 69%, 71% and 71%. Blocks were counterbalanced between subjects.

Experiment. Participants viewed and made predictions for 117 images across three blocks of 39. During the experiment, we referred to the classification system as 'Cee1'. Participants were not provided any information about Cee1's overall accuracy. For each image, participants were shown an image with a label and asked to make a Yes/No prediction if Cee1 would correctly label the image. They were also asked to predict the percentage confidence score Cee1 had in its prediction, and provide comments (optional) explaining their rationale. Cee1 was 'correct' if the predicted label had the highest ranked confidence. The first block was a pretest and participants received no feedback as to if their predictions were correct. At the end of the first block, participants could view a web page with all of the images and their top five assigned labels. In the second and third blocks, participants were shown Cee1's top five assigned labels with confidence scores (see Fig. 1).

Figure 1. Image, predicted label, and top five Inception labels with confidence scores.

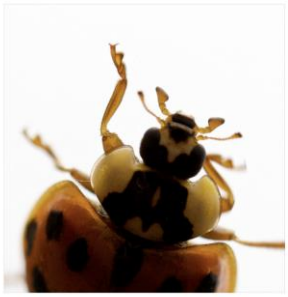
Image and Correct Label	Rank	Classifier Labels	Confidence
 ladybug	1	ladybug	65.63
	2	chocolate sauce	4.83
	3	ant	3.18
	4	bee	2.86
	5	ice cream	2.77

Figure 2. Example of a difficult prediction

Image and Correct Label	Rank	Classifier Labels	Confidence
 ladybug	1	nail	42.82
	2	picket fence	8.86
	3	birdhouse	8.3
	4	ladybug	2.57
	5	padlock	1.99

Participants. Seventy-nine Mechanical Turk participants were recruited to perform this task. Participants were located in the U.S. and high-performing (95% or higher acceptance rate for previous work). They were paid $12 for completing this task, with a $4 bonus given as an incentive to participants who were in the top 30% on the final experiment block.

Outcomes scoring. Inception was considered to be correct only when the correct image label was the top ranked label. Participants were scored as to whether they correctly predicted Inception's classification performance (correct or incorrect). Participant estimates of Inception's confidence were also collected, but are not analyzed in this paper.

Comment analysis. A coding scheme was used to categorize and understand comments users made while performing the image classification prediction task. Participants made 5263 written comments over the course of the task. We developed the scheme using a grounded theory approach based on analysis of a prior think-aloud study of subjects performing the same task and on responses from a pilot version of the current study. Codes are listed in Table 1. Of the 5263 annotated comments, both raters coded 800 randomly chosen comments. Interrater reliability is reported below. In all of the cases where raters disagreed, the ratings were discussed until consensus was achieved.

RESULTS AND DISCUSSION

Good initial predictions

Participants' predictions in the pre-test blocks were much better than chance. With no feedback and minimal training, participants accurately predicted the classifier's accuracy 69.2% (.083) of the time. The closeness of this number to the classifier's accuracy rate of 70% is, as far as we can tell, coincidental. The high accuracy rate suggests that whatever naive model the participants brought to the experiment was a good first approximation.

Small improvements in predictions

Participants made small improvements in their ability to predict classifier accuracy over the three counterbalanced blocks. The average prediction score over the three counterbalance blocks were 69.2% (.083), 71.6% (.058) and 73.2% (.066). We conducted a multilevel regression on correct predictions controlling for individual as a random factor and determined that this small increase was statistically significant at $p<.001$ level. We conducted this analysis using the LME4 package in R, conducting one analysis with only participants as a random predictor, a second with participants plus set number included as an integer (1-3), and performing an ANOVA of the difference.

Minor individual differences

As a follow-up analysis, we wanted to determine whether we could identify better and worse performers, as a first step to identifying expert strategies. As an initial test of whether there were reliable individual differences, we calculated a correlation between individuals' accuracy in the test blocks they viewed second and third. Individuals who had found important patterns should score higher in both, and poor-performing individuals should score lower in both, creating an overall positive correlation. The result was a non-significant positive correlation of -.13. We determined that it would not be meaningful to separately analyze high and low scorers.

Task difficulty

Overall, the level of learning observed was less than anticipated. The study was originally planned as a comparison of learning from different types of feedback. The learning effects were small enough in what was intended as the high feedback condition that we did not run the intended comparisons with less feedback. Instead, analysis focused on understanding users' mental models formed in this difficult task.

Examination of the stimuli illustrates why this was such a difficult task. Figure 2 is one of several examples of a stimuli where every participant guessed incorrectly, (as the research team would have).

All participants predicted that Inception would correctly identify the ladybug in this photo. Human observers can separate the red ladybug object from the background and consider this a relatively easy classification. Instead, the highest-ranked label was 'nail'. It is possible that the large black vertical shape, which humans recognize as a shadow, was what drew the label of nail. It is possible that the textured surface with dark lines drew the label; this texture may have contributed to the 'picket fence' label. Or perhaps this mis-categorization has some other primary reason, has many reasons, or defies easy human-understandable explanation. Listing the top five labels

Table 1. Comment codes

Code	Description	Frequency Observed	Percent agreement and Cohen's Kappa	Example Comment
Object	Comments about the object to be classified, e.g., its features, place in a hierarchy, or distance from a prototype	205	91.94%, .66	*The dots will make it easy to tell that this is a leopard.*
Scene	Comments about the composition of the scene in the image e.g., field of view, aspect, number of objects	243	92.74%, .72	*The bottom of the sax is cut off*
Image	Comments about the image quality, e.g., contrast, monochrome/color, resolution	208	98.39%, .92	*It's dark, not enough contrast*
System	Comments about the system's ability to classify the image	281	91.13%, .68	*I think the computer will get it right, but won't be very confident in it.*
Overall Impression	Comments about general properties of the image not otherwise coded	176	90.32, .52	*It is a clear picture with nothing else in the picture.*
Human	Comments about human ability to classify the image, either in general or relative to the subject making the comment	39	99.19, .80	*Took me a sec to see the bee, honestly.*
Mental Model	Comments about the state of the subject's mental model of the system, e.g. what properties of the object or image the classifier used or about the subject's confidence in his or her mental model	23	99.19%, 0.00	*Not sure how well it will be identified as a monarch without color*
Ground truth	Comments about the accuracy of the "true" label presented as the target for the classifier	14	97.58%, -.01	*I actually do not think that is a ladybug, so it won't be number 1*
Context	Comments about the human context of the scene in the image, e.g., object function, relative size, relationships among objects in the scene	6	100%, undefined	*Is this the same as the ladybug? They aren't in their natural environment.*

provided the user some useful feedback with which to build a mental model.

Qualitative aspects of user mental models

The rest of this analysis focuses on qualitative study of user comments, to understand the emerging mental models of participants doing this task. The codes clearly separated into two tiers, frequently used (Object, Scene, Image, System, and Overall Impression) and infrequently used. Participants made comments across multiple categories, showing a diversity of observations and strategies. Of 62 participants with at least one comment in the random sample, 58 used more than one code type, and 49 used four or more.

Object. A large percentage of the comments referred to the object in the photo. The presence of comments in this category was expected, not only because object recognition is a prominent aspect of human perception, but also because the task itself primed awareness of the central objects. Comments also reflected other expected aspects of human perception. Many comments referred to presence or absences of what the participant believed to be prototypical features of the object, such as the shape of the violin or the presence of wings, e.g., "*Not enough sax[ophone] elements to ID*". Many other comments implied a comparison to an object prototype, either a match, *"clearly a Siamese cat"*, or the potential for mismatches, "*This looks like a lot of different instruments."*

Scene. The large number of comments on the scene often reflected aspects of human perception as well. Many comments referred to the difficulty of separating objects from background; "*Sea of green in the image. Grasshopper blends in.*" The presence of multiple objects, in the same or different classes was expected to cause confusion. Participants also commented on camera angle and distance.

Image. Image comments often corresponded to the image versions that were created using the black and white, low-contrast, and paint-filter. Low-contrast images were often noted as 'dark'. Interestingly, none of the comments referred to a paint filter, even though many people have seen this applied. Image property comments on these were 'blurry' or 'pixellated.' There were also many comments on the black and white image property, e.g. "*I think the lack of color might affect its labeling ability.*" These image conditions were noted at a frequency similar to object and scene comments.

System. These comments referred to the classifier that was the focus of the study. This was important because it showed some separation between capabilities of the individual's perceptual system and of the system's in the subject's mental model. Comments often contrasted what the human saw with what the system might perceive: "*Easy to see ladybug but could label leaf.*" Participants sometimes commented on their evolving system model, e.g. "*It has done pretty consistently well with the insect if the image is good.*" Most of the time (69%) the system code was used in conjunction with one or more other codes.

Human. Reference to human abilities seemed to be a sign of sophistication; the participant seems to recognize that human perception is relevant but only as a proxy: "*It's very difficult to see that this is a saxophone. Even I can't really tell that it's a saxophone, and I've seen them countless times.*"

Mental model. Participants sometimes referred to their own evolving mental model of the classifier. "*Based on shape and pattern, but I'm starting to wonder how much of a role the color plays.*" The Human and Mental Model codes were used infrequently, but this should not be taken as evidence that the participants were not aware of their own mental models.

Ground truth. Comments in this category questioned the veracity of the image labeling: "*You can't see the complete instrument and it may be labeled wrong.*" Participants rarely did this. Images with questionable labeling were mostly avoided during stimuli selection, so there might have been more of these comments in a randomly selected image set.

Context. Context comments were about the semantic context of the photo, e.g. "*Is this the same as the ladybug. They aren't in their natural environment.*" This was the least frequently observed code. Generally, these comments would be questionable, since the classifier has no background

knowledge of the world. Infrequent use of this kind of reasoning might be a mark of sophistication.

DISCUSSION AND CONCLUSION

Humans can work more effectively with powerful AI systems when they can predict likely failures and form useful mental models of how the systems work. To study this, we asked participants to try to predict the outcomes from a high-performing but opaque image classifier. This was a short task (137 images total) and participants had to form their mental models quickly, without being able to test a large number of hypotheses about what cues the classifier may be using.

Subjects' initial guesses about how well the classifiers would do were much better than chance. Participants guessed correctly on 69% of image classifications without any training or feedback. Their initial guesses seemed to be based at least partially on their own perceptions, discussed below.

Participants were not, however, able to substantially improve their predictions of the image classifier's performance over the 137-item trial. There was a small but statistically significant learning effect across the three sets, as mean correctness went from 69% in the pre-test to 73% in the third block.

Implicit and explicit anthropomorphism

Insights into participant thought processes were gained by reading and categorizing comments they wrote while doing the task using a simple one-level coding scheme (Table 2).

We argue that participants often used *effective anthropomorphism* to make their guesses. Participants used their own visual system as a proxy for the image classifier, to good effect. Comments frequently included comments about clarity, contrast and visibility. Comments often implied human-like concepts of prototypes, hierarchical classifications of objects, and key features of objects.

There appeared to be both *implicit* and *explicit anthropomorphism* at work. Implicit anthropomorphism would mean that participants applied their own perceptual abilities to the AI without necessarily realizing they were doing so. But there is evidence that many participants also used explicit anthropomorphism. Participants talked about the system, frequently. They also discussed their own or general human abilities, indicating some cognitive separation of human and classifier abilities. The 'mental model' tag indicated awareness that participants were forming a mental model of the system as they did the task.

These conclusions are somewhat tentative. This study was not originally designed to look at anthropomorphism; it was designed to study learning. The secondary analysis of comments was interesting enough, however, to suggest some ideas about mental model formation and anthropomorphism. These observations should be confirmed with more targeted studies of the phenomena.

Some anthropomorphic assumptions turn out to be correct

Many assumptions participants made about image classifiers may be accurate, if not now then in future generations of these systems. There is a fascinating current line of research into how convolutional neural network-based image classifiers do what they do. The systems are designed in layers that represent sequential processing steps. Early layers identify specific features of objects, such as 'edge detectors', which in later layers are combined into holistic labels (Nguyen et al., 2016). There is evidence that some image classifiers form abstracted prototypes of object classes, and that these prototypes may have some hierarchical characteristics, (Alsallakh et al., 2018). There is evidence that image classifiers make use of context in interpreting some images, although this usage is (likely) less sophisticated than human top-down processing abilities. It is an interesting question how far these similarities will go as classifiers become more complex.

Participants' anthropomorphism may also have been effective because of the actual task that the classifier was trained to do. While we call many such systems image classifiers, they are not trained to compare objects in images to some platonic ideals. Training data for these systems are typically drawn from large sets of human-labeled images. In other words, the classifier has been trained to put the same labels on images as people do. Whether this aspect of the machine's task relates to anthropomorphic assumptions could be another question for further research.

Future work

Future work should more directly test the hypothesis about implicit and explicit anthropomorphism versus other possible strategies. will include exploring how participants operationalize their mental models through selecting new images to retrain an existing image classifier to improve its performance. We will examine consistency between participant-articulated mental models and image selection, overlap of image sets and models across participants, and improvements in model performance.

REFERENCES

Alsallakh, B., Jourabloo, A., Ye, M., Liu, X., & Ren, L. (2018). Do Convolutional Neural Networks Learn Class Hierarchy? *IEEE Transactions on Visualization and Computer Graphics*, *24*(1), 152–162.

Damiano, L. & Dumouchel, P. (2018). Anthropomorphism in human-robot co-evolution. Frontiers in Psychology 9, article 468.

Epley, N., Waytz, A. & Cacioppo, J.T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism. Psychological Review, 114 (4), 865-886.

Gluck, M. A. (2002). How do People Solve the "Weather Prediction" Task?: Individual Variability in Strategies for Probabilistic Category Learning. *Learning & Memory*, *9*(6), 408–418.

Khemlani, S. S., & Johnson-Laird, P. N. (2011). The need to explain. *Quarterly Journal of Experimental Psychology*, *64*(11), 2276–2288.

Olah, et al., (2018). The Building Blocks of Interpretability. Distill, 2018.

Olah, et al., (2017). Feature Visualization. Distill, 2018.

Rouse, W. B., & Morris, N. M. (1986). On Looking Into the Black-Box - Prospects and Limits in the Search for Mental Models. Psychological Bulletin, 100(3), 349–363.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2014). Going Deeper with Convolutions. *ArXiv:1409.4842*.

Talone, A. B., Phillips, E., Ososky, S., & Jentsch, F. (2016). An Evaluation of Human Mental Models of Tactical Robot Movement. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*(1), 1558

Tullio, J., Dey, A. K., Chalecki, J., & Fogarty, J. (2007). How it works: A field study of non-technical users interacting with an intelligent system Conference on Human Factors in Computing Systems. 31–40.

Yang, Q., Suh, J., Chen, N.-C., & Ramos, G. (2018). Grounding Interactive Machine Learning Tool Design in How Non-Experts Actually Build Models. *DIS 2018* (pp. 573–584). New York, New York, USA: ACM.

Zhang, Y., Larlus, D., & Perronnin, F. (2014). What makes an Image Iconic? A Fine-Grained Case Study. *ArXiv: 1408:4325*.