Two Perils of Binary Categorization: Why the Study of Concepts Can't Afford True/False Testing

Greg Jensen[1] **and Drew Altschul**[2]

[1]**Department of Psychology, Columbia University, New York , NY, USA**
[2]**Department of Psychology, University of Edinburgh, Edinburgh, United Kingdom**

Keywords: Concepts, Categorization, Machine Learning, Animal Cognition, Comparative Cognition

INTRODUCTION

Many claims about concept learning in animals rely on binary categorization tasks (Herrnstein et al., 1976; Freedman et al., 2001; Marsh and MacDonald, 2008). When subjects exceed chance levels of performance, they are alleged to have learned "the concept." Critics are quick to point out that although subjects have learned *something*, confounds may explain performance more simply (Katz et al., 2007; Wright and Lickteig, 2010; Zentall et al., 2014). Despite a growing literature on both sides, supporters of "concept learning in animals" seem no closer to persuading the skeptics, while skeptics are no closer to persuading proponents. This rift hinges on disagreements over the strength of the evidence.

Results from dichotomous classification procedures are inescapably ambiguous: They represent the weakest possible evidence for concepts in animals, for reasons unrelated to the validity of the theories they are aim to test. One shortcoming of this approach is the *tailor-made classifier*, which may arise during training. Effectively, "teaching to the test" undermines claims about animals' general knowledge. Another shortcoming is the *lucky guess*, which manifests during testing. A simplistic response during the testing phase will yield many rewards due to guessing alone, making it difficult to asses the precise content of learning. These shortcomings are independent, such that either or both might confound an experiment.

The Tailor-Made Classifier

The risk of animal subjects 'outsmarting' their minders has been with us since Clever Hans. Whatever the aims of our experimental paradigms, the influence of extraneous information must be minimized so that results reflect the intended empirical test.

Concept learning presents the scrupulous researcher with a challenge: How does one identify (much less control for) the extraneous features of a stimulus? Our understanding of how the brain categorizes stimuli remains limited (Freedman and Assad, 2011), but there is also no consensus about what constitutes a *feature*. The list of stimulus attributes that might be used to categorize stimuli includes overall descriptive statistics ("presence the color green"), low-level structural details ("T-shaped edge junctions"), patterning ("the presence/absence of tiled features"), functional interpretation ("that looks like food"), ecological indicators ("bright color = poison"), and any of the potential interactions between levels (cf. Spalding and Ross, 2000; Marsh and MacDonald, 2008). As such, the content of learning is subject to multiple interpretations.

A *classifier* is an algorithm (however simple or complex) that identifies a stimulus as belonging to some discrete category. In general, classifiers must undergo training, becoming sensitive to category-relevant features by trial and error. A classifier may also have innate knowledge (such an instinct to treat some stimuli as threatening), and may be unable to detect or exploit certain features.

Herein lies the problem: The aim is to uncover general conceptual aptitudes, and if training only requires that two categories be identified, then the classifier need only identify some difference that distinguishes those two categories. The result is a tailor-made classifier: Tailored by the specifics of the binary training paradigm, and optimized solely for that dichotomous discrimination. Just as a bespoke suit is tailored upon request to fit a single person, a tailor-made classifier is only effective at the discrimination

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.688v1 | CC-BY 4.0 Open Access | rec: 12 Dec 2014, publ: 12 Dec 2014

it was trained for. This is Clever Hans in a nutshell: A (cognitively) cheap trick that yields rewards but provides little insight into a generalized form of knowledge.

When faced with this problem, researchers have frequently chosen to narrow the scope of the features available. A set of images might have colors removed, luminances matched, occluders introduced, and noise added (e.g. Basile and Hampton, 2013). Such studies are valuable because they help reveal which features can used by the classifier. However, regularized stimuli cannot rule out the possibility of a tailor-made classifier, because so many potential 'features' might provided the basis for the classification. Furthermore, insofar as the resulting stimuli are 'unnatural,' they generalize poorly to how stimuli are categorized in ecological contexts.

The Lucky Guess

Independent of the classifier (tailor-made or otherwise), the clarity of the evidence depends on how learning is eventually tested. If subjects must make dichotomous choices (e.g. 'face' vs. 'house,' or 'same' vs. 'different'), then a naive animal will be rewarded half the time. If the positions of the stimuli are counterbalanced (and they usually are, to prevent bias), then this naive animal needn't even randomize its responses; uniformly and insensitively choosing 'left' yields a steady stream of rewards.

If a subject's classifier functions even modestly, this rate of reward can be exceeded. However, it is difficult to assess what proportion of correct responses are genuine classifications and what proportion are merely *lucky guesses*. Accuracy of 70% on a binary test could mean that the subject is guessing at random more than half the time (e.g. 40% correct classifications, 30% lucky guesses, 30% unlucky guesses). If a 50% reward rate is deemed satisfactory to the subject, then responding quickly and mindlessly may prove the most favorable strategy. High guessing rates undermine the researcher's ability to make general statements about stimulus properties, particularly given the difficulty in determining which characteristics are used by the classifier.

Guessing is much less effective when tests requires more complex responses. If a subject must take a set of n stimuli and assign each to one of n categories, the odds of guessing correctly drop as n increases. This has two benefits. On the one hand, guesses are more likely to include at least one error. This improves the signal-to-noise ratio in trying to evaluate the characteristics of a subject's classifier, effectively making every correct sequence of responses more informative. On the other hand, the reward gradient will be better correlated with accuracy: Poor performance will yield far fewer rewards, providing an incentive to attend to the task and to produce high-quality responses.

A DEMONSTRATION BY SIMULATION

These two confounding factors are relevant regardless of the complexity of the classifier. In education (as in machine learning), teaching to the test yields poor general learning and T/F exams are poor measures of the depth of learning. Rather than provide a rhetorical argument based on theory, we offer a concrete simulation using the *bag-of-features* classifier (O'Hara and Draper, 2011) provided in the Computer Vision System toolbox for Matlab v2014b (Mathworks, 2014). Despite relying strictly on low-level features, this approach performs well with photographic stimuli. To represent a "cognitive limit," we limited all classifiers to no more than 100 clusters of features. The Caltech-101 image bank provided 9665 stimuli belonging to 102 categories (Fei-Fei et al., 2007). Half the images were used to train the classifier, and the other half were set aside as a novel "validation set" for testing.

Training a Tailor-Made Classifier

The ten categories in the Caltech 101 with the most images were ordered by size. The classifier was trained and subsequently validated using the first two of these categories, then the first three, and so forth up to ten. Because the classifier was limited to 100 clusters, its criteria became more general as the number of categories increased. The accuracy for each category, as well as the overall average, is plotted in Figure 1 (left).

Some image categories continue to perform well as additional categories are added: Airplanes, leopards, and 'easy' faces were categorized correctly over 85% of the time. However, other categories did less well when the classifier was forced to generalize. In particular, performance for the category of 'background images' steadily deteriorated, presumably due to the lack of consistent discrete features. If this algorithm was being studied with only three categories, however, this deficiency would not be apparent: Backgrounds were categorized correctly 90% of the time when competing with only two other

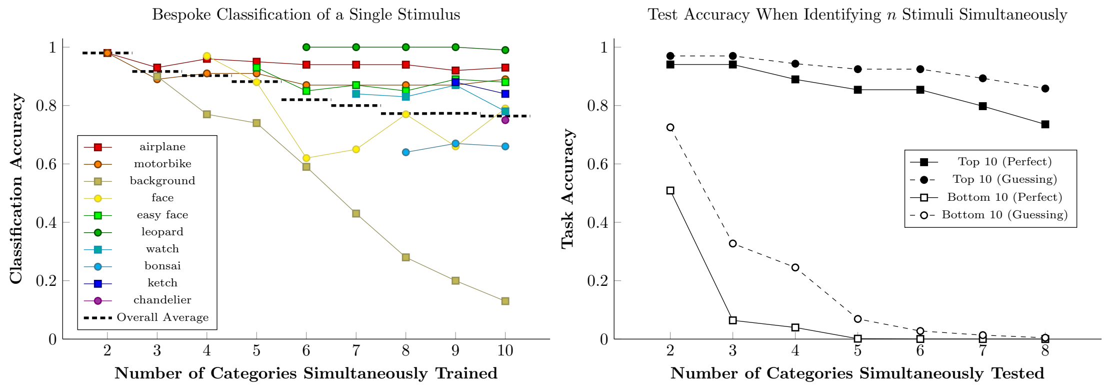

Figure 1. Performance of the bag-of-features classifier using 100 feature clusters. (Left) Classification accuracy given training on the ten largest categories in the Caltech 101 sample set. Colored lines show accuracy for specific categories, while dashed black lines show overall accuracy for each level of training complexity. (Right) Accuracy by a classifier trained on 102 categories during a test in which n stimuli must be classified correctly for a trial to be 'correct.' Performance for the classifier's ten best (black) and worst (white) categories was gauged. Solid lines indicate cases in which classification was done perfectly, while dashed lines indicate cases where correct responses required at least one guess.

possible choices. Although the bag-of-features algorithm cannot perform abstract generalizations, it can apparently identify the abstract class of 'backgrounds' by process of elimination. It is only when backgrounds compete with many other stimuli that the algorithm reveals its weakness.

Measuring the Benefits of Guessing

We retrained the classifier using all 102 image categories in parallel. The classifier was then tested using the following procedure: Exactly one novel image was drawn from each of n categories. The classifier had to match *every* image with its corresponding category. Performing entirely at random yields correct responses with a frequency of $\frac{1}{n!}$. Sometimes, classifiers judge multiple images as belonging to the same category (e.g. identifying both an picture of a bonsai and a picture of a forest as 'bonsai' rather than 'bonsai' and 'background' respectively). In these cases, our algorithm randomly assigned one of the images to the identified category, and then deduced that the remaining images must belong to whichever categories remained unaccounted for.

Figure 1 (right) shows performance when the classifier's ten highest-performing categories (black) and ten lowest-performing categories (white) were tested in this way. Solid lines show the trials in which every stimulus was correctly identified without guessing, while dashed lines show those trials that were correct given at least one guess. Although there is a clear distinction between high and low performance, there is always a benefit for guessing. In the two-item test, a poor classifier guessed its way to 72% accuracy, a level that would be considered 'high' in many published studies. In three- and four-category tests, almost every correct trial for the poor classifiers involved at least some guessing.

There is a temptation to view every 'correct response' as a case in which performance was indicative of mastery, but even a poor classifier that yields only vague hunches provides enough information for performance to exceed chance quite considerably. This 'slightly-informed guessing' is responsible for many correct responses made by a poor classifier. The best defense against guessing is to increase the complexity of the test, which makes each trial much more informative.

RECOMMENDATIONS

Our simulation demonstrates perils of narrow category training and of simplistic tests. When a subject (or an algorithm) is trained on only a handful of categories, learning is likely to overspecialize, failing to capture the classifier's general aptitudes. Similarly, when even a highly general classifier is tested on only a few categories at a time, correct trials frequently result from informed guessing rather than from robust representation of all extent categories. We demonstrate these problems separately, but the two can easily act in concert. When a study displays both confounds, it is nearly impossible to judge whether performance arises from any abstract understanding of the stimuli.

The best defense against the possibility of a tailor-made classifier is to increase the number of categories that are trained in parallel. While various clever tricks may permit pictures of faces to be distinguished from pictures of houses, such trickery is more difficult given three categories, still more difficult given four, and so on.

Many studies that trained more than two categories (e.g. Herrnstein et al., 1976; Sigala, 2009; Vonk, 2013) nevertheless tested only one or two stimuli at a time. Others have required that subjects match a stimulus to one of four categories (Bhatt et al., 1988; Lazareva et al., 2004). Although an improvement, such match-to-sample procedures reward random responses on $\frac{1}{n}$ trials, and informed guessing remains an effective approach for a poor classifier.

Contrary to the recommendations of Katz et al. (2007), we recommend that test conditions require subjects to identify more than one stimulus category during each trial. Unfortunately, few validated methods provide an appropriate level of response complexity. One candidate is the simultaneous chain (Terrace, 2005), which has been used to test serial and numerical cognition. Another candidate is the ALVIN procedure (Washburn and Gulledge, 1995), albeit adapted to make use of novel categorical stimuli. Recovering from the weaknesses of prior concept studies will require that researchers raise the bar, and give their animal subjects the opportunity to succeed (or fail) on their own cognitive merits.

DISCLOSURE/CONFLICT-OF-INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

GJ and DA both contributed the conceptualization, analysis, and writing of this piece.

ACKNOWLEDGEMENT

The authors wish to thank Herbert Terrace and Vincent Ferrera for their feedback. Support was provided by NIH grant 5-R01-MH081153-06 awarded to V. Ferrera and H. Terrace.

REFERENCES

Basile, B. M. and Hampton, R. R. (2013). Monkeys show recognition without priming in a classification task. *Behavioural Processes*, 93(q):50–61.

Bhatt, R. S., Wasserman, E. A., Reynolds, W. F., and Knauss, K. S. (1988). Conceptual behavior in pigeons: Categorization of both familiar and novel examples from four classes of natural and artificial stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*, 14(3):219–234.

Fei-Fei, L., Fergus, R., and Perona, P. (2007). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70.

Freedman, D. J. and Assad, J. A. (2011). A proposed common neural mechanism for categorization and perceptual decisions. *Nature Neuroscience*, 14(2):143–146.

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312–316.

Herrnstein, R. J., , Loveland, D. H., and Cable, C. (1976). Natural concepts in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 2(4):285.

Katz, J. S., Wright, A. A., and Bodily, K. D. (2007). Issues in the comparative cognition of abstract-concept learning. *Comparative Cognition & Behavior Reviews*, 2:79–92.

Lazareva, O. F., Freiburger, K. L., and Wasserman, E. A. (2004). Pigeons concurrently categorize photographs at both basic and superordinate levels. *Psychonomic Bulletin & Review*, 11(6):1111–1117.

Marsh, H. L. and MacDonald, S. E. (2008). The use of perceptual features in categorization by orangutans (pongo abelli). *Animal Cognition*, 11(4):569–585.

O'Hara, S. and Draper, B. A. (2011). Introduction to the bag of features paradigm for image classification and retrieval. *arXiv Preprint*, arXiv:1101.3354.

Sigala, N. (2009). Natural images: A lingua franca for primates? *The Open Neuroscience Journal*, 3(1):48–51.

Spalding, T. L. and Ross, B. H. (2000). Concept learning and feature interpretation. *Memory & Cognition*, 28(3):439–451.

Terrace, H. S. (2005). The simultaneous chain: A new approach to serial learning. *Trends in Cognitive Sciences*, 9(4):202–210.

Vonk, J. (2013). Matching based on biological categories in orangutans (pongo abelii) and a gorilla (gorilla gorilla gorilla). *PeerJ*, 1:e158.

Washburn, D. A. and Gulledge, J. P. (1995). Game-like tasks for comparative research: Leveling the playing field. *Behavior Research Methods, Instruments, and Computers*, 27(2):235–238.

Wright, A. A. and Lickteig, M. T. (2010). What is learned when concept learning fails: A theory of restricted-domain relational learning. *Learning & Motivation*, 41(4):273–286.

Zentall, T. R., Wasserman, E. A., and Urcuioli, P. J. (2014). Associative concept learning in animals. *Journal of the Experimental Analysis of Behavior*, 101(1):130–151.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.688v1 | CC-BY 4.0 Open Access | rec: 12 Dec 2014, publ: 12 Dec 2014