Informing Machine Perception With Psychophysics

By Justin Dulay, *Student Member IEEE*
Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556 USA

Sonia Poltoratski
Zoox, Inc., Foster City, CA 94404 USA

Till S. Hartmann
Zoox, Inc., Foster City, CA 94404 USA

Samuel E. Anthony
Independent Researcher

Walter J. Scheirer, *Senior Member IEEE*
Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556 USA

Gustav Fechner's 1860 delineation of psychophysics, the measurement of sensation in relation to its stimulus, is widely considered to be the advent of modern psychological science. In psychophysics, a researcher parametrically varies some aspects of a stimulus and measures the resulting changes in a human subject's experience of that stimulus; doing so gives insight into the determining relationship between a sensation and the physical input that evoked it. This approach is used heavily in perceptual domains, including signal detection, threshold measurement, and ideal observer analysis. Scientific fields, such as vision science, have always leaned heavily on the methods and procedures of psychophysics, but there is now growing appreciation of them by machine learning researchers, sparked by widening overlap between biological and artificial perception [1], [2], [3], [4], [5]. Machine perception that is guided by behavioral measurements, as opposed to guidance restricted to arbitrarily assigned human labels, has significant potential to fuel further progress in artificial intelligence (AI).

Digital Object Identifier 10.1109/JPROC.2024.3380905

In essence, psychophysical measurements of human behavior represent a richer source of information for supervised machine learning. What has been missing thus far from algorithms that learn from labeled data is a reflection of the patterns of error (i.e., the difficulty) associated with each data point used at training time. With knowledge of which samples are easy and which are hard, some measure of consistency can be achieved between the model and the human reference point. The true advantage of doing this stems from the human ability to solve perceptual tasks such as object recognition in an astonishingly fast and accurate way [6]. Human visual ability developed over millennia with changes in evolutionary genetic predisposition and thousands of hours of "pretraining" for object recognition tasks during development. By leveraging a more powerful learning system—the brain—it is possible to improve machine learning training in new ways.

In this Point of View article, we advocate for an alternative to traditional supervised learning that operationalizes the science of psychophysics (Fig. 1). To begin, it is helpful to define a couple of terms related to this new combination of psychophysics and machine learning, which will be used throughout this article. We define *informing* machine learning with psychophysics as adding behavioral information at any stage of the machine learning training pipeline to improve model performance. Similarly, we consider *guiding* as a description of improving a model's performance on test data toward higher supervised learning metrics of accuracy, thus instantiating the concept of informing within a chosen task domain that the model operates in.

There are several ways to combine psychology principles and machine learning models. First, neuroscience-inspired blocks can be used within artificial neural network architectures to varying degrees of success [7], [8], [9]. Second, performing psychophysics experiments on trained models can provide insight into human-model similarity [10], [11], [12], [13], [14]. Third, the strategy might be to get the performance and features of machine learning algorithms to be more human-like [15]. Fourth, psychophysical measurements can directly inform the machine learning model without changing the underlying architecture [2], [5]. In this article, we emphasize this last approach because it modularly fits into the training regime of many machine learning algorithms without requiring extensive modification of the algorithm or changing its underlying training objective. It is also compatible with the other three approaches.

We can view the choice of a psychophysical measurement type as a hyperparameter and the psychophysical measurements themselves as additional labels for data points to be used during training. In traditional supervised learning, performance is limited by the arbitrary labels reflecting class membership, which are the only source of information providing guidance on how to treat individual samples during training. Psychophysically informed supervised learning is a more complete learning pipeline because of the measured behavioral information. This is akin to the idea of optimal experiment design [16] because training data are systematically modified in order to optimize a chosen loss function.

In the rest of this article, we will take a brief tour of psychophysics for machine learning, including the problem space of perception where these ideas apply, as well as the expanding body of work related to psychophysically informed machine learning. To highlight the feasibility of gathering and using behavioral measurements in a new machine learning domain, we demonstrate how this training regime works in practice with a series of experiments related to handwritten character classification. As we will see, we are just scratching the surface of what can be achieved with this exciting interdisciplinary area of psychophysically informed machine learning.

I. PSYCHOPHYSICS FOR MACHINE LEARNING

A. Problem Space of Perception

Hans Moravec's famous AI paradox pointed out that the perceptual tasks that humans accomplish effortlessly have been among the most challenging to model. As he wrote, "it is comparatively easy to make computers exhibit adult-level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility" [17]. Indeed, recent advances in computational power have yielded tremendous advances in cognitive tasks that require extensive mental effort for human participants, such as advanced strategy games [18] or machine translation [19]. However, machine learning has continued to struggle to perform perceptual tasks that appear intuitive to humans but may have ambiguous or ill-defined "ground truth," such as medical image interpretation [20] or pro-social driving behavior [21].

Classification models derive complex latent representations from data without the need for rule-based assignment and can leverage psychophysical data alongside a class label. For instance, given a photograph of a chair, an additional psychophysical measurement associated with the photograph would relate information about the latent space of "chairness" to complement the extracted features and human-assigned label of "chair." This approach is particularly powerful in a machine learning context in which the correct label or solution is conceptually defined by humans, rather than an absolute ground truth. An example of this would be a model of the subjective assignment of "first impressions" made about the personality of a face in an image [1].

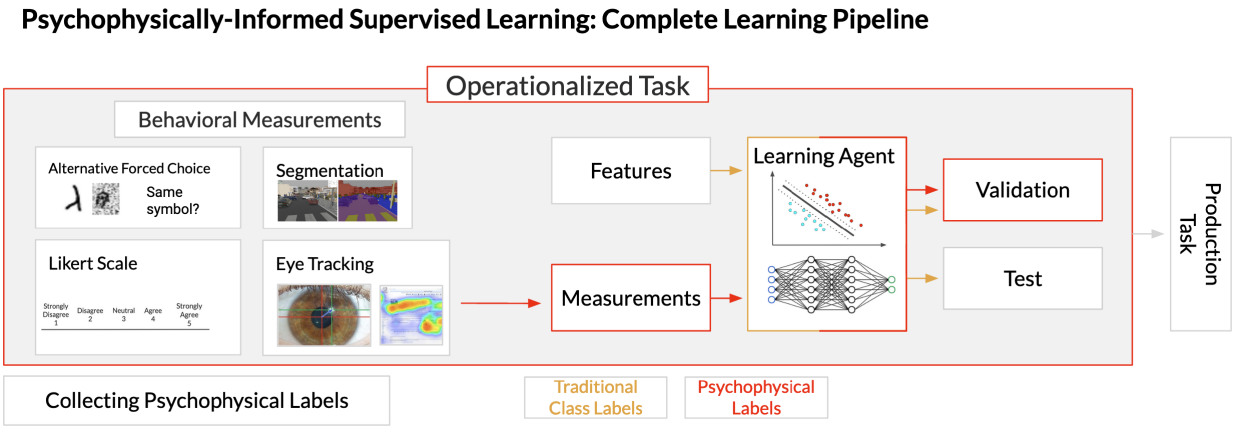

Fig. 1. *When performing similar visual tasks, humans and machine agents both solve for some latent representation of features. However, at present, the human capacity for this is superior. The central component of psychophysically informed learning is collecting quantifiable latent information from human experiments on visual recognition tasks and augmenting the* training *regime of machine learning models with it. A learning agent with a closer representational space to humans for a visual task, which is in part learned from the psychophysical measurements, solves that task in a way that is better than an agent without access to those measurements.*

B. Informing Machine Learning With Psychophysical Data

The goal of psychophysically informed data collection is to reveal additional information about the underlying latent representational space that yields the traditional annotations or labels that humans produce for machine learning datasets. This can be done by measuring information about each label's difficulty, confusability of label pairs, or integration of information over time. To cover the latent space effectively, experimental stimuli should effectively cover the sample space of the task, and experiments constructed from the stimuli should span a range of difficulty. Then, experimenters should select a response modality best matched to the machine learning goal and provide careful instructions that focus participant performance on the critical measurement. For example, a task utilizing reaction time would yield more accurate data from keypresses than mouseclicks, and if participants are explicitly instructed to "perform the task as quickly and accurately as possible, without taking breaks during trials."

In recent years, psychologists have effectively ported many in-lab study protocols to online crowdsourcing sites such as Amazon Mechanical Turk, demonstrating that data quality can be comparable [22] while allowing for rapid data collection from large numbers of participants [23]. However, there is an ongoing debate about the viability of such data in some circumstances [24]. The existing work in combining psychophysics with machine learning has found circumstances where crowdsourcing is effective. When conducting crowdsourced psychophysical experiments, researchers must be careful to preserve data quality. Current risks include using chatbots to solve tasks [25] and subjects rushing through tasks without trying (evident in identical reaction times or unnaturally fast ones).

Psychophysical labeling of the data provides an extra dimension beyond the usual supervised label. In a classification mode, a loss function can use this information to improve the learning process. For example, psychophysical labels can be incorporated into the loss to force the learning process to have more consistency with human perception. Consider a loss function where data points with associated low latency in response time result in high error for incorrect model predictions and data points with high latency yield lower error for mistakes. In other words, the model should not make mistakes on easy samples but is allowed to miss some of the hard ones in the same pattern humans do. Alternatively, there could be some advantage to leveraging the psychophysical information in a way that is *inconsistent* with human behavior but still improves the model performance. One way would be to reverse the error emphasis so that the training regime puts a higher priority on getting difficult samples correct.

In a regression mode, Likert scale (where the user makes a response on a discrete scale, usually from 1 to 5) data could be used to directly inform training to match human judgment, with additional psychophysical labels for the loss function as needed. For example, a neural network-based regressor can make direct use of averaged Likert scale response scores. As with experimental design in psychology, the options for modeling here are numerous.

C. Domains That Have Benefited From This Approach

Several domains have been investigated by researchers looking for ways to use psychophysics in machine learning (Fig. 2), spanning the four strategies we introduced in the beginning of this article. A loss

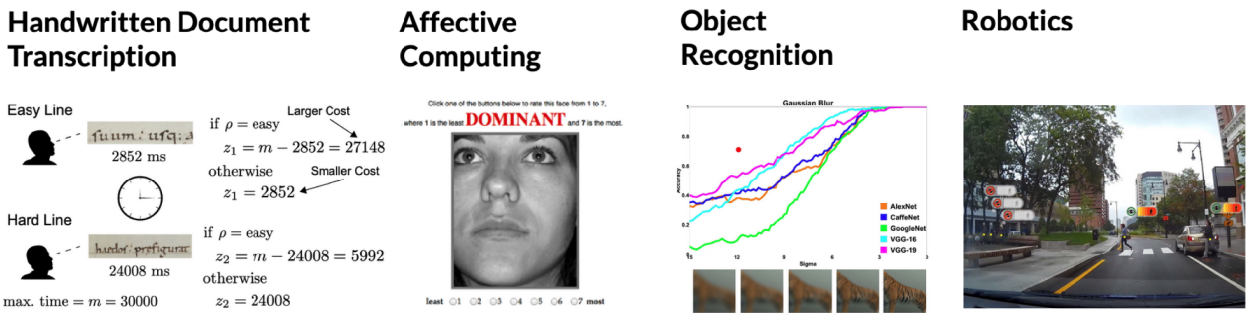

Fig. 2. ***Psychophysical labels have been utilized in multiple domains. Supervised model training for handwritten document transcription benefits from measurements of human reading behavior collected via crowdsourcing site workers or expert readers [5]. Affective computing tasks can leverage forms of annotation such as Likert-scale ratings [26]. Machine learning algorithms for object recognition also benefit from psychophysical evaluation and conditioning [4], [27]. The field of robotics, especially pertaining to autonomous vehicles, benefits from labels derived from behavioral measurements when considering social settings such as pedestrian crossings [28]. In general, when considering a perceptual task, a supervised learning model can always benefit from a larger label space.***

function able to use measurements from crowdsourced psychophysics experiments was first introduced by Scheirer et al. [2] for the domain of human biometrics. They conducted a series of behavioral experiments using the psychology crowdsourcing platform `TestMyBrain.org` where participants were presented with a two-alternative forced-choice question about whether or not a face was present in a given stimulus. Two-alternative forced choice is a common way to gather psychophysical data for recognition-based tasks because the operative step of recognition is binary—selecting a given positive sample type in a relative latent representation space and rejecting a given negative sample [29]. Stimuli were designed to test the impact of different controllable conditions such as noise and occlusion on face detection, which provides increased coverage of the full difficulty and scope of the problem compared to unperturbed labeling, maximizing the size of an informative label space. Different from the typical treatment of labeled samples in supervised machine learning, they found that difficult samples (e.g., a heavily occluded face) where the participants chose correctly after a relatively long period provided additional information for training support vector machine (SVM) classifiers with a loss function applying penalties based on perceptual measurements. Adding psychophysical measurements increased the robustness of the *label space*, which led to a state-of-the-art model for face detection.

In the domain of handwritten document transcription, loss functions incorporating psychophysical data for artificial neural network training have also been explored. Grieggs et al. [5] measured the reaction time of expert readers for documents of varying age and language, and used those measurements as labels in different loss functions that could emphasize easy or difficult samples. Observations were made about differences in reader behavior between expert and novice groups, with implications for other data labeling tasks. As with the original SVM work, this strategy was able to yield state-of-the-art performance for handwritten document transcription.

It has been demonstrated that humans adeptly make complex judgments about personality traits in miniscule amounts of time [30]. In the domain of affective computing, using two-alternative forced-choice or Likert scale ratings and regression models, it is possible to model this phenomenon using machine learning. The ChaLearn Looking at People First Impressions Challenge Competition [31] focused on models for the Big 5 personality traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism). McCurrie et al. [26] looked at a set of different traits, including trustworthiness, dominance, IQ, and estimated age. Work by Rojas et al. [1] has looked at modeling some of these traits in a classification mode. This research is a natural extension of laboratory testing in social psychology and is facilitated by psychophysical measurements.

In the domain of object recognition, an agent attempts to disentangle a learned manifold on some latent representation space of learned data [6]. Psychophysical evaluation has played an important role in evaluating biological and artificial vision, including their similarities and differences. Perturbed stimuli (e.g., rotated objects) activate neurons at different levels than canonical object views of the same stimuli. This same effect was observed in artificial neural networks [11]. Likewise, temporal processing in the brain has been modeled [32], [33], [34], [35] and validated with strategies that compare models with human performance [12], [13]. However, differences also exist between models and biological reference points. For instance, while some artificial neural networks generalize better than humans on some types of noise,

humans outperform them on many noisy recognition tasks [10], [36], [37]. This work demonstrates that the incorporation of human behavioral measurements within the label space of specific recognition tasks where the artificial agent typically fairs poorly can be beneficial. RichardWebster et al. [11], [27] introduced a framework to evaluate different types of image perturbations (blurring, rotation, and resolution) and their effects on both artificial and human performance for object and face recognition tasks, respectively. Zhang et al. [4] have suggested the use of human gaze measurements for improving performance in various object-related tasks, especially in a reinforcement learning context. Also, Dulay and Scheirer [38] observed that models can learn psychophysical representations of task difficulty that can transfer between tasks.

Kumar et al. [39] observed an inverse relationship between the metric of "perceptual scores" and accuracy, where such scores are computed in a deep feature space. While perceptual scores are a documented form of feature similarity strictly in models, we are more concerned here with the process of enhancing model accuracy with rigorous and easy-to-incorporate methods and procedures from the field of psychology. Importantly, better model accuracy can be achieved in ways that are not necessarily consistent with human responses, thus affirming the observation of Kumar et al. [39] across models and human reference points.

Finally, in the domain of robotics, psychophysics has been positioned as a means to create more generalizable embodied intelligence in robotic systems [40]. By simulating an artificial environment with many potential scenarios using perturbed inputs, a robotic system can generalize by learning to perform on stimuli that could potentially appear in the wild. This borrows from human-in-the-loop learning for autonomous systems but remains distinct in that the measurements for the scenarios are taken from people before the model is trained. Furthermore, it has also been suggested that the models used for autonomous driving can benefit from the human perception of pedestrians in uncertain situations (e.g., a pedestrian at a crosswalk who is not in motion but may intend to cross), which reveals more information about the situation at hand [28].

II. CASE STUDY: OPTICAL CHARACTER RECOGNITION INFORMED WITH PSYCHOPHYSICS

Optical character recognition (OCR) is a popular supervised learning task where the objective is to classify the characters within images of text. Sometimes those images are of poor quality, rendering the task more challenging to the learning agent while providing a representative example of text in the wild. Humans, equipped with a rich latent understanding of text recognition, typically outperform artificial agents on nontrivial OCR tasks. In this illustrative case study, we conducted a series of human behavioral measurements in crowdsourced experiments to operationalize the training stage of a supervised deep learning agent with psychophysical data.

A. Dataset Preparation and Behavioral Experiments

We implemented both the psychophysical tasks and the OCR machine learning task using a subset of the Omniglot dataset [41]. The dataset contains images of handwritten characters from hundreds of typesets, many of which a typical crowdsourced study participant would be unfamiliar with. In order to prepare a stimulus dataset for the human behavioral experiments, we selected 100 random classes of distinct characters represented in the images from the original Omniglot dataset. To augment these data, we generated a counterpart sample for each image with a deep convolutional generative adversarial network (DCGAN) [42] using the pretrained weights for inference to increase intraclass variance and the sample size per class. This resulted in a dataset of 100 classes with 40 instances per class for the psychophysical stimulus dataset to be viewed by the participants.

We conducted a series of four different psychophysical behavioral experiments on variations of a two-alternative forced-choice task with human participants. For each experiment of this particular task, the participant viewed two different images from the stimulus dataset with a prompt asking whether the images represent the same symbol or not. The first image of the pair was chosen at random, while the second was chosen from the same class or a different one with a probability of 0.5.

The following items correspond to the partitions within Table 1, which contains the experimental results.

1) The first experiment was a control experiment where the images from the original stimulus dataset were not perturbed. As a baseline, this task presented instructional prompts that were not tailored to psychophysics tasks, but rather asked participants for input on non-perturbed inputs—tasks machine learning models typically perform well on. Users made their responses using a cursor, which is typical in labeling tasks but does not yield reliable reaction-time estimates.
2) The second experiment used the exact same images as the control experiment, but we modified the instructional prompts slightly. Participants were instructed to complete the task "as quickly and accurately as possible" and they were allowed to complete the task by pressing an *F* or *J* key. We consider these keys to be easy ones to press based on most keyboard layouts. We considered this prompt set to be easier to understand than the first.

Table 1 Test Time Top-1 Accuracy Reflects Substantial Benefit When Using Reaction Time as an Additional Label. We Conducted a Two-Way ANOVA Considering the Experimental Protocols and Loss Functions as Variables to Assess If the Changes in Accuracy Are Significant. The Results Have a p-Value From Two-Way ANOVA of 2.02×10^{-6} When Considering the Main Effect of Using Reaction Time as a Component of the Loss Function. The F-Statistic for the Same Main Effect Was 17.51 for This Statistically Significant Result of the Two-Way ANOVA. The Error Bars Reflect the Standard Error From the Mean for the Model Accuracies Across Fivefold. Using the Averaged Accuracy Score as a Label Rarely Yielded Substantial Benefit. However, We See Improved Performance When Using Reaction Time as a Psychophysical Parameter in All Cases

Control experiment	Train Accuracy	Test Accuracy	95% C.I.
Cross Entropy	0.741 ±0.005	0.705 ±0.004	0.078
Avg. Accuracy	0.743 ±0.005	0.692 ±0.008	0.055
Avg. Reaction Time	**0.754 ±0.005**	**0.719 ±0.004**	0.055
Different Prompts	Train Accuracy	Test Accuracy	95% C.I.
Cross Entropy	0.732 ±0.005	0.705 ±0.004	0.078
Avg. Accuracy	0.723 ±0.003	0.697 ±0.008	0.062
Avg. Reaction Time	0.731 ±0.005	**0.729 ±0.005**	0.062
Blurred Images	Train Accuracy	Test Accuracy	95% C.I.
Cross Entropy	0.691 ±0.005	0.642 ±0.005	0.062
Avg. Accuracy	0.643 ±0.008	0.542 ±1.005	0.107
Avg. Reaction Time	**0.710 ±0.003**	**0.668 ±0.006**	0.068
Noisy Images	Train Accuracy	Test Accuracy	95% C.I.
Cross Entropy	0.672 ±0.006	0.602 ±0.005	0.062
Avg. Accuracy	0.641 ±0.007	0.592 ±0.016	0.110
Avg. Reaction Time	**0.732 ±0.004**	**0.680 ±0.005**	0.062

Fig. 3. **Are these the same character?** ***An example two-alternative forced-choice OCR task as seen from the participant's view. (d) and (f) Character pairs where the class labels differ; (a)–(c) and (e) represent the same class pairing. The blurred and noisy images lead to more informative psychophysical labels for operationalization within the machine learning task during training.***

3) The third experiment incorporated the condition of Gaussian blur. A randomly chosen image from one of the 100 classes was blurred using one of five different kernels (also chosen randomly with respect to the level of perturbation), and the other image that was paired with it was left unaltered. We expanded the range of experiment difficultly to avoid a ceiling effect, a form of scale attenuation in which the maximum performance measured does not reflect the true maximum of the independent variable. In this case, we expect a measurement ceiling if the task is too easy for participants and maximally accurate responses lose their relationship to task difficulty.
4) The fourth experiment was conducted like the third experiment, but with Gaussian noise instead of blurring. Likewise, there were five different levels of Gaussian noise that could be applied, selected at random. Refer to Fig. 3 for sample depictions of this task.

Participants for the behavioral experiments were recruited on Amazon Mechanical Turk. Each participant completed 100 two-alternative forced-choice trials, and each of the four experiments had 1000 unique participants. A two-alternative forced-choice task efficiently determines the implicit difficulty of a sample pairing within a dataset. Each pair was shown at least three times and up to five times. The reaction time for each task was recorded by measuring the interval between the first presentation of the stimuli and the participant's recorded response. Since the model will only be doing single-image classification, we only considered the human reaction time for an image from its most difficult pairing. This is how the information from the forced-choice task transfers to the classification task. Spam from dishonest subjects and incomplete response sets were manually pruned.

Each experiment formed a psychophysically annotated dataset that was used later in the machine learning task. The resulting four datasets included all of the image pairs shown in each two-alternative forced-choice instance, the responses of the participants, the average accuracy of participants on each image pairing, and the average reaction time of participants on each image pairing. Each image pairing was distributed approximately evenly across all participants in each experiment. The averaged accuracies and reaction times per pairing were calculated across all responses for that instance, where the number of responses per pairing varied slightly but not significantly. Each instance within the dataset is thus the average across the responses for that particular image pairing within the dataset.

B. Loss Function Formulation

The psychophysical loss utilized data collected from human behavioral experiments in addition to traditional supervised learning data. It expanded upon a traditional supervised learning pipeline. In this case study, we made use of a standard ResNet50 deep

neural network model [43] and cross-entropy loss. For all experiments, we used the same hyperparameters for the model.

The cross-entropy loss is defined as

$$\mathcal{L} = -\left(\sum_j y_j \log(\hat{y}_j)\right)$$

where $\hat{y}_j$ is a model prediction and y_j is the traditional class label associated with it at the jth index in the dataset. In order to incorporate the psychophysical labels into cross-entropy loss, we normalized and scaled the measurements to fit within the expected range of the loss function values. Furthermore, we only considered modifying the behavior of the loss function on model outputs where the prediction was incorrect. Here, reaction time or accuracy is used as a proxy for sample difficulty. Easy samples that the model classifies incorrectly result in stronger penalties for the model during training. We made use of the averaged reaction times and averaged accuracies separately from one another; we did not combine the two in a given loss function. To use these labels, we defined a psychophysical penalty in the manner of Grieggs et al. [5]

$$z_i = m - r_i$$

where z_i is the penalty, m is the maximum value for either reaction time or accuracy, and r_i is the psychophysical label (either reaction time or accuracy) at the ith index of the set of psychophysical label data pairings. Next, we incorporate z_i into the cross-entropy loss

$$\mathcal{L} = -\left(\sum_j y_j \left(\log(\hat{y}_j) z_i c\right)\right)$$

at the jth index in the dataset, where c is a scaling factor for the psychophysical penalty to modulate its impact (set at 0.5 in our experiments).

C. OCR Classifier Experiments

The study concluded that psychophysical loss improves the top-1 accuracy of the dataset by 1.1% points for the control experiment protocol and as much as 8% on others—a substantial improvement for a machine learning endeavor. The ResNet50 architecture used in these experiments was pretrained on ImageNet. We trained three models based on this architecture for each of the four psychophysical datasets from the behavioral experiments: a set of equally unperturbed images (control experiment), re-worded prompts for the control experiment set, blurred images combined with originals, and noisy images combined with originals. The ResNet models used the average human reaction time or average accuracy data as discrete values gathered from the data to be used during the training process as weighting terms in the loss function, rather than a differentiable continuous variable, with respect to the following models.

1) The first model was a standard ResNet50 with normal cross-entropy loss.
2) The second model substituted regular cross-entropy loss for the psychophysical loss using average accuracy.
3) The third model substituted regular cross-entropy loss for the psychophysical loss using average reaction time.

We trained each model for 20 epochs. In order to report accuracy fairly, we repeated model training five times with a different random seed. The results reported in Table 1 reflect the mean accuracy of each run along with standard error.

In addition, we conducted a two-way ANOVA test over the experimental results of the case study. The case study combined experimental protocols with different loss functions; therefore, we conducted the two-way ANOVA with two variables: 1) the four experimental protocols and 2) the three loss functions. The goal of the two-way ANOVA was to determine the statistical significance of changes in machine learning model accuracy when applying different protocols to the data (such as perturbing the inputs in systematically random ways) and changing the loss functions (such as including human reaction times in the loss calculation). The four experimental protocols were a control with no human data included, the modification of annotator prompts, the blurring of images, and the addition of random noise to the images. The three loss functions were the control loss cross entropy, the inclusion of average human accuracy to the loss, and the inclusion of average human reaction time to the loss function.

We report the ANOVA details in Table 1. In this case study, the statistical analysis reveals that incorporating human reaction times into the loss space for ResNet-50 deep learning models, which yielded the best results, shows a significant difference in accuracy on supervised image classification tasks. These results have a p-value from two-way ANOVA of 2.02×10^{-6} when considering the main effect of using reaction time as a component of the loss function. In contrast, accuracy labels did not always outperform the control. Therefore, when integrating these new labels into machine learning training, it remains important to assess the effectiveness for the task. In this case, reaction time was the more informative measurement type. This has been shown in the literature for training artificial neural networks using psychophysical data [5]. However, there is no guarantee that this will generalize to all tasks.

III. CONCLUSION

Psychophysical labels from human behavioral experiments have been shown to improve the performance of supervised learning models in many different domains in the literature. We conducted a case study to demonstrate how quickly this strategy can be adapted to a new domain. More

work needs to be done to develop similar strategies for different modes of learning, including unsupervised and reinforcement learning. By improving training regimes or policy estimators in these fields, generalization may be achieved more effectively than with traditional strategies. In all, we hope that this work inspires future conversation and research at the intersection of psychology and computer science.

DATA AND CODE AVAILABILITY

All data and code used in this article can be found at: https://github.com/dulayjm/PyTorch-Psychophysics-Learning.

Acknowledgment

This work was supported in part by the U.S. National Science Foundation under Grant BCS:1942151. This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the University of Notre Dame IRB under Application No. 18-01-4341. ■

REFERENCES

[1] Q. M. Rojas, D. Masip, A. Todorov, and J. Vitria, "Automatic prediction of facial trait judgments: Appearance vs. structural models," *PLOS ONE*, vol. 6, no. 8, 2011, Art. no. e23323.

[2] W. J. Scheirer, S. E. Anthony, K. Nakayama, and D. D. Cox, "Perceptual annotation: Measuring human vision to improve computer vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1679–1686, Aug. 2014.

[3] S. Escalera et al., "ChaLearn looking at people challenge 2014: Dataset and results," in *Proc. ECCV Workshops*, Sep. 2014, pp. 459–473.

[4] R. Zhang et al., "AGIL: Learning attention from human for visuomotor tasks," in *Proc. ECCV*, Sep. 2018, pp. 663–679.

[5] S. Grieggs et al., "Measuring human perception to improve handwritten document transcription," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6594–6601, Oct. 2022.

[6] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?" *Neuron*, vol. 73, no. 3, pp. 415–434, Feb. 2012.

[7] F. P. de Lange, M. Heilbron, and P. Kok, "How do expectations shape perception?" *Trends Cognit. Sci.*, vol. 22, no. 9, pp. 764–779, Sep. 2018.

[8] J. Dapello, T. Marques, M. Schrimpf, F. Geiger, D. D. Cox, and J. J. DiCarlo, "Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations," *BioRxiv*, pp. 1–30, Jun. 2020.

[9] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," 2018, *arXiv:1811.12231*.

[10] R. Geirhos, C. R. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," in *Proc. NeurIPS*, 2018, pp. 1–13.

[11] B. RichardWebster, S. E. Anthony, and W. J. Scheirer, "PsyPhy: A psychophysics driven evaluation framework for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2280–2286, Sep. 2019.

[12] M. Schrimpf et al., "Brain-score: Which artificial neural network for object recognition is most brain-like?" *BioRxiv*, pp. 1–9, Sep. 2018.

[13] B. RichardWebster, J. Dulay, A. DiFalco, E. Caldesi, and W. J. Scheirer, "Psychophysical-score: A behavioral measure for assessing the biological plausibility of visual recognition models," 2023, *arXiv:2210.08632*.

[14] C. Gontier, J. Jordan, and M. A. Petrovici, "DELAUNAY: A dataset of abstract art for psychophysical and machine learning research," 2022, *arXiv:2201.12123*.

[15] R. Geirhos, K. Meding, and F. A. Wichmann, "Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 13890–13902.

[16] C. Gontier, S. C. Surace, I. Delvendahl, M. Müller, and J.-P. Pfister, "Efficient sampling-based Bayesian active learning for synaptic characterization," *PLoS Comput. Biol.*, vol. 19, no. 8, 2023, Art. no. e1011342.

[17] H. Moravec, *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA, USA: Harvard Univ. Press, 1988.

[18] O. Vinyals et al., "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[19] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.

[20] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," in *Classification in BioApps*. Cham, Switzerland: Springer, 2018, pp. 323–350.

[21] N. Lacetera, M. Macis, and R. Slonim, "Will there be blood? Incentives and displacement effects in pro-social behavior," *Amer. Econ. J., Econ. Policy*, vol. 4, no. 1, pp. 186–223, Feb. 2012.

[22] L. Germine, K. Nakayama, B. C. Duchaine, C. F. Chabris, G. Chatterjee, and J. B. Wilmer, "Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments," *Psychonomic Bull. Rev.*, vol. 19, no. 5, pp. 847–857, Oct. 2012.

[23] N. Stewart, J. Chandler, and G. Paolacci, "Crowdsourcing samples in cognitive science," *Trends Cognit. Sci.*, vol. 21, no. 10, pp. 736–748, 2017.

[24] S. Haghiri, P. Rubisch, R. Geirhos, F. Wichmann, and U. von Luxburg, "Comparison-based framework for psychophysics: Lab versus crowdsourcing," 2019, *arXiv:1905.07234*.

[25] V. Veselovsky, M. H. Ribeiro, and R. West, "Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks," 2023, *arXiv:2306.07899*.

[26] M. McCurrie, F. Beletti, L. Parzianello, A. Westendorp, S. Anthony, and W. J. Scheirer, "Predicting first impressions with deep learning," in *Proc. IEEE FG*, May 2017, pp. 518–525.

[27] B. RichardWebster, S. Y. Kwon, C. Clarizio, S. E. Anthony, and W. J. Scheirer, "Visual psychophysics for making face recognition algorithms more explainable," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 252–270.

[28] M. Milford, S. Anthony, and W. Scheirer, "Self-driving vehicles: Key technical challenges and progress off the road," *IEEE Potentials*, vol. 39, no. 1, pp. 37–45, Jan. 2020.

[29] N. Prins and F. Kingdom, *Psychophysics: A Practical Introduction*, 2nd ed. New York, NY, USA: Academic, 2016.

[30] J. Willis and A. Todorov, "First impressions: Making up your mind after a 100-ms exposure to a face," *Psychol. Sci.*, vol. 17, no. 7, pp. 592–598, Jul. 2006.

[31] V. Ponce-López et al., "Chalearn LAP 2016: First round challenge on first impressions-dataset and results," in *Proc. ECCV Workshops*, Oct. 2016, pp. 400–418.

[32] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," 2016, *arXiv:1605.08104*.

[33] E. Watanabe, A. Kitaoka, K. Sakamoto, M. Yasugi, and K. Tanaka, "Illusory motion reproduced by deep neural networks trained for prediction," *Frontiers Psychol.*, vol. 9, Mar. 3389, Art. no. 340023, doi: 10.3389/fpsyg.2018.00345.

[34] A. Gomez-Villa, A. Martín, J. Vazquez-Corral, and M. Bertalmío, "Convolutional neural networks deceived by visual illusions," 2018, *arXiv:1811.10565*.

[35] T. Fel, I. F. R. Rodriguez, D. Linsley, and T. Serre, "Harmonizing the object recognition strategies of deep neural networks with humans," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 9432–9446.

[36] H. Jang, D. McCormack, and F. Tong, "Noise-trained deep neural networks effectively predict human vision and its neural responses to challenging images," *PLOS Biol.*, vol. 19, no. 12, Dec. 2021, Art. no. e3001418.

[37] R. Geirhos et al., "Partial success in closing the gap between human and machine vision," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 23885–23899.

[38] J. Dulay and W. J. Scheirer, "Using human perception to regularize transfer learning," 2022, *arXiv:2211.07885*.

[39] M. Kumar, N. Houlsby, N. Kalchbrenner, and E. D. Cubuk, "Do better ImageNet classifiers assess perceptual similarity better?" 2022, *arXiv:2203.04946*.

[40] N. Sünderhauf et al., "The limits and potentials of deep learning for robotics," *Int. J. Robot. Res.*, vol. 37, nos. 4–5, pp. 405–420, 2018.

[41] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, Dec. 2015.

[42] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

ABOUT THE AUTHORS

Justin Dulay (Student Member, IEEE) received the bachelor's degree in computer science from Saint Louis University, St. Louis, MO, USA, in 2021. He is currently pursuing the master's degree with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA.

His research interests include computer vision and machine learning.

Sonia Poltoratski received the Ph.D. degree in cognitive neuroscience from Vanderbilt University, Nashville, TN, USA, in 2017.

She was a Postdoctoral Researcher at Stanford University, Stanford, CA, USA. She is currently a Data Scientist specializing in visual perception at Zoox., Inc., Foster City, CA, USA.

Till S. Hartmann received the Ph.D. degree in neuroscience from Rutgers University, New Brunswick, NJ, USA, in 2011.

He was a Postdoctoral Fellow at Harvard Medical School, Boston, MA, USA. He is currently a Staff Software Engineer at Zoox., Inc., Foster City, CA, USA.

Samuel E. Anthony received the Ph.D. degree in psychology from Harvard University, Cambridge, MA, USA, in 2018.

He is an Independent Researcher, a Blogger, and a Board Member of the Many Brains Project.

Walter J. Scheirer (Senior Member, IEEE) is the Dennis O. Doughty Collegiate Professor of Engineering with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA. His research interests within the field of computer science include artificial intelligence, computer vision, machine learning, and digital humanities.

Prof. Scheirer is a Global AI Leader. He is serving as the Chair for the IEEE Computer Society Technical Community on Pattern Analysis and Machine Intelligence and a Board Member for the Computer Vision Foundation.