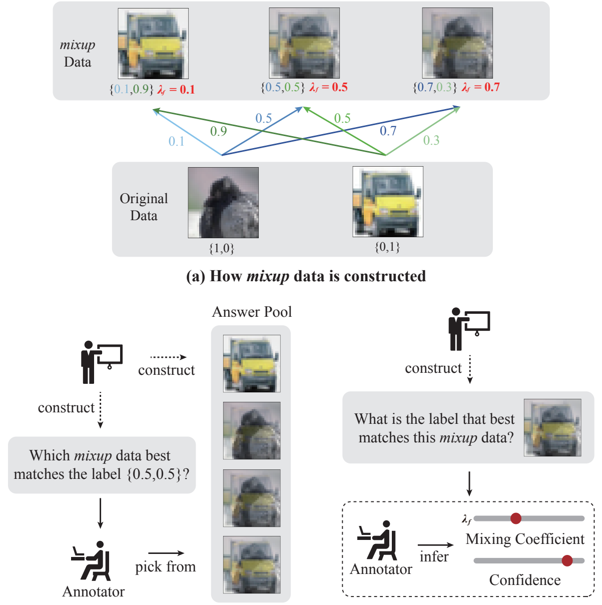
Human-in-the-Loop *Mixup*

Katherine M. Collins* [1] **Umang Bhatt**[1,2] **Weiyang Liu**[1,3] **Vihari Piratla**[1] **Ilia Sucholutsky**[4] **Bradley Love**[2,5]
Adrian Weller[1,2]

[1]University of Cambridge
[2]The Alan Turing Institute
[3]Max Planck Institute for Intelligent Systems
[4]Princeton University
[5]University College London

Abstract

Aligning model representations to humans has been found to improve robustness and generalization. However, such methods often focus on standard observational data. Synthetic data is proliferating and powering many advances in machine learning; yet, it is not always clear whether synthetic labels are perceptually aligned to humans – rendering it likely model representations are not human aligned. We focus on the synthetic data used in *mixup*: a powerful regularizer shown to improve model robustness, generalization, and calibration. We design a comprehensive series of elicitation interfaces, which we release as `HILL MixE Suite`, and recruit 159 participants to provide perceptual judgments along with their uncertainties, over *mixup* examples. We find that human perceptions do not consistently align with the labels traditionally used for synthetic points, and begin to demonstrate the applicability of these findings to potentially increase the reliability of downstream models, particularly when incorporating human uncertainty. We release all elicited judgments in a new data hub we call `H-Mix`.


mixup Data
{0.1,0.9} $\lambda_f = 0.1$
{0.5,0.5} $\lambda_f = 0.5$
{0.7,0.3} $\lambda_f = 0.7$
0.1
0.9
0.5
0.5
0.7
0.3
Original Data
{1,0}
{0,1}
Answer Pool
construct
construct
Which *mixup* data best matches the label {0.5,0.5}?
Annotator
pick from
construct
What is the label that best matches this *mixup* data?
Annotator
infer
λ_f
Mixing Coefficient
Confidence

(a) How *mixup* data is constructed

(b) Elicitation Setting I: Endorse a synthetic image to match a label

(c) Elicitation Setting II: Infer the label for a synthetic image

Figure 1: Framework overview. A) Synthetic data generating process used in *mixup*; B) and C) depict elicitation settings: B) participants endorse a synthetic image to match a label, C) participants infer the label for a synthetic image and provide their uncertainty in the corresponding inference.

1 INTRODUCTION

Synthetic data is proliferating, fueled by increasingly powerful generative models, e.g. [Goodfellow et al., 2014a, Dhariwal and Nichol, 2021]. These data are not only consumed directly by people – but, as training predictive models on synthetic data has been found to unlock tremendous advances in machine learning (ML) [Silver et al., 2016, de Melo et al., 2022, Emam et al., 2021, Jordon et al., 2022], synthetic data is increasingly employed to train algorithms serving as engines of many applications humans may interact with. However, it is not always clear whether human perceptual judgments of synthetically-generated data match the generative process used to create them.

Aligning networks to match humans' perceptual inferences could be a way to further ensure model reliability, trustworthiness, downstream performance, and robustness [Nanda et al., 2021, Chen et al., 2022, Fel et al., 2022, Sucholutsky and Griffiths, 2023]. If these data are *not* aligned with human percepts, then performance potentially could be improved by altering such signals to better match the richness of human judgments: this has proven effective when aligning

*Correspondence to: kmc61@cam.ac.uk

Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023), PMLR 216:454–464.

models with human probabilistic knowledge and perceptual uncertainty [Collins et al., 2022a, Sanders et al., 2022, Sucholutsky et al., 2023]. We argue that one ought to *verify* whether synthetic data aligns with human perception, and if not, explore whether training with *human-relabeled* examples improves model performance.

In this work, we take a step in this direction by focusing on *mixup* [Zhang et al., 2018]: a method whereby a model is trained only on synthetic, linear combinations of conventional training examples. We focus on *mixup* for three key reasons. First, the generative process for synthetic *mixup* examples is very simple, and provides us with direct access to the "ground truth" generative model parameters; that is, we have precise control over the mixing coefficient used to create the mixed image. This enables us to compare any discrepancy between human perceptual judgments and this parameter explicitly. A generative model of the likes of a generative adversarial network (GAN) [Goodfellow et al., 2014a] or diffusion model [Ho et al., 2020] does not as easily permit these kinds of precise comparisons. Second, despite this simplicity, *mixup* is a powerful and popular training-time method that has been leveraged to address model fairness [Chuang and Mroueh, 2020], improve model calibration [Thulasidasan et al., 2019, Zhang et al., 2022], and increase model robustness via regularizing the form of category boundaries learned implicitly [Zhang et al., 2020, Verma et al., 2022]. *mixup* is frequently used as a strong benchmark for many new data augmentation and regularization techniques [Hendrycks et al., 2019, 2022]. Third, prior work in human categorical perception – revealing that humans show non-linear "warping" effects along category boundaries [Harnad, 2003, Folstein et al., 2013, Goldstone and Hendrickson, 2010] – suggests that humans *will* differ in their percepts from the linear category boundaries encouraged by *mixup*.

To that end, we consider whether *mixup* labels match human perception, and if not, how the labeling scheme can be improved to better align with human intuition – and human uncertainty – to potentially enhance model performance. We focus on two flavors of elicitation: 1) having participants "construct" a midpoint between categories by selecting from a set of synthetic images, and 2) eliciting traces of humans' broader category boundary across a range of mixed images by having participants directly intervene on the synthetic label, along with their uncertainty in their judgments. We design three online elicitation interfaces to address these questions, which we offer as The Human-in-the-Loop Mixup Elicitation Suite (`HILL MixE Suite`). We collect judgments from over 150 humans on these synthetically combined images, which we release in a dataset we call "Human Mixup" or `H-Mix`[1]. We then demonstrate one of the possible use cases of this data: as adjusted training data for deep networks to improve model generalization, calibration, and adversarial robustness. We depict our general framework in Fig. 1. Our data (`H-Mix`) and general elicitation paradigm (e.g., `HILL MixE Suite`) could support a range of downstream applications: from serving as new training labels for machine learning or benchmarking model alignment to auditing synthetic data, and informing cognitive science studies, among others. We see our work as a step in the exciting direction of a human-centric perspective on synthetic data powering many ML algorithms, which emphasizes the potential utility of *human* uncertainty in human-in-the-loop systems.

2 PROBLEM FORMULATION

2.1 DECOUPLING DATA AND LABEL MIXING IN *MIXUP*

We first review *mixup* [Zhang et al., 2018] and explicate the recipe by which synthetic examples are created. We employ the nomenclature and notation around "*mixup* policies" from [Liu et al., 2021b]. We assume access to a finite set of N samples $\{(x_1, y_1), (x_2, y_2, \cdots, (x_N, y_N)\}$. *mixup* training consists of constructing synthetic training examples $(\tilde{x}, \tilde{y})$ via linear combinations of pairs of the training observations $(x_i, y_i), (x_j, y_j)$ for $i, j \in [1, N]$, corresponding to the following data and label mixing functions:

$$\text{Data Mixing: } f(x_i, x_j, \lambda_f) = \lambda_f x_i + (1 - \lambda_f) x_j = \tilde{x} \quad (1)$$

$$\text{Label Mixing: } g(y_i, y_j, \lambda_g) = \lambda_g y_i + (1 - \lambda_g) y_j = \tilde{y} \quad (2)$$

where λ_f and λ_g are defined as the ***data mixing coefficient*** and ***label mixing coefficient***, respectively. We refer to the combined images x_i, x_j and their labels y_i, y_j as the ***endpoints***. For a specified mixing coefficient λ, we denote the resultant image as $\tilde{x}$. *mixup* typically assumes $\lambda_f = \lambda_g$. We instead decouple the data and label mixing functions to permit a more general formulation where the data and label mixing functions can have different coefficients.

2.2 HUMAN-IN-THE-LOOP *MIXUP*

Our decoupling allows us to probe whether human percepts align with either the mixing policy over the observations (f) or the targets (g). Human alignment of these mixing policies could be important for several reasons. First, we may want to understand how well the synthetic data used to power many models deployed on the web matches human perceptual judgments, thus ensuring model trustworthiness. Second, given that these policies do afford *mixup* downstream niceties–such as improved generalization, robustness, and calibration– we believe it is worth exploring whether modulating such data to be more human-aligned can yield similar, or better, performance boosts. We, therefore, pose

[1] All data, elicitation interfaces, and experiment code will be included in our repository.

two questions to separate groups of human participants to better elucidate alignment of the *mixup* synthetic data construction:

RQ1: What $\tilde{x}$ do participants believe matches a given $\tilde{y}$?

RQ2: Conditioned on $\tilde{x}$, what do humans perceive as $\tilde{y}$?

Unless otherwise noted, we focus on the setting where we maintain the structural form of f and g; that is, they are each parameterized by a single mixing coefficient. We discuss alternative functional forms which may more flexibly capture the richness of human percepts of these synthetically-constructed images in the Supplement.

3 SELECTING A MATCHING MIDPOINT (RQ1)

We first consider holding g fixed and *creating* a perceptually-aligned input. We liken this setting to counterfactual data creation from [Kaushik et al., 2019].

3.1 PROBLEM SETTING

In our setup, we inform participants that they will observe samples combined from particular categories y_i, y_j. We fix the label mixing coefficient, λ_g (here, to 0.5 – but our procedure could be extended to arbitrary mixing coefficients) and ask participants to construct a viable $\tilde{x}$ that would be perceived as the λ_g mixture of the categories. Ideally, we may want to see what kind of example the participant may select from the full space of possible examples (in our case, images); for simplicity, we restrict that participants choose a $\tilde{x}$ from a set of M pre-constructed linear interpolations which we refer to as $\{\tilde{x}_j\}_{j=1}^{M}$, which we refer to as $\tilde{X}_M$. Each $\tilde{x}_j$ is the result of executing f for a given λ_f. Here, we consider a sweep of over the mixing coefficients $[0.0, 0.1, ...0.9, 1.0]$. From their selected image, we can uncover how their perception of the data-generating process differs relative to what was actually used to create said selected image.

3.2 ELICITATION PARADIGM

We design two means of eliciting people's selection of a $\tilde{x}$:

1. Interface 1 (`Construct`): participants use their keyboard to iterate over $\tilde{X}_M$ (ordered), where key presses increment or decrement j by one such that $\tilde{x}_j$ are cycled through at increments of 0.1. One mixed example is displayed on the screen at a time. Participants press "Next" when they are happy with the selected $\tilde{x}_j$.

2. Interface 2 (`Select-Shuffled`): participants see all $\tilde{x} \in \tilde{X}_M$ on the screen at once. Mixed examples are *shuffled* and thus presented in an unordered fashion. Participants indicate their selection by clicking on the $\tilde{x}_j$ they think best matches λ_g.

Example interfaces, and design rationales, are depicted in the Supplement. As mentioned, participants are explicitly told the categories being combined (y_1, y_2) and are asked to indicate the image that they think is most likely to be perceived as the 50/50 combination of the mixed images by *100 other crowdsourced workers*. Such elicitation language is drawn from [Chung et al., 2019], following a recommended practice in high-fidelity human subject elicitation whereby participants are asked to assume a third-person perspective when responding [Prelec, 2004, Oakley and O'Hagan, 2010].

Stimuli and Participants We focus on a random subset of the `CIFAR-10` test images, a dataset containing low-resolution images drawn from ten categories of objects and animals (e.g., truck, ship, cat, dog) [Krizhevsky et al., 2009]. We use the test set as this permits downstream comparisons against `CIFAR-10H`: an expansive set of approximately 51 human annotators' judgments about each example [Peterson et al., 2019, Battleday et al., 2020]. From each unique category combination (e.g., truck-dog, ship-cat, cat-dog), we sample 6 random images from each of the categories and linearly combine them in pixel-space. We sample 249 such image pairings, and for each, we sweep over the space of 11 mixing coefficients incremented by 0.1 between $\lambda_f = 0.0$ and $\lambda_f = 1.0$ (totaling 2739 synthetically mixed images in total). We recruit a total of 70 participants from Prolific [Palan and Schitter, 2018] and hosted on Pavlovia. 45 participants were allocated to `Construct`, which was subdivided into two conditions based on the starting point of the selection: 23 participants started at the $\lambda_f = 0.9$ mixing coefficient, and 22 participants were assigned always starting at $\lambda_f = 0.1$. The remaining 25 participants were allocated to `Select-Shuffled`. Further details are included in the Supplement.

3.3 INVESTIGATING DATA MIXING ALIGNMENT

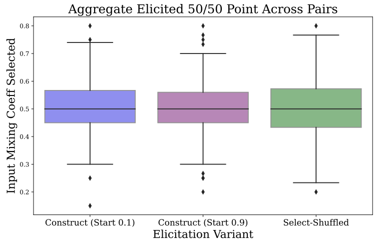


Figure 2: Averaging human participants' selections per image pair reveals the typical pair is minimally relabeled.

We find that, in aggregate, humans' selections indicate alignment with the underlying mixing coefficient (see Fig. 2),

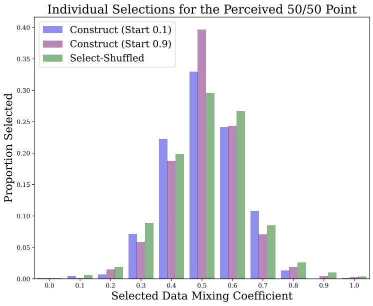


Figure 3: Participants do not always endorse the 50/50 point suggesting misalignment in the data labeling policy. The bar plot depicts extracted mixing coefficient of individuals' selections for the perceptually-aligned midpoints.

which is stable across elicitation methodology. However, we cannot conclude from these data that the *mixup* data policy is aligned with humans. If we look at the selections made by individual humans, we see that a substantial portion endorsed a $\tilde{x}$ which differed from that which would naturally be assumed in *mixup* (see Fig. 3). Example image pairs that yield high relabeling across interface types are shown in Fig. 4. We identify 9 such image pairs that are highly relabeled (which we define as $|\lambda_h - 0.5| \geq 0.15$, where we let λ_h be the mixing coefficient used to generate the $\tilde{x}$ selected by humans) across interface types. This picture suggests that indeed human percepts are *not* consistently aligned with the synthetic data construction process – and that perhaps with a larger set of stimuli, more such examples can be recovered. Note, there are a total of 101 image pairs that are endorsed by at least one interface as in need of high relabeling. More work is needed to elucidate whether discrepancies in relabeling were induced by the varied interface design or simply individual differences among the participants recruited.

Takeaways These data suggest that while in general, the 50/50 combined image is recoverable – at an individual level, such percepts are more nuanced. Our data, which we include as part of `H-Mix`, indicate systematic differences in perceptions of synthetically-constructed data. These differences emerge somewhat robustly across elicitation types. We next turn to richer traces of humans' perceptual representations of these synthetically-generated data.

4 ELUCIDATING ALIGNMENT OF THE LABEL MIXING POLICY (RQ2)

The above elicitations have focused only on the 50/50 point; however, *mixup* trains on synthetically-generated images sampled for a wide range of mixing coefficients. It, therefore, warrants study to analyze human perceptual alignment over a richer spectrum of mixing coefficients. We consider instead eliciting humans' judgments over what the label mixing coefficient λ_g ought to be. Studying the alignment of g could push forward a deeper understanding of what the data often used to train *mixup* and similar methods even means to humans, and potentially further motivate the design of alternative relabeling schemes (see Section 5). We therefore now focus on utilizing human input to design a perceptually-aligned target *mixup* policy g_h.

4.1 PROBLEM SETTING

We assume f is a linear mixing policy over inputs employed in [Zhang et al., 2018]. To form our human-aligned target policy, we want to find a function $g_h(y_i, y_j, \lambda) = \tilde{y}$ such that $\tilde{y}$ perceptually corresponds to the associated mixed input $f(x_i, x_j, \lambda) = \lambda x_i + (1 - \lambda)x_j = \tilde{x}$. How do we get $\tilde{y}$ from people efficiently?

We consider matching λ_g to what humans *infer* λ_f to be. In this setup, we assume humans are aware of the generative processes f and g_h, and are shown the mixed image $\tilde{x}$ and underlying labels y_i, y_j. People are then tasked with forming a probabilistic judgment as to what the underlying mixing coefficient is that generated the observed image $\tilde{x}$ when given the underlying y_i, y_j – e.g., judging $P(\lambda_f|\tilde{x}, y_i, y_j)$.

If human perception is aligned to the underlying linear *mixup* policies, then the human predicted mixing coefficient λ_h should be equivalent to λ_f, rendering $\lambda_f = \lambda_g = \lambda$ a sensible mixing scheme. However, if human estimates are not aligned, we may consider setting $\lambda_g = \lambda_h$ to make g yield a $\tilde{y}$ which best corresponds to humans' percepts of $\tilde{x}$.

4.2 ELICITATION PARADIGM

To elicit such information, we design a new interface where subjects infer the mixing coefficient between two given labels. We show each worker a mixed image and tell them the categories that were mixed to generate the image. Participants also provide us with their *uncertainty* in their inference. As some image combinations appear quite convoluted, we reason that subjects' confidence in their inference – or lack therefore – may provide interesting signals as to the perceptual sensibility of the mixed images. We run our relabeling experiment on $N = 81$ participants again through Prolific [Palan and Schitter, 2018]. Further details are included in the Supplement.

Stimuli selection Similar to Section 3.2, we sample images to mix from `CIFAR-10` [Krizhevsky et al., 2009]. We do so in a class-balanced fashion: 46 mixed images are sampled for each of the 45 possible class combinations, resulting in 2070 total stimuli. Each mixed image is formed by constructed by selecting a data mixing coefficient $\lambda_f \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$.

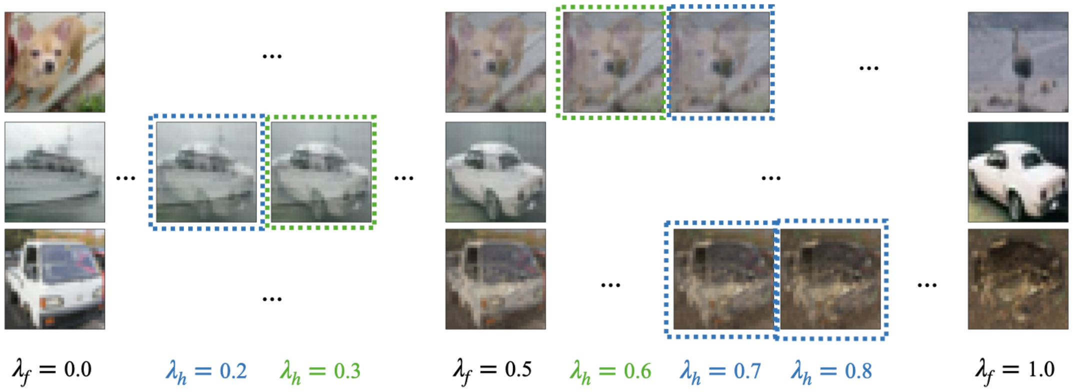


Figure 4: Example image pairs where substantial relabeling of the 50/50 point was recommended across all interface types. Synthetic images highlighted in blue received the most endorsements from participants across all interface types, with images in green receiving the second most. For row three, participants were split equally between two selections. The mixing coefficient (λ_f or λ_h) used to construct the images is shown along the bottom.

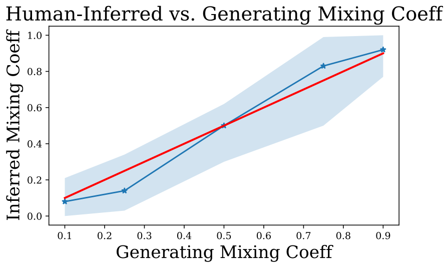
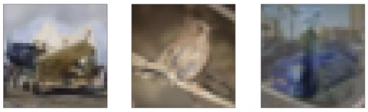


Figure 5: We uncover a sigmoidal relationship between humans' inferred mixing coefficient (λ_h, blue) as compared to the mixing coefficient used to generate the image (λ_g, red) suggestive of misalignment. We depict the median, along with the 25th and 75th percentiles. The red line indicates what the exact parallel between λ_h and λ_f would look like (highlighting perceived human deviation).



	Dog, Airplane	Bird, Cat	Automobile, Bird
Generating λ_f	0.25, 0.75	0.5, 0.5	0.5, 0.5
Human-Inferred λ_h	0.42, 0.58	0.99, 0.01	0.87, 0.13

Figure 6: Examples of average human relabelings of the generating mixing coefficient reveal discrepancies.

4.3 VALIDATING THE MIXING COEFFICIENT AGAINST HUMAN RESPONSES

We now compare the human-inferred mixing coefficient against the generating coefficient and analyze participants' uncertainty in such inferences. We also conduct a preliminary exploration into the relationship between participants' predicted uncertainty and the ambiguity of the underlying images being combined.

4.3.1 Relationship between Generating Mixing Coefficient and Alignment

We consider whether participants recover the data mixing coefficient: in Fig. 5, we show the median relabeling for images at given coefficients. We observe a non-linear, roughly sigmoidal structure to human relabelings, consistent with past research in human categorical perception [Harnad, 2003, Goldstone and Hendrickson, 2010, Folstein et al., 2013, Destler et al., 2019]. The aggregate recovery of the 50/50 point corroborates our findings in RQ1. However, we find that the picture is nuanced: wide confidence bounds suggest there are mixed images for which inferred mixing coefficients are substantially different from the parameterization assumed in *mixup*. Qualitative inspection of averaged relabelings for particular images (Fig. 6) – and across category pairs (Fig. 7) – also reveals such misalignment. We recommend future work to investigate why particular category pairs, for this dataset, are yielding different boundaries.

4.3.2 Analyzing Human Uncertainty

We next look closer at the reported human uncertainty in the mixing coefficient. First, we investigate whether human uncertainty estimates depend on the mixing coefficient assigned. Indeed, we do observe that humans' uncertainty

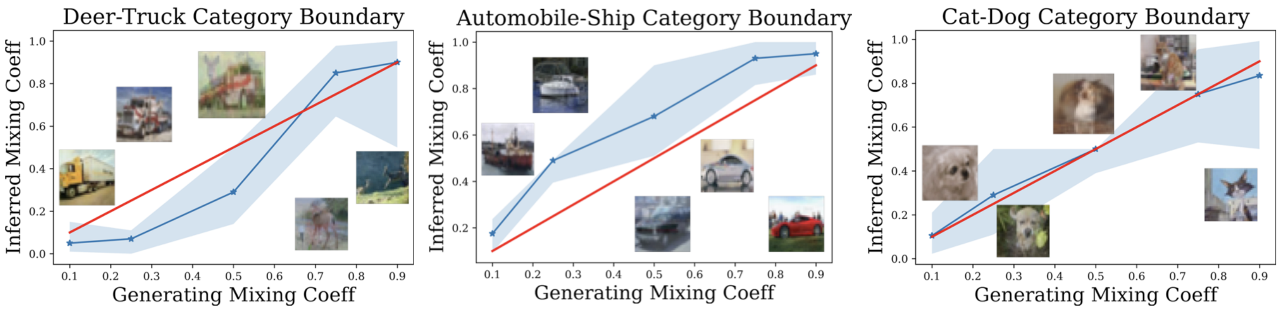


Figure 7: "Category boundaries" elicited from humans display a diverse structure. Many – though not all – deviate from linearity assumed in *mixup*. We overlay examples of synthesized stimuli, ordered by the λ_f used to create them.

tracks with the mixing coefficient (see Table 1); participants have the lowest confidence (i.e., highest uncertainty) for images generated from $\lambda_f = 0.5$.

Table 1: Participants' average reported confidence, or uncertainty, in their inference of the mixing coefficient (higher confidence means less uncertainty). Error bars indicate standard deviation across participants. The mixing coefficient here is computed as $|0.5 - \lambda_f|$ due to symmetry (a mixing coefficient of 0.1 is as extreme as 0.9).

Mixing Coefficient	**Reported Confidence**
0.1	0.79 ± 0.17
0.25	0.72 ± 0.20
0.5	0.63 ± 0.20

Additionally, while intuitive, we probe whether there are specific predictors of when and why a mixed image may be hard to label – e.g., perhaps images which are naturally ambiguous become even more muddled when combined. We use the entropy of the `CIFAR-10H` labels as a measure of image "ambiguity"[Peterson et al., 2019, Battleday et al., 2020]. Recall, `CIFAR-10H` labels are constructed from many annotator's judgments about the most probable image category; entropy is therefore computed over the frequencies of these class selections and captures some sense of the amount of disagreement between annotators.

We compare humans' elicited confidence in their mixing coefficient, and the amount of relabeling ($|\lambda_h - \lambda_f|$) against the entropy of the `CIFAR-10H` labels of the images being combined. We find in Fig. 8 that if both endpoints are high entropy under `CIFAR-10H` (where we consider "high" being entropy ≥ 0.5), participants report markedly lower confidence in their inference than if both endpoints have low entropy (entropy ≤ 0.1). However, we do not find a significant effect of endpoint entropy and amount of relabeling. This suggests that the ambiguity of the underlying images being mixed plays some role in determining when the resulting synthetic image may be hard to label, but there remains a question as to what can predict high amounts of relabeling from participants. We leave these questions for future investigation.

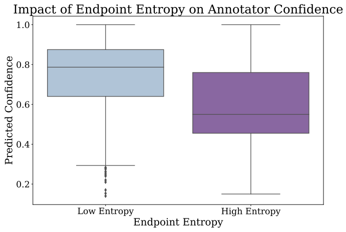


Figure 8: Uncertainty reported by annotators in their inference of λ, as a factor of whether the combined labels y_i, y_j are high or low entropy. Entropy is measured over the `CIFAR-10H` human-derived labels.

We can go further in the study of human uncertainty over *mixup* examples by directly eliciting soft labels from each individual over the entire space of possible classes, inspired by [Collins et al., 2022a]. We include a preliminary investigation into eliciting richer forms of human uncertainty over *mixup* examples, which lend additional nuance to the discrepancy between human perceptual judgments and the synthetic labels classically used in *mixup*, in the Supplement. In particular, our primary finding is that people sometimes place probability mass on classes which are *different* from the endpoint classes being combined (see Figure 9). We include the elicited soft labels in our release of `H-Mix`; these soft labels are small-scale at present (from $N = 8$ participants, see Supplement), and we have not yet explored their computational implications, but see grounds for leveraging richer forms of human uncertainty in this vein as ripe for future work.

Takeaways Our dataset, `H-Mix`, highlights discrepancies between humans' internal models of synthetically generated data compared to what is traditionally used in *mixup*. We

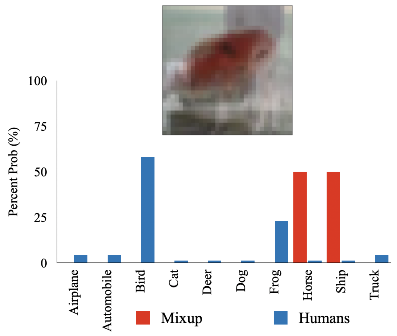


Figure 9: Example combined image ($\lambda_f = 0.5$; horse/ship) which has been relabeled by humans (blue) through the elicitation of individual soft labels our soft label elicitation (see Supplement); here, we average the individual soft labels derived from the two different participants who annotated the image. The label which would be used by *mixup* is shown in red.

observe variable labeling policies on a category-pair basis and uncover a likely relationship between the strength of the mixing coefficient and ambiguity of the underlying combined images with participants' reported uncertainty in their judgments. We also preview even richer discrepancies between human percepts and the functional form of the *mixup* mixing policies (see Supplement).

5 LEARNING WITH HUMAN RELABELINGS *AND UNCERTAINTY*

In addressing RQ1 and RQ2, this work illuminates that human perceptual judgments do not consistently recover the parameters of the generative model traditionally used to construct data in *mixup*. These findings beg the question: if we instead align synthetic examples with human perceptual judgments, how would this impact model performance? Such a question is important to consider in the pursuit of more trustworthy ML systems: better generalization, robustness, calibration, and a richer understanding of whether the models are trained on human-aligned data could all potentially engender more stakeholder trust [Zerilli et al., 2022].

To that end, we consider two initial empirical studies of the impact of training on human perceptual judgments of synthetic examples: one, wherein we compare training models with varied forms of labels on the specific set of 2070 mixed images from `H-Mix`, and another where we go beyond the collected examples and consider a first attempt at constructing a generic human-aligned label mixing policy. Here, we focus on the data collected for RQ2; i.e., for given $\tilde{x}$ how should we change $\tilde{y}$. We encourage leveraging and scaling the data collected in RQ1 for future work.

5.1 RELABELING DIRECTLY WITH H-MIX

Setup We train a PreAct ResNet-18 [He et al., 2016] and VGG-11 [Simonyan and Zisserman, 2014] over 7,000 regular `CIFAR-10` images (following the split used by [Collins et al., 2022a]) combined with the 2,070 synthetically mixed images where we vary the labels. While we would ideally study human relabelings for every synthetic image that could be generated with f, we only have labels for a small subset and instead compare using our labels versus traditional *mixup* labels over a *finite, augmenting set* of combined images. 5 seeds are run per variant per architecture. Results are averaged across architectures.

Evaluation We evaluate a suite of metrics over 3,000 examples from `CIFAR-10H`, a dataset containing labels from many humans over the `CIFAR-10` test set [Peterson et al., 2019]. We compare: cross entropy between the model-predicted and the human-derived label distributions (CE), model calibration following [Hendrycks et al., 2022] and robustness to the Fast Gradient Sign Method (FGSM) adversarial attack [Goodfellow et al., 2014b], again following the set-up of [Collins et al., 2022a].

Leveraging Human Relabelings for ML Training We first compare learning with our averaged human-inferred mixing parameters against the classical *mixup* labels over the same 2070 synthetically-mixed images. We include sanity checks with completely random and uniform labels for the synthetic examples, as well as a baseline not including any synthetic examples ("No Aug"). Interestingly, we find in Table 2 that aligning the mixed example labels with averaged human labels yields *worse* model performance. We think these results are worth highlighting: it is not always the case that aligning models to human perception yields performance gains, possibly due to the recently discovered U-shaped relationship between representational alignment and generalization [Sucholutsky and Griffiths, 2023].

The Value of Human Uncertainty Information However, the human-inferred λ_g alone does not capture the richness of human perceptual judgments over synthetic images: participants at times reported being uncertain in their inferences. Therefore, we account for human uncertainty (ω) in the inference of the synthetic data generating parameter to construct softer $\tilde{y}$ (see Supplement for details). We find substantial performance boosts come from leveraging human uncertainty. Such data suggest that indeed, aligning models in accordance with human perceptual inferences could have advantages – and suggests that confidence could offer a potent modulator signal worth considering eliciting. This is in line with core ideas from Dempster-Shafer Theory [Shafer, 1976, Dempster, 1967], that soft labels should be expressed as one set of values representing the mixture weights, and a second associated set of values representing uncertainty about that mixture (i.e., belief and plausibility).

Table 2: Comparing performance when varying the form of the synthetic labels on the 2070 mixed images. Results averaged over 5 seeds, with error bars depicting 95% confidence intervals (CIs) across seed performance.

Label Type	CE	FGSM	Calib
Regular			
(No Aug)	2.02±0.12	13.12±2.65	0.28±0.011
+ Random	2.11±0.13	12.81±2.84	0.24±0.014
+ Uniform	2.16±0.14	12.71±2.79	0.25±0.012
+ *mixup*	1.65±0.11	10.62±2.44	0.23±0.005
+ Ours			
(Relabel)	1.78±0.12	11.69±2.90	0.24±0.009
(Relabel & ω)	**1.48±0.06**	**8.89±1.59**	**0.19±0.001**

5.2 GENERALIZING RELABELING

So far, we have focused on varying the labels of a presupposed augmenting set of mixed images; however, the set was comparatively small (2070 images) and therefore does not directly mimic the *mixup* learning paradigm. In practice, *mixup* is typically applied over the entire dataset; that is, on each batch, a new mixing coefficient is sampled, resulting in often entirely new images being generated per batch. It is infeasible to consider recruiting human participants to relabel every such image. Automated human-aligned labeling policies are therefore worth considering. We argue that our data offers a prime starting point to explore such questions.

We offer a preliminary alternative label mixing policy based on the human data we have collected in `H-Mix`. Inspired by the non-linearities we observe at a category level, we use `scipy.curve_fit` to fit a logistic function per category pair. For each batch, we swap in our label mixing policy to map from the sampled generating mixing coefficient to an approximately more human-perceptually aligned coefficient. Such fits only account for humans' relabelings, not their uncertainty. Accounting for human confidence in automated label policies is a ripe direction for future work.

Setup We follow the same ensembling and evaluation methodology laid out in Section 5.1, but now run traditional *mixup* following [Zhang et al., 2018] where generating mixing coefficients are sampled from a $Beta(1, 1)$ distribution (i.e., uniform on $(0, 1)$).

Results We observe (see Table 3) a striking parity in performance across models. These data highlight that with the addition of even a relatively small number of human annotations through `HMix` to alter the labeling policy, we find that robustness to adversarial attacks increases at negligble cost to performance or calibration. As in [Sucholutsky and Griffiths, 2023], human representation alignment may be useful for other downstream, untested tasks: training on more human-aligned data-generating policies could induce functional fits that are preferable to stakeholders even if we see no objective improvement along particular performance measures. We recommend such studies for future work.

Table 3: Training with mixing policies fitted per category pair, compared against full *mixup*. Results averaged over 5 seeds, with 95% CI error bars.

Label Policy	CE	FGSM	Calib
mixup	**1.15±0.08**	7.46±2.40	**0.10±0.01**
Human-Fits (Ours)	1.16±0.08	**7.32±2.27**	**0.10±0.01**

Takeaways Human perceptual judgments can be leveraged to construct alternative synthetic data-generating policies to train ML systems; however, such induced methods of aligning with (approximations) of human perception are not automatic salves. Our results highlight that constructing more human-aligned label policies, particularly through capturing and representing human uncertainty, is promising, but more work is needed before generalizing conclusions.

6 DISCUSSION

(Mis)alignment of Mixup Examples Through a series of novel user studies, we uncover that the synthetic examples used in *mixup* do not consistently align with humans' perceptual representations. We find indications that participants' *uncertainty* in their inferred mixing coefficients tracks with the degree of ambiguity of the original images that are combined. As we have begun to explore empirically, such relabeling may impact downstream model performance: realigning mixup labels with humans' reported judgments can impact learning, with human uncertainty seemingly poised to provide a strong supervisory signal – corroborating [Peterson et al., 2019, Collins et al., 2022b, Sucholutsky et al., 2023]. The collation of humans' inferences of the *mixup* generative parameters could also be used to benchmark whether models are aligned with human percepts, say if `H-Mix` is used as a held-out or probe set [Gruber et al., 2018]. We recommend such directions for future work, particularly those focused on the uncertainty elicitation in `H-Mix`. We release additional soft labels over mixed examples which further highlight human perceptual misalignment (see Supplement).

Scaling Human-Centric Data Relabeling A key challenge for human-centric relabeling of synthetically-generated data (not unique to *mixup*) is that a nearly infinite variety can be generated. It is not reasonable to expect humans to judge *all* possibilities, nor to provide their uncertainty over all labels.Any attempt at human-in-the-loop relabeling faces the obstacle of identifying which examples to relabel, and how to handle cases that cannot be relabeled. While we take steps to address the latter through fitting generic functions per class pair that enable sampling of arbitrary mixing coefficients, we highly encourage researchers to consider leveraging our H-Mix to develop alternative human-grounded automated synthetic data policies.

To address the former, we encourage looking to smarter ways to select examples to query people over – rather than random selection as we have done – such as [Liu et al., 2021a, 2017]. Additionally, our results raise the related question: are there particular relabelings that are *hurting* model performance? Prior works have demonstrated how cleaning data can reduce model error [Pleiss et al., 2020]. We encourage future work in this direction in the context of H-Mix. Additionally, our results raise the related question: are there particular relabelings that are *hurting* model performance? Prior works have demonstrated how cleaning data can reduce model error [Pleiss et al., 2020]. We encourage future work in this direction in the context of H-Mix.

Limitations Thus far, we only consider human validation and relabeling of *mixup* labels for a single image classification dataset, CIFAR-10. This dataset is low-resolution. Thus, the endpoint images – and the combinations of images – can be ambiguous and challenging to interpret. It is possible that we may find humans to be more, or less, aligned with the generative parameters for different image datasets, or for entirely different data modalities, e.g., audio or video. We encourage the application of the HILL MixE Suite paradigm to other datasets. Moreover, as we have many category pairs – arising even from just 10 categories – we do not have a substantial number of synthetic examples *per* category pair (i.e., 46 synthetically-mixed images for each of the 45 category pairs). This could impact the stability of the category boundaries we elicit, e.g., potentially leading to breaks of monotonicity (see Supplement A). Further, as with many web-based human elicitation studies, it is not always clear whether the responses returned arise from individual differences in perception, participant noise, or malicious behavior [Lease, 2011, Gadiraju et al., 2015]. We also do not train participants to provide calibrated uncertainty; uncertainty judgments included in H-Mix – while empirically useful for training – could be infused with classical biases in humans' probabilistic self-reports [Lichtenstein et al., 1977, Tversky and Kahneman, 1996, O'Hagan et al., 2006, Sharot, 2011]. We also highlight that, aside from repeat trials, we are unable to capture whether participants' percepts fluctuate – such instability is certainly a possibility when considering cognitive neuroscience research around perceptual dominance [Blake and Logothetis, 2002].

Extending to New Synthetic Data Paradigms In this work, we focused on the synthetic data classically used in *mixup*, as the simplicity of the data generating process – a single mixing coefficient parameter – enables us to precisely compare human versus traditional parameterizations of the synthetic data construction process. We hope our work spurs further study of aligning synthetic data generation with human perception and motivates the design of more human-aligned synthetic data to improve ML systems, particularly those focused on the interplay between model and human *uncertainty*. We release the code of all interfaces included in our HILL MixE Suite, which we hope will empower researchers with additional tools to investigate humans' percepts over synthetically-constructed data. For instance, our Select-Shuffled interface could readily be extended to elicit stakeholders' preferences, in the form of selection, over any collection of constructed synthetic examples. As demonstrated in [Ouyang et al., 2022], scalable human preference elicitation has wide utility.

7 CONCLUSION

Through a series of human participant elicitation studies, we find that the synthetic examples generated via *mixup* differ in fundamental ways from human perception, suggesting misalignment of the data and label mixing policies. We offer early indications that collating humans' percepts of these synthetic examples could impact model performance, particularly when modulated *by elicited human uncertainty*. Our work further motivates the design of automated relabeling procedures for synthetic examples which leverage elicited human data (e.g., training a model to predict a likely human's mixing coefficient) to sidestep inherent issues with scaling human annotation over the space of possible synthetic examples, particularly in eliciting and utilizing human uncertainty. Synthetic data of all kinds are proliferating: we encourage more researchers to consider these data from a human-centric perspective; i.e., investigating whether the samples align with human percepts, and if not, whether altering labels – specifically via human uncertainty – can yield safer, more reliable models with improved generalization.

ACKNOWLEDGMENTS

We thank (alphabetically) Federico Barbero, Alan Clark, Krishnamurthy (Dj) Dvijotham, Benedict King, Tuan Anh Le, Haitz Sáez de Ocáriz Borde, N. Siddharth, Joshua Tenenbaum, Richard E. Turner, Vishaal Udandarao, and Lio Wong for helpful discussions. We also thank our participants on Prolific, and our reviewers for incredibly helpful feedback.

KMC is supported by a Marshall Scholarship. UB acknowledges support from a JP Morgan AI PhD Fellowship and from DeepMind and the Leverhulme Trust via the Leverhulme Centre for the Future of Intelligence (CFI). BL acknowledges support under the ESRC grant ES/W007347, and AW acknowledges support from a Turing AI Fellowship under grant EP/V025279/1, The Alan Turing Institute, and the Leverhulme Trust via CFI. IS is supported by an NSERC fellowship (567554-2022).

References

R. M. Battleday, J. C. Peterson, and T. L. Griffiths. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature communications*, 11(1):1–14, 2020.

R. Blake and N. K. Logothetis. Visual competition. *Nature Reviews Neuroscience*, 3(1):13–21, 2002.

V. Chen, U. Bhatt, H. Heidari, A. Weller, and A. Talwalkar. Perspectives on incorporating expert feedback into model updates. *arXiv preprint arXiv:2205.06905*, 2022.

C.-Y. Chuang and Y. Mroueh. Fair mixup: Fairness via interpolation. In *ICLR*, 2020.

J. J. Y. Chung, J. Y. Song, S. Kutty, S. R. Hong, J. Kim, and W. S. Lasecki. Efficient elicitation approaches to estimate collective crowd answers. In *CSCW*, 2019.

K. M. Collins, U. Bhatt, and A. Weller. Eliciting and learning with soft labels from every annotator. In *HCOMP*, 2022a.

K. M. Collins, C. Wong, J. Feng, M. Wei, and J. B. Tenenbaum. Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. *arXiv preprint arXiv:2205.05718*, 2022b.

C. M. de Melo, A. Torralba, L. Guibas, J. DiCarlo, R. Chellappa, and J. Hodgins. Next-generation deep learning based on simulators and synthetic data. *Trends in Cognitive Sciences*, 26(2):174–187, 2022.

A. P. Dempster. Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics*, 38(2):325 – 339, 1967. doi: 10.1214/aoms/1177698950. URL https://doi.org/10.1214/aoms/1177698950.

N. Destler, M. Singh, and J. Feldman. Shape discrimination along morph-spaces. *Vision Research*, 158:189–199, 2019.

P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021.

Z. Emam, A. Kondrich, S. Harrison, F. Lau, Y. Wang, A. Kim, and E. Branson. On the state of data in computer vision: Human annotations remain indispensable for developing deep learning models. *arXiv preprint arXiv:2108.00114*, 2021.

T. Fel, I. Felipe, D. Linsley, and T. Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

J. R. Folstein, T. J. Palmeri, and I. Gauthier. Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex*, 23(4):814–823, 2013.

U. Gadiraju, R. Kawase, S. Dietze, and G. Demartini. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *CHI*, 2015.

R. L. Goldstone and A. T. Hendrickson. Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1):69–78, 2010.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014a.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

L. Z. Gruber, A. Haruvi, R. Basri, and M. Irani. Perceptual dominance in brief presentations of mixed images: Human perception vs. deep neural networks. *Frontiers in Computational Neuroscience*, 12:57, 2018.

S. Harnad. Categorical perception. In *Encyclopedia of Cognitive Science*, volume 67. MacMillan: Nature Publishing Group, 2003.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. pages 770–778, 2016.

D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *ICLR*, 2019.

D. Hendrycks, A. Zou, M. Mazeika, L. Tang, B. Li, D. Song, and J. Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *CVPR*, 2022.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen, and A. Weller. Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*, 2022.

D. Kaushik, E. Hovy, and Z. C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

A. Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

M. Lease. On quality control and machine learning in crowdsourcing. In *AAAI Workshops*, 2011.

S. Lichtenstein, B. Fischhoff, and L. D. Phillips. Calibration of probabilities: The state of the art. *Decision making and change in human affairs*, pages 275–324, 1977.

W. Liu, B. Dai, A. Humayun, C. Tay, C. Yu, L. B. Smith, J. M. Rehg, and L. Song. Iterative machine teaching. In *ICML*, 2017.

W. Liu, Z. Liu, H. Wang, L. Paull, B. Schölkopf, and A. Weller. Iterative teaching by label synthesis. In *NeurIPS*, 2021a.

Z. Liu, S. Li, D. Wu, Z. Chen, L. Wu, J. Guo, and S. Z. Li. Unveiling the power of mixup for stronger classifiers. *arXiv preprint arXiv:2103.13027*, 2021b.

V. Nanda, A. Majumdar, C. Kolling, J. P. Dickerson, K. P. Gummadi, B. C. Love, and A. Weller. Exploring alignment of representations with human perception. *arXiv preprint arXiv:2111.14726*, 2021.

J. E. Oakley and A. O'Hagan. Shelf: the sheffield elicitation framework (version 2.0). *School of Mathematics and Statistics, University of Sheffield, UK*, 2010.

A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain Judgements: Eliciting Expert Probabilities*. John Wiley, Chichester, 2006.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, 2022. URL `https://arxiv.org/abs/2203.02155`.

S. Palan and C. Schitter. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky. Human uncertainty makes classification more robust. In *ICCV*, 2019.

G. Pleiss, T. Zhang, E. Elenberg, and K. Q. Weinberger. Identifying mislabeled data using the area under the margin ranking. *NeurIPS*, 2020.

D. Prelec. A bayesian truth serum for subjective data. *Science*, 306(5695):462–466, 2004.

K. Sanders, R. Kriz, A. Liu, and B. Van Durme. Ambiguous images with human judgments for robust visual event classification. In *NeurIPS*, 2022.

G. Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

T. Sharot. The optimism bias. *Current biology*, 21(23): R941–R945, 2011.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

I. Sucholutsky and T. L. Griffiths. Alignment with human representations supports robust few-shot learning. *arXiv preprint arXiv:2301.11990*, 2023.

I. Sucholutsky, R. M. Battleday, K. M. Collins, R. Marjieh, J. Peterson, P. Singh, U. Bhatt, N. Jacoby, A. Weller, and T. L. Griffiths. On the informativeness of supervision signals. *UAI*, 2023.

S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *NeurIPS*, 2019.

A. Tversky and D. Kahneman. On the reality of cognitive illusions. *Psychological Review*, 103(3):582–591, 1996.

V. Verma, S. Mittal, W. H. Tang, H. Pham, J. Kannala, Y. Bengio, A. Solin, and K. Kawaguchi. Mixupe: Understanding and improving mixup from directional derivative perspective, 2022. URL `https://arxiv.org/abs/2212.13381`.

J. Zerilli, U. Bhatt, and A. Weller. How transparency modulates trust in artificial intelligence. *Patterns*, 3(4):100455, 2022.

H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

L. Zhang, Z. Deng, K. Kawaguchi, A. Ghorbani, and J. Zou. How does mixup help with robustness and generalization? In *ICLR*, 2020.

L. Zhang, Z. Deng, K. Kawaguchi, and J. Zou. When and how mixup improves calibration. In *ICML*, 2022.