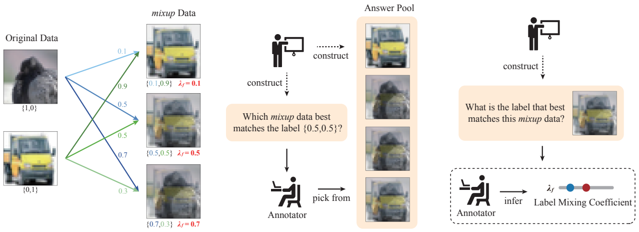
Human-in-the-loop *mixup*

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

[1] University of Cambridge [2] The Alan Turing Institute [3] Max Planck Institute for Intelligent Systems [4] University College London
{kmc61, usb20, wl396, vp421}@cam.ac.uk, b.love@ucl.ac.uk, aw665@cam.ac.uk

Abstract

Synthetic data is proliferating and powering many advances in machine learning. However, it is not always clear if synthetic labels are perceptually sensible to humans. The web provides us with a platform to take a step towards addressing this question from a human-centric perspective, through online elicitation. We design a series of elicitation interfaces, which we release as `HILL MixE Suite`, and recruit 159 participants, to provide perceptual judgments over the kinds of synthetic data constructed during *mixup* training: a powerful regularizer shown to improve model robustness, generalization, and calibration. We find that human perception does 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. We release all elicited judgments in a new data hub we call `H-Mix`.


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

(a) How *mixup* data is constructed

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

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

Figure 1: Overview of our framework. 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.

Introduction

Synthetic data is proliferating, fueled by increasingly powerful generative models, e.g. (Goodfellow et al. 2014; Dhariwal and Nichol 2021). These data are not only consumed directly by people (e.g., users of the web) – 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.

A human-centric stance on synthetically-generated data may be important for many reasons. Since synthetic examples increasingly form the crux of training data, it is worth considering whether such labels reflect human beliefs. Aligning networks to match humans' perceptual inferences could be a way to further ensure model reliability and trustworthiness (Nanda et al. 2021; Chen et al. 2022; Fel et al. 2022). 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 models with human probabilistic knowledge (Collins, Bhatt, and Weller 2022; Sanders et al. 2022). More broadly, connections to cognitive science can improve our understanding of algorithms for generative models (de Melo et al. 2022; Chandra et al. 2022; Marjieh et al. 2022). Additionally, an improper understanding of how synthetic data is generated could leave users open to manipulation or gamification (Brundage et al. 2018). 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 like a generative adversarial network (GAN) (Goodfellow et al. 2014) or a diffusion model (Ho, Jain, and Abbeel 2020) does not permit these kinds of precise comparisons as easily. 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, Palmeri, and Gauthier 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 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. 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. 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 which powers many of the ML algorithms on the web.

Problem Formulation

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.

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 reason. 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 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 $\tilde{y}$?

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 Appendix.

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, Hovy, and Lipton 2019). Such an approach will let us begin to study how humans perceive the data and labeling generative policies used in *mixup*.

Problem Setting

In our set-up, 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.

Elicitation Paradigm

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

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

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 given 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.

Why do we consider both interfaces? We reason that the first interface could be prone to ordering effects – an astute participant could realize that they can count out where the midpoint is located. This led us to design the second variety (Select-Shuffled) wherein the participant sees all images shuffled simultaneously. We hypothesize that Construct could induce responses biased by the participant's starting position. To probe this, we run two subvariants wherein participants start from either $\lambda_f = 0.1$ or $\lambda_f = 0.9$.

Example interfaces are depicted in the Appendix. As mentioned, participants are explicitly told the categories being combined (y_1, y_2) and are asked to indicate the image they think that 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, Peterson, and Griffiths 2020). From each 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 incrementing by 0.1 between $\lambda_f = 0.0$ and $\lambda_f = 1.0$ (totalling 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 sub-divided 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 Appendix.

Investigating Data Mixing Alignment

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

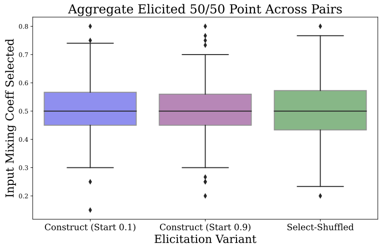


Figure 2: Averaging humans participants' selections per image pair reveal the typical image pair is minimally relabeled.

To our surprise, we find remarkable agreement in the averaged selections per image pair across interface varieties; the median difference in the aggregated selections per image pair is approximately $0.05 - 0.06$ across pairwise comparisons of interfaces. Interestingly, this is equivalent to the median standard deviation amongst annotators per interface. We also do not find a strong effect of starting position for Construct. This is encouraging – and suggests that our general framework is somewhat robust to interface structure.

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 images pairs which 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 stimuli set, more such examples can be recovered. Note, there are a total of 101 image pairs which 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.

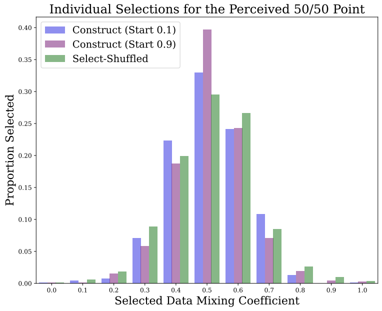


Figure 3: Participants do not always endorse the 50/50 point suggesting misalignment in the data labeling policy. Bar plot depicts extracted mixing coefficient of individuals' selections for the perceptually-aligned midpoints.

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.

Elucidating Alignment of the Label Mixing Policy (RQ2)

The above elicitation 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.

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 $\tilde{x}$.

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 *confidence* 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.

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, result in 2070 total stimuli. Each mixed image is formed by constructed by selecting a data mixing coefficient $\lambda_f \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$.

Human subject experiment We run our relabeling experiment on $N = 81$ participants again through Prolific (Palan and Schitter 2018). Further details are included in the Appendix.

Validating the Generating Mixing Coefficient against Human Responses

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

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 data mixing coefficients. We observe – in aggregate – a non-linear, roughly sigmoidal structure to humans' relabelings, consistent with past research in human categorical perception (Harnad 2003; Goldstone and Hendrickson 2010; Folstein, Palmeri, and Gauthier 2013; Destler, Singh, and Feldman 2019). The aggregate recovery of the 50/50 point corroborates our findings in RQ1. However, as we highlighted in Section 3, we find that the picture is nuanced: wide confidence bounds illustrate that there are mixed images for which participants' inferred mixing coefficients are substantially different to the parameterization assumed in *mixup*. Further, qualitative inspection of example averaged relabelings for particular images (Fig. 6) – and across category pairs (Fig. 7) – reveals such misalignment. We recommend future work to better study why particular category pairs, for this dataset, are yielding different boundaries.

Takeaways Our dataset, H-Mix, highlights discrepancies between humans' internal models of synthetically generated data compared to what is traditionally used in *mixup*. We observe variable labeling policies on a category-pair basis and uncover a likely relationship between the ambiguity of the combined images and participants' reported confidence in their judgments (see Appendix).

Exploring the Impact of Learning with Human Relabelings

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 the synthetic examples with human perceptual judgments, how does this impact model performance? Such

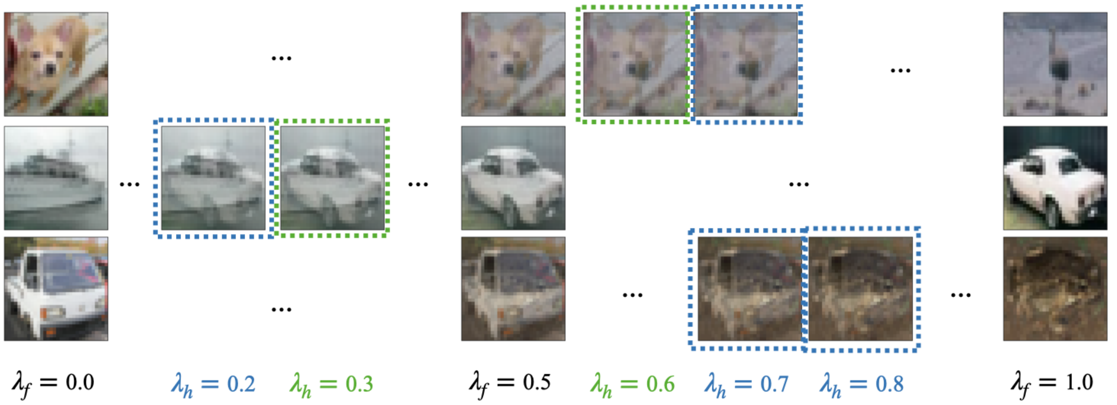


Figure 4: Example image pairs where substantial relabeling of the 50/50 point was recommended across all interface types. Synthetic images highlighted in blue received the most endorsements from participants across all interface types, with images in green receiving second most. For row three, participants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

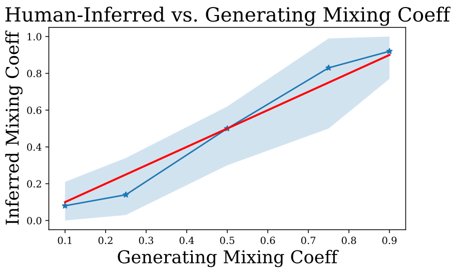


Figure 5: We uncover a sigmoidal relationship between humans' inferred mixing coefficient (λ_h, blue) as compared to the mixing coefficient used to generate the image (λ_g, red) suggestive of misalignment. We depict the median, along with the 25th and 75th percentiles. The red line indicates what exact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

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 at least trained on human-aligned data could all potentially engender more stakeholder trust (Zerilli, Bhatt, and Weller 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 an 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.

Relabeling Directly with `H-Mix`

Setup We train a ResNet and VGG variety (PreAct ResNet-18 (He et al. 2016), and VGG-11 (Simonyan and Zisserman 2014)) over 7,000 regular `CIFAR-10` (following the split used by (Collins, Bhatt, and Weller 2022)) combined with the 2070 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 model 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, Shlens, and Szegedy 2014), again following the set-up of (Collins, Bhatt, and Weller 2022)

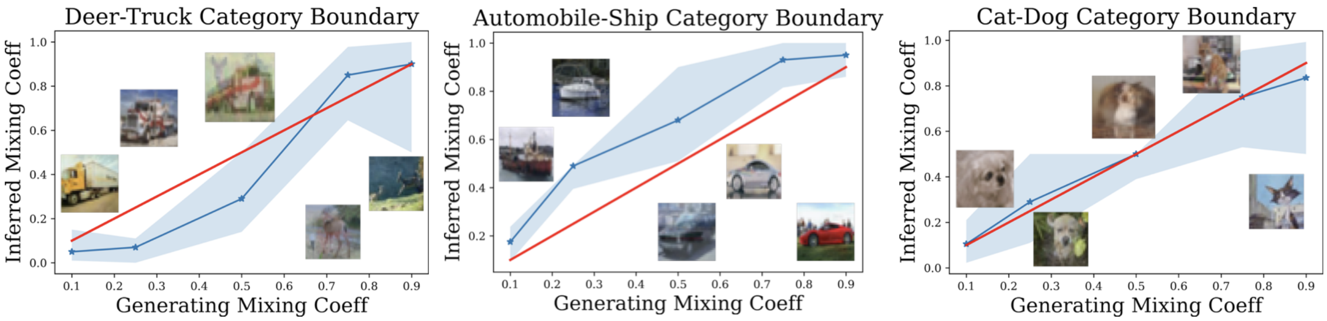


Figure 7: "Category boundaries" elicited from humans display a diverse structure. Many – though not all – deviate from linearity assumed in *mixup*. We overlay example synthesized stimuli shown to participants, ordered by the λ_f used to create them.

Leveraging Human Relabelings for ML Training We first compare learning with our averaged human-inferred mixing parameters against using 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 1 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 measurable performance gains.

However, the human-inferred λ_g alone does not capture the richness of humans' perceptual judgments over the synthetic images: participants at times reported being uncertain in their inferences. Therefore, we next account for human uncertainty (ω) in the inference of the synthetic data generating parameter to construct softer $\tilde{y}$ (see Appendix for details). We find substantial performance boosts fall out of leveraging human confidence. 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.

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 confidence. 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 2) a striking parity in performance across models. These data highlight that constructing more human-aligned data simulators is not necessarily a harm to downstream performance and perhaps could be beneficial. Note, we are only looking at performance on a small set of possible metrics, and a relatively small set of held-out data ($3,000$ examples). It is quite feasible that 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.

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

Related Work

Our work connects most closely to human-in-the-loop data augmentation and the expansive literature surrounding human categorical perception from the cognitive science community, as well as ongoing efforts in the machine learning community to develop more efficacious *mixup*-based data and label mixing functions.

Table 1: Comparing performance when varying the form of the synthetic labels on the 2070 mixed images.

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

Table 2: Training with mixing policies fitted per category pair, compared against full *mixup*.

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**

Human-in-the-Loop Data Augmentation

Incorporating expert feedback into the learning procedure has received increasing attention (Chen et al. 2022). In particular, previous work has considered incorporating humans "in the loop" for data augmentation. For instance, Dataset-GAN (Zhang et al. 2021) employs human participants to label GAN-generated images and feeds these back to the model to generate more synthetic data. (Kaushik, Hovy, and Lipton 2019) similarly incorporate human feedback by having humans *create* counterfactual samples, and has been shown to be an efficient method to adjust model behavior (Kaushik et al. 2021). Other works have considered employing humans to provide "rationales" about examples to improve data-efficiency and downstream modeling performance (Zaidan, Eisner, and Piatko 2007). Here, we marry these ideas in the context of *mixup* by eliciting data and label mixing function parameters to align with human percepts.

Human Categorical Perception

In cognitive science, eliciting humans' judgments over synthetically-constructed examples is a tried-and-true method to characterize human category boundaries (Newell and Bülthoff 2002; Folstein, Palmeri, and Gauthier 2013; Feldman 2021; Folstein, Gauthier, and Palmeri 2012). Such studies often reveal a non-linear structure of humans' percepts. For instance, in the audio domain, the identification of vowel categories has been found to demonstrate "warping" close to prototypical category members – known as the "perceptual magnet effect" (Kuhl 1991; Feldman, Griffiths, and Morgan 2009). Similar nonlinearities have been found in the perception of boundaries between face identities (Beale and Keil 1995) and the transitions between 3D shapes (Newell and Bülthoff 2002; Destler, Singh, and Feldman 2019). Our linearly interpolated stimuli are similar in spirit to the morphological trajectories used in these works, as well as other synthetically-combined images (Oliva, Torralba, and Schyns 2006). (Gruber et al. 2018) also consider 50/50 mixed images; however, their elicitation involves open-ended judgments which does not permit the same kind of data and label mixing alignment studies as our methods more directly eliciting human-inferred generative parameters. Our work also connects to other non-linear perceptual phenomena encountered in the visual domain; namely, binocular rivalry, whereby present participants with a different image in each eye has been shown to induce oscillatory percepts (Blake and Logothetis 2002; Tong, Meng, and Blake 2006).

Other *mixup*-Based Synthetic Data Schemes

Many alternative *mixup* data and label mixing functions have been proposed (Verma et al. 2019; Yun et al. 2019; Kim, Choo, and Song 2020; Kim et al. 2020; Hendrycks et al. 2022). Closest to our work, (Sohn et al. 2022) highlight particular issues with the linear interpolation in label space on the learned topology of the model's category boundaries and instead utilize a Gaussian Mixture Model (GMM)-based relabeling scheme to construct "better" labels than those used in baseline *mixup*. Additional work on learning better pseudo-labels over *mixup* samples have been proposed (Arazo et al. 2020; Cascante-Bonilla et al. 2020; Sohn et al. 2020; Qiu et al. 2022). Similarly, Between-class (BC) learning (Tokozume, Ushiku, and Harada 2017, 2018) proposes hand-crafted adjustments to label construction to better align with human perception based on waveform modulations; however, to our knowledge, no previous works have *directly* considered incorporating humans in-the-loop for either the construction of *mixup* samples, or associated relabeling.

Discussion

(Mis)alignment of Synthesized Examples

Through a series of user studies, we uncover that human perception of the synthetic images and corresponding labels constructed in *mixup* does not consistently align with the generative parameters often used to form said synthetic examples. We find indications that participants' *confidence* 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: re-aligning mixup labels with humans' reported judgments can impact learning, with human confidence seemingly poised to provide a strong supervisory signal. 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.

Scaling Human-Centric Data Relabeling

A key challenge for human-centric relabeling of synthetically-generated data (not unique to *mixup*) is that a near infinite variety can be generated. It is not reasonable to expect humans to judge *all* possibilities. Any attempt then at human-in-the-loop relabeling faces the obstacle of identifying which examples to relabel, and how to handle cases which 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 use for querying people – rather than random selection as we have done – such as (Liu et al. 2021a, 2017), could be beneficial. 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 Appendix 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 have calibrated confidence; confidence judgments included in H-Mix – while empirically useful for training – could be infused with classical biases in humans' probabilistic self-reports (Lichtenstein, Fischhoff, and Phillips 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). Lastly, we emphasize that all of our studies have considered US-based, English-speaking participants. It is important to leverage the power and breadth of the web to consider how perceptual judgments of a more diverse cohort of humans may be differentially aligned with the mixing parameters (Díaz et al. 2022). We encourage further exploration of individual perceptual differences, and their impact of learning, in our synthetic data paradigm.

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. 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.

Conclusion

Through a series of online 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 human confidence. 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 confidence. Synthetic data of all kinds is proliferating: we encourage more researchers to consider these data from a human-centric perspective, investigating whether such examples align with human percepts, and if not, whether modulating labels and learned representations to better match human percepts can yield better performing models.

References

Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N. E.; and McGuinness, K. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International Joint Conference on Neural Networks*.

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

Beale, J. M.; and Keil, F. C. 1995. Categorical effects in the percxeption of faces. *Cognition*, 57(3): 217–239.

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

Brundage, M.; Avin, S.; Clark, J.; Toner, H.; Eckersley, P.; Garfinkel, B.; Dafoe, A.; Scharre, P.; Zeitzoff, T.; Filar, B.; et al. 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

Cascante-Bonilla, P.; Tan, F.; Qi, Y.; and Ordonez, V. 2020. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. *arXiv preprint arXiv:2001.06001*.

Chandra, K.; Li, T.-M.; Tenenbaum, J.; and Ragan-Kelley, J. 2022. Designing Perceptual Puzzles by Differentiating Probabilistic Programs. In *SIGGRAPH*.

Chen, V.; Bhatt, U.; Heidari, H.; Weller, A.; and Talwalkar, A. 2022. Perspectives on Incorporating Expert Feedback into Model Updates. *arXiv preprint arXiv:2205.06905*.

Chuang, C.-Y.; and Mroueh, Y. 2020. Fair Mixup: Fairness via Interpolation. In *ICLR*.

Chung, J. J. Y.; Song, J. Y.; Kutty, S.; Hong, S. R.; Kim, J.; and Lasecki, W. S. 2019. Efficient Elicitation Approaches to Estimate Collective Crowd Answers. In *CSCW*.

Collins, K. M.; Bhatt, U.; and Weller, A. 2022. Eliciting and Learning with Soft Labels from Every Annotator. In *HCOMP*.

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

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

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

Díaz, M.; Kivlichan, I.; Rosen, R.; Baker, D.; Amironesei, R.; Prabhakaran, V.; and Denton, E. 2022. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *FAccT*.

Emam, Z.; Kondrich, A.; Harrison, S.; Lau, F.; Wang, Y.; Kim, A.; and Branson, E. 2021. On The State of Data In Computer Vision: Human Annotations Remain Indispensable for Developing Deep Learning Models. *arXiv preprint arXiv:2108.00114*.

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

Feldman, J. 2021. Mutual Information and Categorical Perception. *Psychological Science*, 32(8): 1298–1310.

Feldman, N. H.; Griffiths, T. L.; and Morgan, J. L. 2009. The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, 116(4): 752.

Folstein, J. R.; Gauthier, I.; and Palmeri, T. J. 2012. How category learning affects object representations: not all morphspaces stretch alike. *Journal of experimental psychology. Learning, memory, and cognition*, 38 4: 807–20.

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

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

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

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

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

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

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

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

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2019. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *ICLR*.

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

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

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

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

Kaushik, D.; Setlur, A.; Hovy, E. H.; and Lipton, Z. C. 2021. Explaining The Efficacy of Counterfactually-Augmented Data. *ICLR*.

Kim, J.; Choo, W.; Jeong, H.; and Song, H. O. 2020. Co-Mixup: Saliency Guided Joint Mixup with Supermodular Diversity. In *ICLR*.

Kim, J.-H.; Choo, W.; and Song, H. O. 2020. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *ICML*.

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

Kuhl, P. K. 1991. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & psychophysics*, 50(2): 93–107.

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

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

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

Liu, W.; Liu, Z.; Wang, H.; Paull, L.; Schölkopf, B.; and Weller, A. 2021a. Iterative Teaching by Label Synthesis. In *NeurIPS*.

Liu, Z.; Li, S.; Wu, D.; Chen, Z.; Wu, L.; Guo, J.; and Li, S. Z. 2021b. Unveiling the Power of Mixup for Stronger Classifiers. *arXiv preprint arXiv:2103.13027*.

Marjieh, R.; Sucholutsky, I.; Langlois, T. A.; Jacoby, N.; and Griffiths, T. L. 2022. Analyzing Diffusion as Serial Reproduction. *arXiv preprint arXiv:2209.14821*.

Nanda, V.; Majumdar, A.; Kolling, C.; Dickerson, J. P.; Gummadi, K. P.; Love, B. C.; and Weller, A. 2021. Exploring Alignment of Representations with Human Perception. *arXiv preprint arXiv:2111.14726*.

Newell, F. N.; and Bülthoff, H. H. 2002. Categorical perception of familiar objects. *Cognition*, 85(2): 113–143.

Nguyen, Q.; Valizadegan, H.; and Hauskrecht, M. 2013. Learning classification models with soft-label information. *Journal of the American Medical Informatics Association*, 21.

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

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

Oliva, A.; Torralba, A.; and Schyns, P. G. 2006. Hybrid Images. *ACM Transactions on Graphics*, 25(3): 527–532.

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

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

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

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

Qiu, Z.; Liu, W.; Xiao, T. Z.; Liu, Z.; Bhatt, U.; Luo, Y.; Weller, A.; and Schölkopf, B. 2022. Iterative Teaching by Data Hallucination. *arXiv preprint arXiv:2210.17467*.

Sanders, K.; Kriz, R.; Liu, A.; and Van Durme, B. 2022. Ambiguous Images With Human Judgments for Robust Visual Event Classification. In *NeurIPS*.

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

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

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

Sohn, J.-y.; Shang, L.; Chen, H.; Moon, J.; Papailiopoulos, D.; and Lee, K. 2022. GenLabel: Mixup Relabeling using Generative Models. *arXiv preprint arXiv:2201.02354*.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *NeurIPS*.

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

Tokozume, Y.; Ushiku, Y.; and Harada, T. 2017. Learning from between-class examples for deep sound recognition. *arXiv preprint arXiv:1711.10282*.

Tokozume, Y.; Ushiku, Y.; and Harada, T. 2018. Between-class learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5486–5494.

Tong, F.; Meng, M.; and Blake, R. 2006. Neural bases of binocular rivalry. *Trends in cognitive sciences*, 10(11): 502–511.

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

Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; and Bengio, Y. 2019. Manifold mixup: Better representations by interpolating hidden states. In *ICML*.

Verma, V.; Mittal, S.; Tang, W. H.; Pham, H.; Kannala, J.; Bengio, Y.; Solin, A.; and Kawaguchi, K. 2022. MixupE: Understanding and Improving Mixup from Directional Derivative Perspective.

Vodrahalli, K.; Gerstenberg, T.; and Zou, J. 2021. Do Humans Trust Advice More if it Comes from AI? An Analysis of Human-AI Interactions. *arXiv preprint arXiv:2107.07015*.

Wei, J.; Zhu, Z.; Luo, T.; Amid, E.; Kumar, A.; and Liu, Y. 2022. To Aggregate or Not? Learning with Separate Noisy Labels. *arXiv preprint arXiv:2206.07181*.

Yun, S.; Han, D.; Chun, S.; Oh, S.; Yoo, Y.; and Choe, J. 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *ICCV*.

Zaidan, O.; Eisner, J.; and Piatko, C. 2007. Using “annotator rationales” to improve machine learning for text categorization. In *NAACL*.

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

Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR*.

Zhang, L.; Deng, Z.; Kawaguchi, K.; Ghorbani, A.; and Zou, J. 2020. How Does Mixup Help With Robustness and Generalization? In *ICLR*.

Zhang, L.; Deng, Z.; Kawaguchi, K.; and Zou, J. 2022. When and How Mixup Improves Calibration. In *ICML*.

Zhang, Y.; Ling, H.; Gao, J.; Yin, K.; Lafleche, J.-F.; Barriuso, A.; Torralba, A.; and Fidler, S. 2021. DatasetGAN: Efficient Labeled Data Factory With Minimal Human Effort. In *CVPR*.

Acknowledgments

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

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.

Analyzing Human Uncertainty

Additionally, while intuitive, we investigate 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, Peterson, and Griffiths 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.

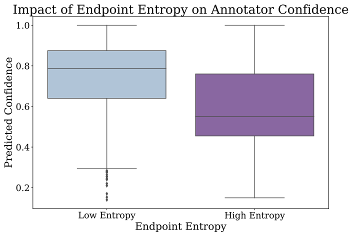


Figure 8: Confidence reported by annotators in their inference of λ, as a factor of whether the combined labels y_i, y_j are high or low entropy. Entropy is measured over the `CIFAR-10H` human-derived labels.

Additional Notes on `H-Mix`

Human Subject Experiments

We include additional details on our human elicitation studies. For all experiments, we require participants speak English as a first-language and reside in the United States.

Elicitation (RQ1) Each participant sees a total of 32 mixed images, where the final two are repeats. Repeats are primarily used here to measure raters’ internal consistency[2]. The median time taken per participant per image as 9.30 and 11.01 seconds for the `Construct` and `Select-Shuffled` interfaces, respectively. A bonus was offered to encourage participants to provide responses which would match what other participants would provide; we applied this bonus to all participants post-hoc resulting in the average participant being paid at a rate of $11.78.

Elicitation (RQ2) Each participant sees $59 - 62$ images, where two images are repeated. Repeats are placed at the end and correspond to the images presented on trials 15 and 20, respectively[3]. The order of the images presented in a batch, as well as the order of the endpoint labels displayed for a given image, are shuffled across participants. We follow the same third-person perspective prompting in Section 3 from (Chung et al. 2019). Participants are asked “what combinations of classes” they thought other participants would say is “used to make” each image, and “how confident” they

[2] Participants’ selections, for each interface type, change by a median of 0.1 in repeat trials, suggesting some inconsistencies in participants’ judgments which persists across elicitation method.

[3] We observe a median difference of 0.03 and 0.05 in the inferred mixing coefficient and confidence on repeat trials, indicative of high intra-annotator consistency.

thought other participants would be in their estimate. Responses are indicated on a slider per question. An example survey screen can be seen in Fig. 12. Subjects took a median of 8.41 seconds per image and were payed at a rate of \$8/hr, with an optional bonus which sought to encourage participants to provide calibrated confidence estimates, similar to that of (Vodrahalli, Gerstenberg, and Zou 2021); the bonus was applied to all participants post-hoc. Each mixed image was seen by at least two different participants each. Our interface is depicted in Fig. 12.

Break from Monotonicty

For users of `H-Mix`, it is worth noting that we do encounter some breaks with monotonicity (see Fig. 9) in a few of the aggregated "category boundaries." We reason this could be in part due to several aspects of our set-up. First, our study involved irregular sampling across the space of mixing coefficients we consider: the 50/50 point is enriched. We ran two phases of elicitation: in the first, we sampled 6 image classes per pair to be shown for three mixing coefficients: 0.5, and one chosen randomly from each of the sets $\{0.1, 0.25\}$ and $\{0.75, 0.9\}$, respectively (810 images of the 2070). All 1260 other images are shown for a single mixing coefficient sampled uniformly from the set. Second, while we have human judgments for over 2000 total images, there are less than 50 synthetic images considered for each category pair, giving any participant noise – or the odd image – greater leverage to impact trends. We encourage others to use HILL-MixE Suite and continue to scale this work and elucidate the stability of the inferred mixing coefficient category boundaries we begin to hint at here.

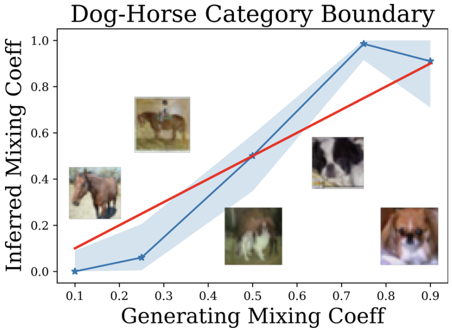


Figure 9: Category boundary elicited from human participants involves a break with monotonicity.

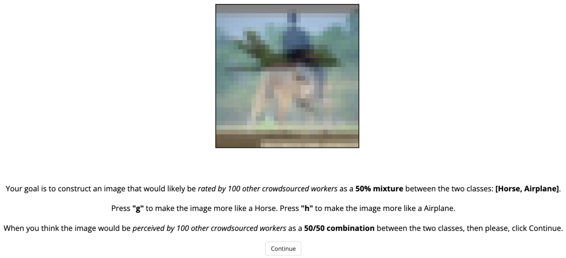
Confidence-Based Smoothing Details

We include further details of our methodology for leveraging human-provided confidence to construct $\tilde{y}$ introduced in Section 5. Human-derived soft labels have been demonstrated to be valuable for learning (Nguyen, Valizadegan, and Hauskrecht 2013; Peterson et al. 2019; Collins, Bhatt, and Weller 2022; Sanders et al. 2022). We transform humans' reported confidence into a smoothing parameter to induce softness using an exponentially-decaying function of human-provided confidence ω: $a * (b^{\omega})$; here, $a = 50, b = 0.0001$. We use the transformed confidence for additive smoothing on the two-category $\tilde{y}$, spread mass accordingly across the full gamut of classes. That is, we use smooth the mass between a completely uniform distribution and a "two-hot" label which uses the human-derived relabeling. Parameters a, b are selected using a held-out set of regular `CIFAR-10` images (from $a \in \{5, 10, 15, 25, 50, 100\}, b \in \{0.00001, 0.0001, 0.001, 0.01, 0.1\}$). We recommend the consideration of alternate smoothing functions, which could, for instance, account for miscalibration in humans' reported confidence.

Further, we compare the impact of learning with aggregated versus de-aggregated participants' predictions. In Section 5, we considered learning with relabelings averaged across participants for a mixed image, and smoothed with confidence reports averaged across participants. Here, we consider instead separating out participants' responses to learn with individual relabelings smoothed by individual confidence, closely related to (Wei et al. 2022). We find in Table 3 that learning with *de-aggregated* data could potentially offer greater performance gains. However, as (Wei et al. 2022) discuss: whether to aggregate can depend on many factors. Our empirical findings support the need for tailoring label construction in context.

Interfaces Included in `HILL MixE Suite`

We display sample pages of the interfaces created and used in this work, which we release as part of `HILL MixE Suite`. Interfaces for Section 3 are shown in Figs. 10 and 11; the interface used Sections 4 is depicted in Fig. 12.

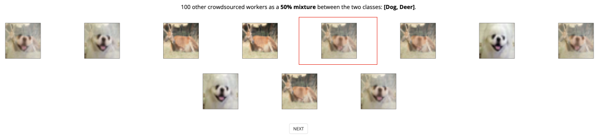


Your goal is to construct an image that would likely be *rated by 100 other crowdsourced workers* as a **50% mixture** between the two classes: **[Horse, Airplane]**.

Press **"g"** to make the image more like a Horse. Press **"h"** to make the image more like a Airplane.

When you think the image would be *perceived by 100 other crowdsourced workers* as a **50/50 combination** between the two classes, then please, click Continue.

Figure 10: Construct interface where participants press arrow keys to select $\tilde{x}$.

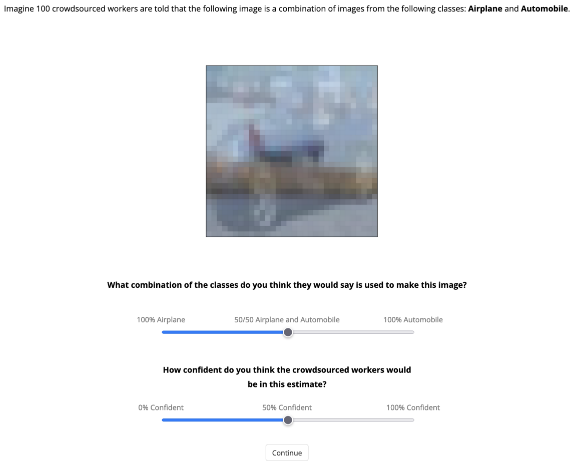


100 other crowdsourced workers as a **50% mixture** between the two classes: **[Dog, Deer]**.

Figure 11: Interface for the selection of a given λ_g from a set of possible mixed images.

Table 3: Varying whether to aggregate when using incorporating human confidence ω in label construction.

Label Type	CE	FGSM	Calib
Ours (Avg with ω)	1.48±0.06	8.89±1.59	**0.19±0.01**
Ours (Separated with ω)	**1.44±0.11**	**8.33±1.92**	**0.19±0.01**



Imagine 100 crowdsourced workers are told that the following image is a combination of images from the following classes: **Airplane** and **Automobile**.

What combination of the classes do you think they would say is used to make this image?

How confident do you think the crowdsourced workers would be in this estimate?

Figure 12: Interface for inferring the *mixup* generative label parameter and providing confidence in such inference.

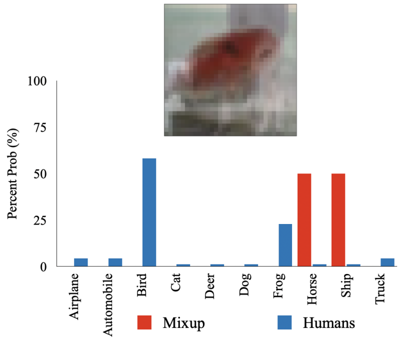


Figure 13: Example combined image ($\lambda = 0.5$; horse/ship) which has been relabeled by humans (blue) using our soft label elicitation. The label which would be used by *mixup* is shown in red.

Alternative Synthetic Example Category Composition Elicitation

Given human participants are uncertain about the underlying mixing coefficient in a number of cases, we consider whether the category composition typically used in *mixup* – e.g., placing mass only on the labels of the images used to form the synthetic combined sample – are reasonable. As demonstrated in Fig 13, the synthetic *mixup* image may look like something else entirely.

We therefore consider a follow-up small-scale human elicitation study wherein we relax the *mixup* assumption that the label mixing function must output a label constructed only from the two classes used to form the mixed image – and instead collect $\tilde{y}$ *directly* by showing the mixed image to human annotators in the form of soft labels. This provides a comparison to the previous human-annotated endpoint label mixing coefficients, and can further inspire useful designs for the label mixing policy.

Study Design

We recruit $N = 8$ participants again from Prolific (Palan and Schitter 2018), yielding soft labels over a total of 100 mixed images. The images are drawn from the same set of stimuli created in Section 4; however, here, we only show images with a mixing coefficient $\in \{0.25, 0.5, 0.75\}$. Participants are told that images are formed by combining other images, and are asked to provide what they think others would see in the image. Participants are asked to specify what others would view as the most probable category with an associated percentage (on a scale of 0-100), an optional second most probable category with a probability, and any categories that would be perceived as definitely not in the image. Again employing the third-person viewpoint framing borrowed from (Chung et al. 2019). We rely on the soft label elicitation interface proposed in (Collins, Bhatt, and Weller 2022) and modify the instructions to be better suited combinations of images. All elicited judgments are included in our data hub, `H-Mix`.

Analyzing Elicited Soft Labels for Synthetic Images

We explore the correspondence between the elicited category compositions of the mixed images with the labels that would be used to generate the mixed image (as would be used in traditional *mixup*; i.e., placing mass only on two categories). While participants did tend to place probability mass on the generating endpoints that correlated with the mixing coefficient used (Pearson $r = 0.52$). Interestingly, we find that participants report thinking that 38.3% (±0.6%) of the probability mass of a label should be placed on *different* classes from those which are used to create the image. This is remarkable and suggests that mixed images *do not* consistently look like the labels used to create them, corroborate similar trends found in (Gruber et al. 2018) wherein humans endorse categories which are not present in the image. Hence, alternative labelings even beyond the kind we explore in the main text may be preferred which are more aligned with human percepts. Examples of such labeled mixed images are shown in Fig. 13.

Takeaways The typical two-category labels used in *mixup* do *not* consistently match human perception. We find that human annotators often assign probabilities to alternate classes when asked to label a mixed image. This suggests that the pursuit of aligning synthetic data labeling to match human perception, at least for the synthetic data constructor used in *mixup*, warrants the design of alternative label mixing functions g_{rich} which yield richer label distributions over a broader range of categories.