Improving Machine Learning Systems by Eliciting and Incorporating Additional Human Knowledge

Katherine M. Collins

Supervisors: Dr. Adrian Weller MBE
Umang Bhatt

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
Master of Philosophy

Magdalene College

August 2022

Dedicated to my supportive, inspiring, and fun family.

Declaration

I, Katherine M. Collins of Magdalene College, being a candidate for the MPhil in Machine Learning and Machine Intelligence, hereby declare that this report and the work described in it are my own work, unaided except as may be specified below, and that the report does not contain material that has already been used to any substantial extent for a comparable purpose.
Word Count: 14809 words
Number of GPU Hours Used: 1521 hours

Software Declaration Standard Python packages for machine learning applications were employed such as `PyTorch`, `scikit-learn`, and `scipy`. The following additional sources were consulted and/or directly leveraged in the code powering this work:

- The base code of *mixup* (Zhang et al., 2017) was used and modified to be run with soft labels.
- Uma et al. (2020) kindly shared their code which we checked to match training procedures.
- Models trained were drawn from their repositories, which originally were sourced from He et al. (2015) and Simonyan and Zisserman (2014).
- The code of Peterson et al. was also inspected and somewhat drawn on to ensure reproducibility of their results, though heavily modified. `CIFAR-10H` data was sourced from their repository.
- Pinto et al. also provided their Mix-MaxEnt code, which was modified to run within our setting.
- Adversarial attack code relied on the `torchattacks` package, though was custom-modified to extract the loss.

- The calibration metric from Hendrycks et al. was sourced from their repository and also modified.
- GloVe embeddings (Pennington et al., 2014) and the `gensim` `Gensim` embeddings were used in tandem with code drawn from the following online tutorial for semantic re-distribution experiments.
- Natural corruption robustness checks used `CIFAR-10-C` created by Hendrycks and Dietterich.
- All human experiments were custom-written ontop of the `jsPsych` platform base (de Leeuw, 2016).

Wherever online programming resources were significantly relied upon, comments are left in the documentation of the code.

Human Subject Experiments Ethics Declaration Human subject experiments were approved by the Department of Engineering Ethics Team under the light-touch scheme. All experiments are run through Prolific (Palan and Schitter, 2018) and hosted with Pavlovia. To align with local regulations, annotators are paid at a base rate of $8/hr with a possible bonus up to a rate of $9/hr.

Paper Declaration The work of this dissertation encompasses and expands on two papers submitted to AAAI Human Computation and Crowdsourcing (HCOMP) Conference. The bulk of the work was done during the thesis period (March to submission; the soft label paper idea came earlier in the year), and has not been submitted elsewhere for academic marking. See Section 1.4 for details on contributions.

Katherine M. Collins
August 2022

Acknowledgements

Thank you to Dr. Adrian Weller for being an absolutely incredible, supportive, and inspiring supervisor. I always appreciate careful questioning of experimental results, boundless ideas, and tips on how to tell a good story. And a textithuge thanks to Umang Bhatt for being an amazing collaborator. I have greatly enjoyed all of our idea sessions, and have appreciated being able to discuss everything from the minutiae of experimental details to broader discussions about academia and larger research questions with you. I am deeply grateful to you and Adrian for helping me see the promise of human-machine teams and providing me with opportunities to already have a chance to engage with ongoing research directions in the space. I am very excited for many more years of collaboration.

Thank you to my other co-authors, Professor Bradley Love and Weiyang Liu, for thoughtful feedback on the *mixup* work, as well as all my past co-authors for their help getting me to where I am in research. Thank you also to the Computational Biological Learning Group, especially Miri Zilka, Jiri Hron, and the Human-AI Reading Group – as well as Krishnamurthy (Dj) Dvijotham, David Krueger, Jeff Shrager, and Marty Tenenbaum.

I especially want to thank Josh Tenenbaum for his endless support and helping me see the rich potential in studying human cognition within machine learning.

A hearty thanks to Professor Richard Turner for not only being an incredible MPhil Programme leader, but for his kindness and support throughout the entire year. Thank you also to Dr. Carl Henrik Ek and Professor Neil Lawrence for helping illuminate the power and promise of probabilistic machine learning.

Thank you to my MPhil coursemates, the Marshall Scholar community, and the Audley Cottage crowd for a fun year. In particular, to Alan Clark, for his detailed comments on the draft. Thank you also to Max Bronckers for his tips on the dissertation, as well as Stuart Burrell for his LaTeX genius.

This dissertation, nor the MPhil as a whole, without the amazing care of Anne Debenham, as well as Kimberly Cole, Rachel Fogg, Derek Matthews, and Stuart Rankin. I also want to thank Professor Bill Byrne for nearly instantly adding GPU hours to my account whenever asked.

Lastly, I want to thank my family for their truly bottomless support and encouragement over not just the course of the dissertation, but my entire life!

Abstract

Data has powered incredible advances in machine learning (ML). Yet, the kinds of data used for training are often hard labels aggregated over humans' annotations, which fail to capture the richness of disagreements across humans, as well as the uncertainty of any individual in their annotation. And while synthetically-generated data has played a key role in ML development, it is not always clear whether the generated data aligns with human perception, which may be important to ensure model trustworthiness. Can additional forms of human knowledge be leveraged to inform the design of better supervisory signals for ML systems to improve downstream generalization, calibration, and robustness?

Here, we review existing works aimed at eliciting and incorporating additional forms of human knowledge and identify two key gaps: 1) the use of soft labels elicited from individual humans, and 2) the alignment of synthetically-generated data with human perception. We address these gaps by conducting several human crowdsourced studies and computationally investigating whether labels constructed using the elicited information enables models trained to enjoy performance gains. However, we find that eliciting rich information from humans is expensive and suffers from scaling challenges. In light of these limitation, we also consider using a small amount of elicited human knowledge to inform the design of automated training procedures.

We therefore contribute to the empirical study of how ML systems could be improved through additional human knowledge. We also release to the community with a new collection of soft labels for `CIFAR-10` – a dataset we release as `CIFAR-10S`, as well as (we believe) the first set of human labelings over synthetically-generated *mixup* training examples.

Table of contents

List of figures

List of tables

Nomenclature

D	Dataset
H	Label entropy
K	Number of categories (in a classification task)
M	Number of human annotators for an instance
N	Number of instances in a provided dataset
P	Label distribution
P_m	Label distribution of an individual human annotator
Ψ	Aggregation algorithm
Θ	Space of possible parameters
α	Mixing coefficient distribution-control parameter
β	Amount of uniform label smoothing
γ	Redistribution smoothing adjustment
κ	Lower-bound entropy clamp
λ_f	Generating mixing coefficient
λ_g	Target mixing coefficient
λ_h	Human-inferred mixing coefficient
$\mathcal{C}$	Concept space
$\mathcal{H}$	Hypothesis class

$\mathcal{X}$	Input domain
$\mathcal{Y}$	Target space
ω	Human confidence
τ	Logistic mixing function
θ	Specific model parameters
$\tilde{x}$	Synthetic input
$\tilde{y}$	Target for a synthetic input
f	Input data mixing policy
g	Target mixing policy
h_θ	Learned model
r	Probability-mass redistribution function
z	Privileged information

Chapter 1

Introduction

1.1 Motivation and Overview

Machine learning (ML) has the potential to profoundly impact many areas of society from healthcare to criminal justice. However, as these systems increasingly permeate many aspects of our life, we want to ensure that they can be trusted and are reliable (Avin et al., 2021; Hendrycks and Dietterich, 2019; Tran et al., 2022; Zerilli et al., 2022). This means, in part, that models should be able to handle unseen data (generalize), produce calibrated predictions which faithfully capture ambiguity (Guo et al., 2017), and be robust to both adversarial and natural perturbations (Hendrycks and Dietterich, 2019; Madry et al., 2017).

How can we achieve these desirable properties? While significant gains have come from new algorithmic developments, for instance in probabilistic machine learning Ghahramani (2015), we argue there is immense, complementary potential in collecting and leveraging other forms of human knowledge to improve machine learning performance. Humans possess knowledge which typically is not in ML systems (Lake et al., 2017; Marcus, 2018; Shneiderman, 2022) and the idea of how human knowledge may differ from that of a machine has a storied history (Turing, 1950).

Here, we focus on ways in which additional human knowledge can be incorporated into the *data* used to train ML systems. As data has been a key driver of ML success, there has been a rise in interest in the datasets used in ML, such as the push for data-centric ML (Lawrence, 2019; Mazumder et al., 2022) and calls for more transparency over data use (Díaz et al., 2022; Gebru et al., 2021; Paullada et al., 2020; Prabhakaran et al., 2021). Exploration into the the kind of information we elicit from humans to form our training sets has potential to substantially impact the development of better ML systems.

Many of the datasets today are formed from the elicitation of hard labels from humans. While combining the annotations of many humans to form a label which represents annotator disagreement (Davani et al., 2022; Uma et al., 2022) has been found to be beneficial in learning (Nguyen et al., 2013; Peterson et al., 2019; Uma et al., 2020), we argue that that these labels are lossy. Constraining each annotator to only provide a single label does not empower an individual to express their confidence in the annotation, nor their belief in any alternative categories. In this work, we study whether ML systems can be improved by training on soft labels elicited from every individual annotator, particularly in the realistic setting **when fewer humans annotators are available**.

Moreover, the majority of works which do center around incorporating additional human knowledge do so over naturally available examples (e.g., an image of cat), rather than synthetically-generated data. Understanding human alignment of synthetic examples is especially important as synthetic training data become even more commonplace in the ML development cycle (de Melo et al., 2022; Jordon et al., 2022). Here, we make a step towards investigating whether and how human perceptual knowledge over synthetically created examples can be leveraged to impact model behavior – both directly and through simulated mimicry of human knowledge. We focus on the data generated when using *mixup* (Zhang et al., 2017).

Our case studies address **the value of different kinds of human knowledge that can be incorporated to improve model performance**, as well as how to handle when we have information from *many* humans. We provide an optimistic, albeit nuanced, view of the prospects of human knowledge-based machine improvements – which contributes to a deeper understanding of how training examples can be elicited and constructed to support the the development of more trustworthy ML systems.

1.2 Contributions

In this dissertation, we contribute:

1. A **taxonomy** of the kinds of human knowledge which have been, and can further be elicited, with a **survey** of methods for aggregating information from many humans.

2. **Three human knowledge elicitation studies** involving the elicitation of under-addressed forms of human knowledge (soft labels from every human over classical `CIFAR-10`(Krizhevsky, 2009) images; human inferences over the generative process of synthetic examples, including human confidence; and human soft labels over said synthetic examples). We release our `CIFAR-10` soft labels in the form of a **new dataset**, dubbed `CIFAR-10S`.

3. A preliminary attempt at **constructing of a "soft label simulator"** using the statistics of `CIFAR-10S`.

4. A thorough empirical **investigation into the most effective way to leverage soft label knowledge from humans** to most improve performance.

5. The first, to our knowledge, **exploration of the correspondence of human percepts over the synthetic training examples** used in *mixup* (Zhang et al., 2017).

6. An **investigation into the impact of aligning *mixup* labels with human percepts** on model performance.

7. **Three extensions of traditional *mixup*** inspired by our human user studies to encourage softness in *mixup*: 1) utilizing a **new clamped entropy-weighted loss**, and 2) mixing over endpoints which leverage **human-elicited and simulated human soft label knowledge**. A third early-stage method of **human-grounded sigmoid-based mixing transformation** is included in the Appendix.

1.3 Thesis Plan

The dissertation is structured as follows:

- **Chapter 2:** We define what we mean by "additional human knowledge" and lay the mathematical groundwork for . We provide a taxonomy for some of the ways human knowledge may be incorporated and review related work on the kinds of human knowledge being elicited for use in ML. We overview literature around how to handle information elicited from many humans.

- **Chapter 3:** We address a gap in the kinds of knowledge elicited from humans, specifically, through the collection of soft labels from every human. We compare the labels we elicit to those constructed by the more standard method of aggregating many humans' elicited hard labels, and study their impact on model performance as a function of number of human annotators and total annotation time. We investigate how best to utilize our elicited soft labels during training.

- **Chapter 4:** A second class of human knowledge is considered, now over *synthetic* examples. We focus on the synthetically created examples used in *mixup* training (Zhang et al., 2017). We study the alignment of human perceptual knowledge and labels traditionally used in *mixup* through a novel human crowdsourcing paradigm. We then explore the impact on model performance of relabeling such examples at training time with human elicited knowledge.

- **Chapter 5:** A follow-up human user study is conducted to address the question of whether humans perceive different category structure in labels than is typically assumed in the labels used in *mixup* to train ML systems. Inspired by our human data, we propose two measures of encouraging "softnesss" in models' predicted distributions. We also introduce a novel simulator for human soft label creation, connecting back with the work of Chapter 3.
- **Chapter 6:** We discuss several exciting directions which arise from this work and address limitations and key challenges in the incorporation of elicited human knowledge within the ML community.
- **Chapter 7:** We conclude with a summary of the dissertation with central takeaways.

1.4 Relation to Submitted Papers

The work in this dissertation encompasses and expands on two papers completed during the timeline of the thesis (March 2022 onwards). Both papers were submitted to AAAI Human Computation and Crowdsourcing (HCOMP) and received positive reception[1].

1. Katherine M. Collins*, Umang Bhatt*, and Adrian Weller, "Eliciting and Learning with Soft Labels from Every Annotator"; **accepted as a full paper**.
2. Katherine M. Collins, Umang Bhatt, Weiyang Liu, Bradley Love, and Adrian Weller, "Human-Annotated *mixup*"; **invited for fast-track publication as a Work-in-Progress paper**.

Contributions I implemented and ran all experiments (human and computational) and made all figures and tables. I also wrote the first pass of both drafts, which were graciously edited and expanded on by the listed co-authors (particularly co-first author and co-supervisor, Umang Bhatt). Some figures, text, and results are included directly in this thesis (specifically in Chapters 2-6); however, the vast majority of computational results have been re-run and expanded for the dissertation report and significant de novo text has been written.

[1] Reviews were received on August 15. While these helped inform some of the writing of the dissertation, all experiments included here were run before receiving the reviews.

Computational extensions have included analyses around annotator aggregation, semantic-based label smoothing, new baselines, additional model architectures trained, alternative automatic labeling schemes (soft label generative model, class-pair mixing), and an additional corruption robustness evaluation metric.

Chapter 2

Background

Chapter Roadmap We first introduce the problem setting considered, before addressing what we mean by "additional human knowledge." We provide an overview the ways in which said human knowledge can be incorporated to improve ML model performance and identify gaps where additional human knowledge could be collected. We discuss various kinds of knowledge that could be collected from individuals and survey best practices in aggregating information across humans. We connect these directions to the specific narrative of this thesis.

2.1 Problem Setting

For the purposes of this thesis, we focus primarily on the incorporation of human knowledge in the supervised learning setting and a $K-$way classification task. That is, there is an observed dataset $\mathcal{D}$ composed of N datapoints with provided features $x_n \in \mathcal{X}$ and K-class target vectors $y_n \in \mathcal{Y}$.

For generality, we let $\mathcal{Y}$ represent a $K-$multiplex, which we can view as a discrete probability distribution over the K-classes, i.e., $\sum_{k=1}^{K} \mathcal{P}(y_n = k|x_n) = 1$ and $\mathcal{P}(y_n = k|x_n) \geq 0$, $\forall k \in \{1, ..., K\}$.

When the label is the traditional one-hot vector $y_n \in \{0,1\}^K \subseteq [0,1]^K$, we call this a **hard label**. In our framework, the label distribution $P(y_n|x_n)$ then has all mass placed on a single class:

$$P_{\text{hard}}(y_n = k|x_n) = 1[y_n = k]$$

where $1[y_n = k]$ is an indicator variable of whether class k has been assigned or not by the annotation process.

Traditionally, a model h is selected from some model class $\mathcal{H}$, parameterized by $\theta \in \Theta$, and optimized to predict $\mathcal{P}(y|x)$ according to some objective $\mathcal{J}$. Though we could consider functions which take side, or auxiliary, information z as input. Within this framework, we consider the plethora of ways in which human knowledge has been, and could further be, incorporated in the aim of improving the performance of h_θ.

2.2 Defining "Human Knowledge"

First, we take a step back to define what we mean by *human knowledge*. We use the term "human knowledge" quite loosely: encompassing humans' commonsense, perceptual judgments, as well as speciality expert factual and procedural information which could be innate or gleaned, for instance, via experience or deliberate education (Turing, 1950). Of course, any machine learning system is the result of multiple forms of human knowledge – ML practitioners select the model class $\mathcal{H}$, the objective function $\mathcal{J}$, and make several other design choices which represent their own knowledge of the task. Here, however, we focus on *non-traditional* forms of human knowledge; primarily those which come not from the ML practitioner themselves.

Why then would we care about incorporating additional human knowledge? Can we not have ML systems trained on data which learn to approximate this information? First, we argue in this thesis that the very manner in which data is collected for training ML systems ought to be investigated — if human-like reasoning is to "fall out" of ML systems, this is due at least partly to the data – necessitating the use of appropriate examples (Emam et al., 2021). Moreover, even with adequate data, it is questionable whether machine learning systems will reach human-like performance (Marcus, 2018), at what rate (Kaplan et al., 2020), and even if feasible, whether that is a future we wish to reach (Bender et al., 2021; Shneiderman, 2022). In any case, in the near – and potentially longer-term, it is worth considering how human knowledge could be leveraged to design ML systems which better function *with* humans (see Discussion in Chapter 6) (Wilder et al., 2020). This has triggered calls for the collection of more human user studies (Shavit et al., 2022) and redressment of regulations around such collection (Kaushik et al., 2022). Similarly, aligning machine learning models with human perception has the potential to improve trustworthiness (Nanda et al., 2021). And beyond the benefits to human alignment that may be derived from incorporating more forms of human input into systems: simply improving the reliability of a machine learning model (e.g., via better data or model priors) ought to confer such benefits as well.

2.3 Incorporating Human Knowledge in ML Systems

As such, we believe it is worthwhile to consider: 1) what kinds of human knowledge to elicit from a single humans, and 2) how best to leverage information across many humans to form effective supervisory signal(s) for ML systems.

2.3.1 Information from a Single Human

Overview

Several forms of information can elicited from any given individual – from a likely category, to confidence in said categorization, and even the selection concepts encompassed by an example. The kinds of human knowledge sought after ought to be informed by their likely use case (Chen et al., 2022). We offer a taxonomy for where we see elicited human knowledge being most applicable: human knowledge can be incorporated in the selection of the hypothesis class $\mathcal{H}$, the space of parameters Θ, and/or the dataset $\mathcal{D}$ (which could comprise information about classical observational or synthetic data; or side information the human has priveledged access to).

While the former the former are exciting, for instance, elicited approximations to experts' probability distributions could form better priors (Fortuin, 2022; Oakley and O'Hagan, 2010), we focus on the former: human knowledge that can be incorporated into the data, e.g., $(x_n, y_n) \in \mathcal{D}$. We consider each of the subareas of eliciting information in the dataspace in turn. We recognize there are most definitely more ways in which human knowledge can inform ML development, we hope this overview helps unify existing work and inspires research into these and others.

Classical Observational Data

Incorporating human knowledge into the classically observed (x, y) pairs is standard practice. Indeed, a powerful driver for ML progress, particularly in the image classification setting, has been the annotation of large-scale datasets (Emam et al., 2021; Fei-Fei et al., 2009; Krizhevsky, 2009). Often, these y_n are the result of asking a human to select a single category they think most represents x_n from a set $k \in \{1...K\}$.

However, let us assume that to produce y_n, humans run some form of inference that samples from an internal probability distribution $P(y_n|x_n)$. Another form of human knowledge could be to aim to elicit a *snapshot* of the entirety of $P(y_n|x_n)$, i.e., $K-1$ probability judgments, or a subset therein from an individual human. In our framework then, we can consider eliciting a **soft label** from an individual:

$$P_{\mathrm{m}}(y_n = k|x_n) = p_k^m$$

where $p_k^m \in [0,1]$ that the label $y_n = k$, assigned by annotator m. In the binary setting, this is equivalent to eliciting an annotators' confidence (discussed more below).

Although humans may not be able to perfectly report their internal probability distribution for a given example (Murray et al., 2015) and such responses p_k^m may not be consistently calibrated (Lichtenstein et al., 1977; O'Hagan et al., 2006; Sharot, 2011; Tversky and Kahneman, 1996), we do not think this is a sufficient reason to avoid eliciting probability judgments from annotators. As noted by O'Hagan et al. (2006) and O'Hagan (2019), human uncertainty can be elicited reliably as long as elicitation is rigorous. Moreover, if an annotator is unsure of their decision, forcing an annotator to compress out all of this uncertainty by specifying one hard label only exacerbates, rather than solves, the challenge of capturing annotator ambiguity. While information about P_m has been studied in the crowdsourcing literature (Chung et al., 2019; Méndez et al., 2022), the incorporation of such human knowledge, specifically into ML systems and when $K > 2$, remains in early days and is a central component of this thesis (see Chapters 3 and 5).

Alternatively, other forms of human knowledge beyond direct probabilities could be elicited to provide more information over an example. For instance, a human select multiple likely labels (Beyer et al., 2020; Chung et al., 2019) or a set of likely labels (Beyer et al., 2020), rank labels (Chen et al., 2021), or provide iterative yes/no answers to hierarchical questions (Branson et al., 2010). Massiceti et al. also demonstrate the benefits of leveraging human annotators to provide speciality observations, e.g., recordings of personal objects by people with visual challenges.

A given human can also express their knowledge through x_n, given y_n. For instance, we view the formation of few-shot linguistic prompts fed to foundation models (Bommasani et al., 2021) and similar large-scale architectures as another manner in which human knowledge can be expressed in "training" data. As an example, Collins et al. (2022) leverage plans and explanations generated by humans to achieve specific goals when constructing effective GTP-3 Brown et al. (2020) prompts; Wong et al. (2021) similarly crowdsource human-generated language data to guide scalable program search.

Synthetic Data

An individual could also be queried for information about *synthetically-generated* examples. Synthesizing effective examples to augment model training has unlocked tremendous ML advances that may not have been possible with limited standard data (de Melo et al., 2022;

Emam et al., 2021; Jordon et al., 2022; Silver et al., 2016). Synthetic data could be generated through a (somewhat) opaque process such as a learned model like a Generative Adversarial Network (GAN) (Goodfellow et al., 2014a), or constructed from some explicit generative process, such as cropping or blurring (Shorten and Khoshgoftaar, 2019), or linearly mixing two examples as is done in *mixup* (Zhang et al., 2017). We refer to synthetically created examples as $(\tilde{x}, \tilde{y})$.

Human knowledge could be elicited and incorporated to improve the fidelity of synthetic data construction. For instance, Zhang et al. consider inquiring for humans for annotations over GAN-generated images to bootstrap the creation of more information. Here, humans express their knowledge over $\tilde{y}$. Alternatively, humans can be queried to provide better $\tilde{x}$. Kaushik et al. 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., 2020). Humans could also be queried to alter the generative process, for instance, inferring the combination factor in the creation of synthetically mixed images in *mixup* – which we address in Chapter 4 in this thesis.

Auxiliary Information

We have reviewed several kinds of rich human knowledge that can be elicited over classical observational and synthetically-generated data. Next we consider human knowledge that lies outside of $\mathcal{X}$ and $\mathcal{Y}$, which could be leveraged to improve model performance.

For instance, we can assume a human annotator has access to side information (z), which could take the form of a belief state or any or classification rule (Vapnik et al., 2015) or other privledged information that only the human has access *a priori* (Mozannar and Sontag, 2020; Sharmanska et al., 2016). This information could be elicited and fed into the model either by augmenting x or modulating the targets y.

Additionally, human knowledge could be used to constrain an intermediate stage of the predictive pipeline. Concept Bottleneck Models (CBMs) present a prime example of this form of human knowledge incorporation (Koh et al., 2020; Margeloiu et al., 2021; Ramaswamy et al., 2022). Here, humans can specify concepts $c \in \mathcal{C}$ which a model should use to represent a given x, that can then be decoded to the target y. $h_\theta(x)$ is then a composition of two functions $h^c_{\theta_c} : \mathcal{X} \to \mathcal{C}$ and $h^y_{\theta_y} : \mathcal{C} \to \mathcal{Y}$ s.t. $h_\theta(x) = h^y_{\theta_y}(h^c_{\theta_c}(x))$ with $\theta = (\theta_c, \theta_y)$. Eliciting human knowledge over c, which could be used to intervene and improve model performance, then provides another compelling case for eliciting other forms of human knowledge in the ML development cycle. Moreover, as pointed out by Ramaswamy et al., interplay with human factors are important when considering deployment, e.g., humans can only process few concepts.

Humans could also report their confidence ω in whatever form of knowledge they provide, whether that be in y or other forms of knowledge like in c. For instance, Nguyen et al. (2014) leverage experts' confidence to train a classifier to predict medical emergency risk, and Steyvers et al. (2022) leverage human confidence in classification to design more effective human-machine teams. Confidence could also be elicited in hierarchical labeling paradigms Branson et al. (2010), wherein annotators are asked to select if their are {Guessing, Probably, Definitely} confident when coming up with their annotation. Expansive work remains in studying the value of incorporating human-elicited confidence in learning. We expand on this research direction in Chapter 4 of this thesis.

2.3.2 Knowledge from Many Humans

We have surveyed various kinds of human knowledge that can be incorporated into ML systems. However, what should we do if we have information about the same example from $M > 1$ different humans? We can consider an aggregation algorithm Ψ which takes all M annotations and returns a consolidated form of human knowledge. Let us focus on elicited labels y as a case study. A common form of Ψ in a $K-$way classification setting is a majority vote (Davani et al., 2022). However, consider the case where annotators disagree: perhaps $\lfloor \frac{M-1}{2} \rfloor$ annotators deem an image to be a dog, and the rest consider the image a cat. Then the image will be marked as a cat, and given the model is trained to map from said image to cat completely, we lose all explicit information about its potential likeness to a dog. Ψ is therefore *lossy*.

Is there a better way to preserve the richness of human knowledge when aggregating across many humans? Several works have considered a Ψ which instead returns a soft label, formed by maintaining the frequencies of many humans' provided hard label (Gordon et al., 2021, 2022; Koller et al., 2022; Peterson et al., 2019; Recht et al., 2019; Sharmanska et al., 2016; Uma et al., 2020, 2022). This label distribution then takes the form:

$$P_{\text{multi}}(y_n = k|x_n) = \frac{1}{M}\sum_{m=1}^{M} 1[y_n^m = k]$$

.

The result is a soft label where $y_n \in [0,1]^K$.

However, the aggregation method does not account for annotator quality. Indeed, crowdsourcing workers may provide inconsistent or low-quality annotations. As such, several works have proposed methods to infer a better "gold" label by accounting for annotator quality and inter-annotator disagreement (Dawid and Skene, 1979; Sharmanska et al., 2016; Smyth et al., 1994; Whitehill et al., 2009; Zhang and Wu, 2018). Dawid and Skene propose an Expectation-Maximization (EM)-based approach to disentangle variability amongst clinicians; Augustin et al. (2017) leverage Bayesian methods to account for crowdsourcing "spammers" when aggregating; and more recently, or Collier et al. consider training an auxiliary neural aggregator, which also utilizing annotator time as a proxy for example difficulty when informing aggregation. Wei et al. also demonstrate the utility of annotator aggreement as a proxy for confidence in the annotation, which is used to smooth labels for effective polyp classification; similar confidence-based aggregation have been developed in Song et al.. Alternative aggregation approaches which involve interactions between humans have also been proposed (Oakley and O'Hagan, 2010; O'Hagan et al., 2006; Thangaratinam and Redman, 2005). While the latter are highly involved, interactive online experiments offer grounds to scale such human-human knowledge acquisition (Miller and Steyvers, 2011).

However, recent literature has suggested that aggregating M humans' knowledge into a single compressed representation may not always be ideal. Wei et al. and Platanios et al. highlight that when M is small, or annotators are very noisy, learning over the original, *de-aggregated*labels is better. This raises the question of *whether* human knowledge should be aggregated in the first place when leveraging as a supervisory signal for ML systems – which we study in Chapters 3 and 4.

Considerations around aggregation also can inform considerations around how many humans to query information for (Lin et al., 2014; Schmarje et al., 2022; Shimizu and Wakabayashi, 2021) and is additional explored in Chapter 3.

2.4 Scope of Thesis

Therefore, in this thesis we focus on addressing some of the gaps noted in the kinds of human knowledge elicited – namely, in eliciting richer probabilistic information from humans (i.e., soft labels and confidence in annotations), as well as knowledge over synthetically constructed examples. We begin to address the question of whether to aggregate human knowledge.

Next We next address the first form of human knowledge that we consider in this work: soft labels elicited from each and every annotator.

Chapter 3

Exploring the Value of Collecting Soft Labels from Every Annotator

The last chapter surveyed the landscape of how additional human knowledge is being incorporated into machine learning systems already, and highlighted the potential for where new forms of human knowledge can be elicited and incorporated in an effort to improve model performance. In this chapter, we address one of the gaps: namely, the utilization of richer representations of humans' uncertainty over targets in the form of soft labels from *each and every* annotator.

Chapter Roadmap This chapter is structured as follows: first, we further motivate why eliciting soft labels from every annotator serves to contribute new human knowledge that is not currently being incorporated into the ML development pipeline; second, we introduce our elicitation framework and analyze the soft labels we get back (a new dataset which we dub CIFAR-10S), highlighting how they fundamentally differ from the kinds of labels typically collected; third, we investigate the impact of training models on these labels and include discussion of the limitations of our elicitation as it relates to use in ML systems.

3.1 Why Care?

As discussed in Chapter 2, label distributions are often constructed with a single hard ordinal label may have been decided on by a single annotator (Passonneau and Carpenter, 2014), or a majority vote of multiple annotators (Sheng et al., 2017); and if they are designed to be soft, this is from aggregating many annotators' hard labels. A prime example of the latter is CIFAR-10H proposed in Peterson et al. (2019) and (Battleday et al., 2020). The authors collect $M \sim 51$ annotators for each of the `CIFAR-10` (Krizhevsky, 2009) test set images. The constructed soft labels have been shown to improve model generalization and robustness (Peterson et al., 2019) – and have since been used for several applications, from benchmarking the reliability of foundation models (Tran et al., 2022), to informing human-machine teaming (Babbar et al., 2022; Straitouri et al., 2022), and expanding the empirical understanding of the impact of labels on performance (Schmarje et al., 2022; Wei et al., 2022b). While the labels are incredibly valuable for the community, they are often touted as representing human "label uncertainty;" (Tran et al., 2022). Although such labels do capture some form of uncertainty, the picture is incomplete as it covers ambiguity *across* humans, and does not capture an individuals' uncertainty.

Eliciting instead *individual* humans' personal probability distribution P_m over the labels[1] could therefore yield different converged labels. Consider the following thought experiment: if every annotator is 51% sure an image is class k and 49% sure it is class ℓ, they will provide class k in the elicitation of Peterson et al. (2019). For our setting, annotators can express their label probabilities directly. This could yield cases where label distributions appear to have all mass on a single class, whereas a possibly more human-aligned and information-rich label distribution can be procured through P_m per human.

A study is then warranted in to whether soft label elicitation converges to more representative label distributions and thereby serve as better supervisory signals for ML systems. We believe we are the first to train using rich soft labels elicited *directly* from annotators by requesting probabilistic judgments per annotator for multi-class problems. Moreover, we posit that fewer annotators may be needed to support model performance with our approach.

3.2 Introducing and Exploring `CIFAR-10S`

We now discuss how we collect our dataset, `CIFAR-10S`. To elicit soft labels from each annotator, we request:

1. The most probable label, with an associated probability
2. Optionally the second most probable label, with an associated probability
3. Any labels which the image is *definitely not*

The most probable and second most probable labels are selected via a radio button, whereas the selection of "definitely not" possible labels is marked through a checkbox to

[1] We acnkowledge this is necessarily an approximation, as discussed in Chapter 2; an individual may never be able to report their *exact* internal distribution. Future work could consider distributions over humans' self-reported label distributions.

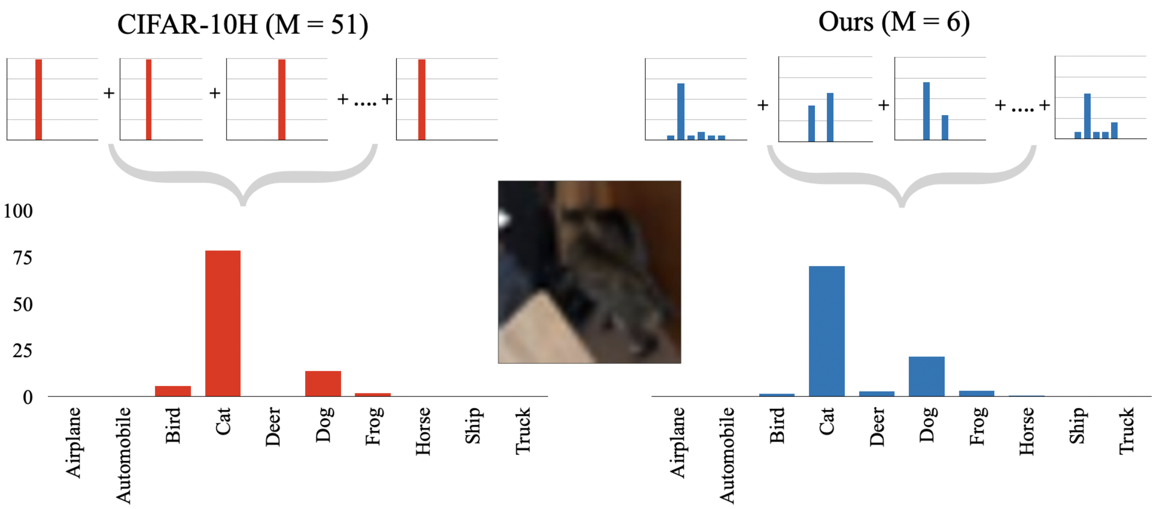


Fig. 3.1 Comparison of kinds of information elicited from M annotators with `CIFAR-10H` vs. our scheme, which enables faster convergence.

allow annotators to select multiple labels. Probabilities are entered in a text box and asked to be between 0 and 100. We do not require that probabilities sum to 100 across the task, as we normalize after by using one of the elicitation practices of O'Hagan et al. (2006). We explore spreading any remaining mass over the labels not marked as impossible.

We additionally request annotators consider how *other* annotators, specifically “100 crowdsourced workers,” may respond. Encouraging annotators to consider a *third-person* perspective has been shown to encourage more representative responses (Chung et al., 2019; Oakley and O'Hagan, 2010), and is partly inspired by Bayesian Truth Serum (Prelec, 2004). Our interface is depicted in Appendix Fig. A.1.

3.2.1 Elicitation Setup

We recruit $N = 248$ participants on Prolific (Palan and Schitter, 2018). We collect labels over a total of 1,000 images from `CIFAR-10H` (Battleday et al., 2020; Peterson et al., 2019) (i.e., 10% of the dataset). Battleday et al. (2020) found that majority of `CIFAR-10H` labels (70%) have very low levels of annotator disagreement, operationalized by label entropy (H):

$$H(y) = -\sum_{k=1}^{K} p(y=k)\log p(y=k)$$

However, to best validate our approach over a limited subset of labels, we elect to enrich the set we show our annotators with the highest entropy examples ($H > 0.25$). However, we

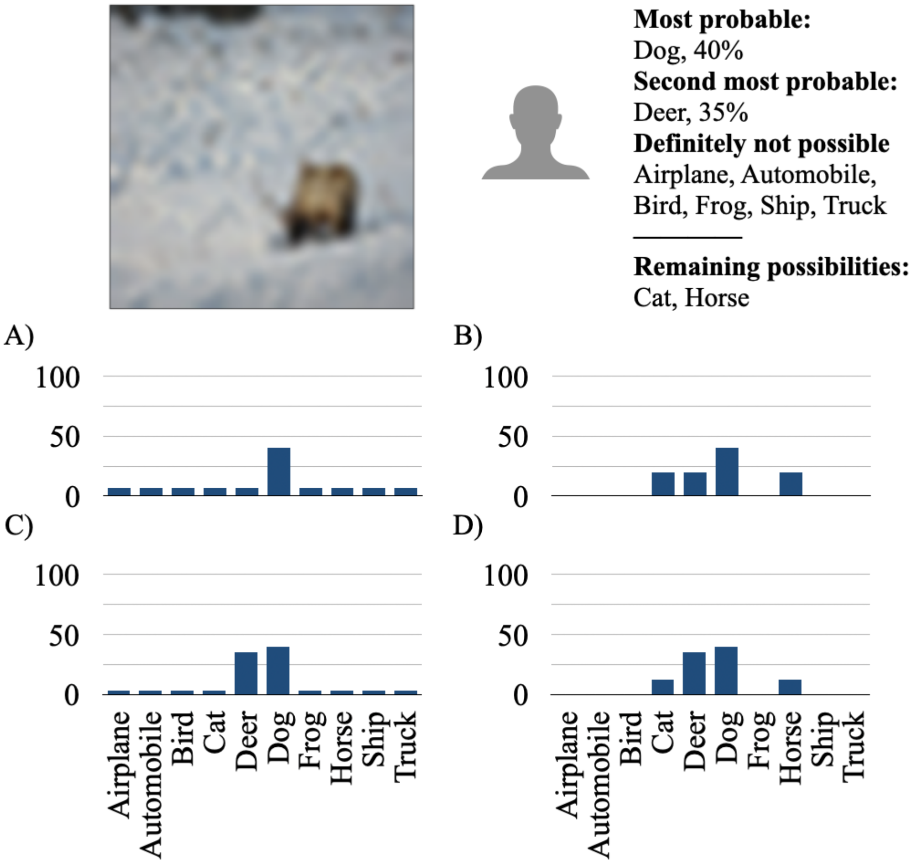


Fig. 3.2 Depiction of constructed label varieties from the information elicited from a single annotator: A) Top 1, Uniform, B) Top 1, Clamp, C) Top 2, Uniform, D) Top 2, Clamp. Note, possible labels are inferred by exclusions.

included three images with low entropy ($H \leq 0.1$) under CIFAR-10H within each batch[2] to ensure a sufficient diversity of ambiguity was shown to each participant.

We follow (Battleday et al., 2020) in *up-sampling* each image to a resolution of 160x160 using Lanczos-upsampling. While this reduces the amount of ambiguity in the traditionally low-resolution CIFAR-10 images, we aim to benchmark our method against (Peterson et al., 2019) as closely as possible and hence follow their transformation. As we observe that an overwhelming proportion of CIFAR-10H images have mass on only two labels (approximately 77.2%), we ask participants to specify only the top two most probable labels and any that are definitely not possible.

[2] With the exception of two of the batches of the 40 batches that contain all higher entropy images due to randomization.

Each participant sees a batch of 27 images, where two such images are repeated as checks for attention and consistency. The order of labels and images was shuffled across participants. We encourage annotators to provide responses they think others would provide, inducing third-person thinking (Chung et al., 2019; Oakley and O'Hagan, 2010). All remaining elicitation use third-person perspective framing.

3.2.2 Constructing Soft Labels

Our elicitation yields multiple pieces of information (first and second most probable labels with specified probabilities, and labels which are deemed to have zero probability) which we can use – or ignore – when forming a soft label. We explore several varieties of soft label constructions.

How to Redistribute Extra Mass? A central question in our elicitation scheme is how to distribute any mass which is left unspecified; for instance, if an annotator marks "truck" as the most probable class with probability 70% and "automobile" as the second most probable class at 20% likely, there is 10% of mass remaining that conceivably could be spread onto other classes.

To handle *redistribution*, we define a function r which takes as input any elicited probabilities from the annotator, and outputs a length K vector $\hat{p}_m$, representing the "completed" set of K probabilities over the label space (where $\sum_{k=1}^{K} \hat{p}_k^m = 1$). We then let:

$$P_m(y_n = k|x_n) = r(\{p_j^m\}_{j=1}^{K'})_k = \hat{p}_m^k$$

We consider two forms of r in this work: 1) **uniform** redistribution whereby the remaining mass is spread equally over the remaining classes, or 2) **clamp** which uses the "definitely not" elicitation to spread the remaining mass equally over only those classes which the annotator did not specify as zero probability. Alternative forms of redistribution which better capture inter-class relationships are warranted (see preliminary exploration in Appendix B.4).

If an annotator specifies 100% of the mass over the top one or two labels but only selects a subset of the remaining labels as definitely not possible, then we posit that the annotator views the unselected classes not having zero probability. Thus, we maintain a small portion of mass γ to be spread over the remaining classes. γ is selected via a held-out set, as discussed in Section 5.1. We do not apply this procedure in the *uniform* redistribution setting, as there we assume no access to the "definitely not" information.

Label Varieties We have 2 x 2 possible soft label construction methods: {most probable only, most probable and second most probable} x {redistribute uniformly, redistribute via clamp}. We use the notation T1 to specify if only the most probable class and its associated probability is used, and T2 if we include information about both the most probable and second most probable categories. We also refer to the redistribution approaches as "clamp" or "unif" following the definitions above. The label that uses *all* elicited information is T2 Clamp, which is the label set we refer to as `CIFAR-10S`. All soft labels, regardless of variety, are normalized to sum to one. Examples of constructed labels from a single annotators' response are shown in Fig. 3.2.

3.2.3 What New Information Do `CIFAR-10S` Labels Provide?

We compare how the structure of our elicited labels in `CIFAR-10S` compares to those in `CIFAR-10H` (Battleday et al., 2020; Peterson et al., 2019). As discussed in Section 3.1, the elicitation of `CIFAR-10H`, and any classical hard label set-up, is lossy.

While `CIFAR-10H` labels may nearly have all mass on a single class, our elicitation yields labels which have mass spread across more classes. This not only captures some of the inherent ambiguity in the image, but has the potential to provide information into inter-class similarity structure. For example, our annotators place mass jointly over "automobiles" *and* the similar "truck" category, whereas a `CIFAR-10H` label may have all mass on the "automobile" category; see Fig. 3.3. We highlight additional examples of label differences in Appendix Fig. B.1 and Appendix Fig. B.2. While we do not study inter-class similarity structure in this work, this direction is ripe for further inquiry.

Takeaways Our elicitation approach enables annotators to express their distribution over image labels, approximating and expanding on the richness of `CIFAR-10H` from far fewer annotators. Even in cases where annotators may agree on the most likely image label, our approach – which enables annotators to express their instantaneous probability judgments over *possible* other categories – yields labels which potentially better represent the distribution over labels.

3.3 On Learning with Soft Labels from Every Annotator

We now explore performance of models trained with the new kinds of soft labels that we construct against those formed from many annotators' hard labels (`CIFAR-10H`). We conduct an extensive investigation into the impact of aggregation, the number of humans whose

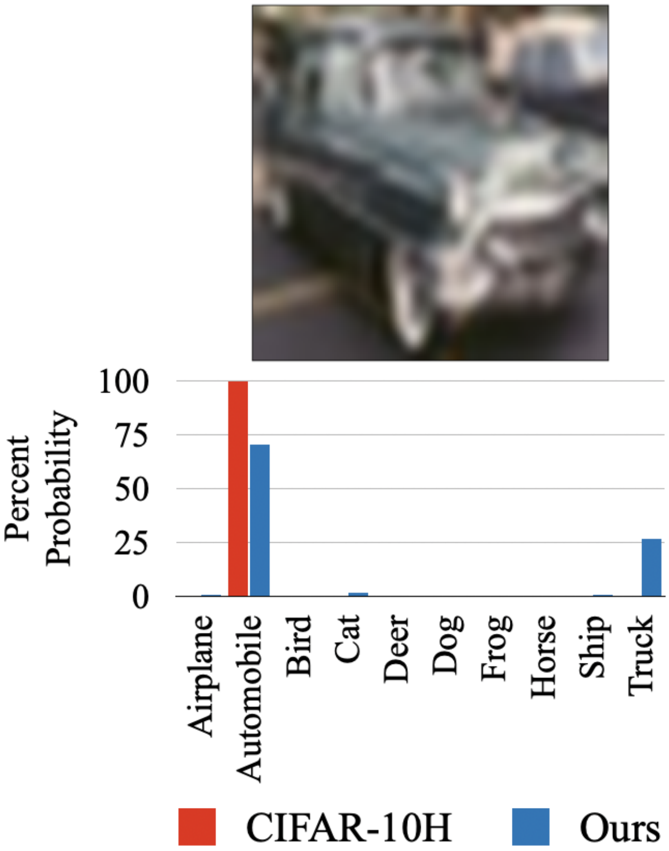


Fig. 3.3 Our soft variant of a `CIFAR-10H` hard label captures inter-class similarities (i.e., trucks and automobiles).

information we supervise with, total annotation time, the kind of information collected, and the number of examples for which soft labels are provided on conferred performance.

Setup Our model and training procedures follow Uma et al. (2020), as they explore learning with `CIFAR-10H` labels and explicate a clear, standardized learning procedure. We employ the same ResNet-34A (He et al., 2016) with the same weight decay (1e-4) and learning rate scheduling: we start with a learning rate of 0.1 and drop by a factor of 1e-4 after epoch 50 and again at 55. We also two additional architectures not included in Uma et al. (2020): VGG-11(Simonyan and Zisserman, 2014) and ResNet-50 (He et al., 2015). Starting learning rates are selected using a validation set of `CIFAR-10` (0.1 for VGG-11 and 0.4 for ResNet-50 from $\{0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5\}$. All models are trained from scratch for a total of 65 epochs and optimize a cross-entropy objective. Experiments are run over 5 seeds, unless

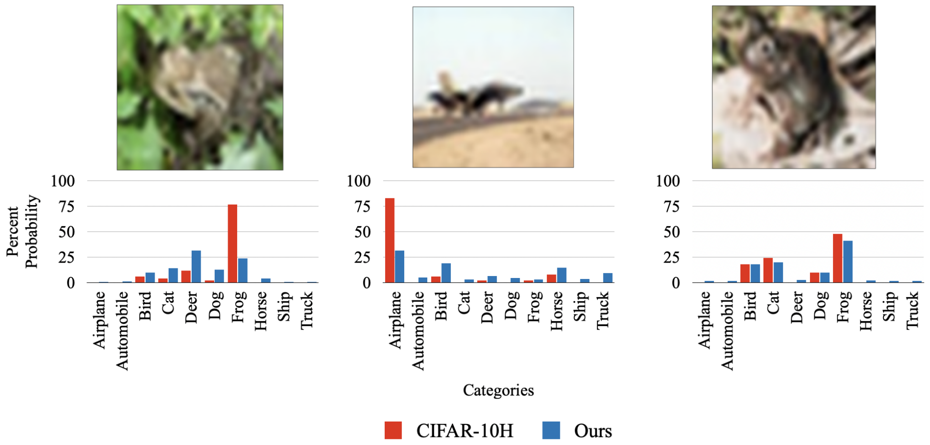


Fig. 3.4 Comparison of our elicited labels against CIFAR-10H. From left to right: two examples with high Wasserstein distance between labels; one example where we recover similarly rich, high entropy labels from 8.5x fewer annotators. The CIFAR-10 labels for these images are frog, airplane, and frog, respectively.

otherwise noted. A redistribution factor of $\gamma = 0.1$ is used to spread extra mass, selected via the same validation procedure from $\{0.0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4\}$.

We follow the same 70/30 split used in (Uma et al., 2020). However, as we have fewer CIFAR-10S labels than CIFAR-10H, to ensure a fair comparison, we consider just training on the CIFAR-10H labels that we have our soft label versions of. Note, we always hold out 100 of our labels to ensure we can evaluate against some variant of our soft labels. As such, we are considering the variation in performance conferred by changing 900 labels. For the remaining examples in the 6,100 – we use a hard version, i.e., the original CIFAR-10 label.

3.3.1 Evaluation Framework

Data

Selection of data used to evaluate models is important to faithfully benchmarking performance; however, evaluation datasets typically used like the CIFAR-10 test set are rife with annotation errors permeat (Northcutt et al., 2021) and as discussed throughout this work, if hard labeled, do not adequately capture human uncertainty. We instead use heldout aggregate soft labels from CIFAR-10H and our CIFAR-10S as test sets. While humans are of course not always correct in their annotations themselves, nor calibrated in their confidence (see 2.3.1), these labels serve to better measure whether models handle example ambiguity. When interpreting results, we note that while elicitation approach CIFAR-10H may converge to different label distributions than ours, we acknowledge that at present, `CIFAR-10H` is a more stable dataset given annotator quality is more tightly controlled (Battleday et al., 2020). We believe when scaled, our labels however could serve as a valuable benchmark; here, as we have few labels collected, we only maintain 100 held-out (which are re-sampled per seed). We note that as the labels we collect are enriched to be more ambiguous, `CIFAR-10S` is a naturally challenging evaluation set.

Metrics

No single metric captures all the qualities we wish to obtain in a trustworthy model Thomas and Uminsky (2022). We therefore consider a suite of metrics, focused on generalization, calibration, and robustness.

1. Generalization: we measure generalization using Cross Entropy (CE) over the soft label, which allows us to capture whether the models' full distribution over the K categories is sensible: $\frac{1}{N}\sum_1^N \sum_1^K P_{\text{eval}}(y_n = k|x_n)\log(m_\theta(x_n)_k)$, where P_{eval} is the discrete label distribution derived from the human soft labels.

2. Calibration: model calibration is scored using the RMSE adaptive-binning method used by Hendrycks et al. (2022) to measure whether models' predictive distributions match their "correctness". Here, "correct" is based on: $\arg\max_{k\in\{1,\ldots,K\}} P_{\text{eval}}(y_n = k|x_n)$.

3. Robustness: loss after a Fast Gradient Sign Method (FGSM) attack is used to measure models' robustness to an adversarial attack (Goodfellow et al., 2014b)[3] Attack strength is run at an $\ell_\infty = 4$ bound following Peterson et al. (2019).

3.3.2 How Best to Learning with Labels from Many Humans?

We first compare our aggregate per-annotator soft labels using all information elicited from annotators (i.e., T2, Clamp) against the *complete* aggregate labels from `CIFAR-10H`, i.e., labels formed from all of their approximately $M = 51$ labelers. Aggregating our soft labels involves:

$$\Psi(\{\hat{p}_m^k\}_{m=1}^M) = P_{\text{agg}}(y_n = k|x_n) = \frac{1}{M}\sum_{m=1}^{M} \hat{p}_m^k$$

[3] We also considered the multi-step Projected Gradient Descent, PGD Kurakin et al. (2016) attack; however, the metric was unstable and warrant further investigation.

We benchmark performance against training on conventional CIFAR-10 hard labels. We also compare against using hard, random, and uniform labels instead. Label smoothing (LS) (Szegedy et al., 2015) as a baseline is discussed in Appendix Section B.1.

As shown in Table 3.1, we recover the findings of (Peterson et al., 2019) that training with soft labels is markedly more advantageous[4]. While better than baselines, training on our soft labels does not outperform labels constructed from many annotators' hard labels.

However, as discussed in Chapter 2, perhaps we are not leveraging the information across many humans well. A naive Ψ is used which does not account for differences in individual humans' calibration nor skill. We hypothesize that learning on *de-aggregated* labels (i.e., just individual labels P_m) may realize the fuller benefits of our collected soft labels. Table 3.2 reveals this is the case: from **8.5x fewer human annotators** than CIFAR-10H, our labels endow trained classifiers with better generalization, calibration, and robustness in nearly all checks.

Not only are these results compelling for highlighting the benefits of eliciting soft labels from every human to support model performance, but they contribute further empirical support that sometimes it is better not to aggregate labels (Wei et al., 2022b). Peterson et al. (2019) found similarly in some settings, which they hypothesize is due to more varied gradient information being available. Alternatively, it could be that our annotators are simply noisier, and keeping annotations separate could enable the model to uncover which labels to "trust" in order to achieve all-around good performance. Indeed, the naive Ψ used in Table 3.1 does not account for differences in individual humans' calibration nor skill. Future work could investigate alternative designs of Ψ, of the kinds discussed in Chapter 2, to construct better training labels; for the remainder of this section, we run all models with de-aggregated labels.

3.3.3 Learning with Labels from Fewer Humans

Thus far, we suggest that eliciting a less common kind of human knowledge from annotators – in the form of soft labels – is more conceptually sound and can advance performance, at least when de-aggregated, and do so from fewer humans queried. We next dig deeper into the hypothesis that richer human-derived soft labels offers a more annotator-efficient path to model performance boosts than the typical many annotator hard label approach of CIFAR-10H.

[4] The good calibration of uniform labels when evaluated on our held-out labels could be due to: 1) the possibility of being calibrated, but "bad" (i.e., nearly always wrong, but representing your confidence in said wrongness), and/or 2) the soft evaluation labels naturally being closer to uniform – given the higher entropy as discussed.

	Label Type	CE	Calibration	FGSM Loss
CIFAR-10H	Hard Labels	2.026±0.18	0.277±0.01	15.455±7.5
	Random Labels	1.77±0.16	0.226±0.02	13.476±1.39
	Uniform Labels	1.599±0.13	0.203±0.02	10.199±3.82
	CIFAR-10H	**1.325±0.07**	**0.201±0.01**	**8.750±1.8**
	Ours (T2, Clamp)	1.369±0.07	0.203±0.01	8.872±1.63
CIFAR-10S	Hard Labels	4.46±0.49	0.425±0.09	15.782±4.67
	Random Labels	3.093±0.53	0.353±0.05	10.697±4.14
	Uniform Labels	2.923±0.26	**0.311±0.06**	11.768±5.18
	CIFAR-10H	**2.558±0.16**	0.313±0.03	**8.416±1.64**
	Ours (T2, Clamp)	2.591±0.19	0.324±0.02	9.116±1.63

Table 3.1 Comparing performance when training on labels aggregated from humans' annotations ($M = 51$ CIFAR-10H annotators, $M = 6$ of ours). Our soft labels here utilize all elicited information from annotators (i.e., T2 Clamp). Different labels are considered over 900 examples. Performance is evaluated over 3000 heldout CIFAR-10H labels (top) and 100 heldout labels from our collection (the T2 Clamp variant of CIFAR-10S, bottom). 95% confidence intervals are included. 5 seeds are run each for three models (ResNet-34A, ResNet-50, VGG-11) and averaged. Bold indicates best performance (lower is better for all metrics).

	Label Type	CE	Calibration	FGSM Loss
10H	CIFAR-10H	1.293±0.08	0.194±0.01	8.577±1.91
	Ours (T2, Clamp)	**1.281±0.06**	**0.184±0.01**	**8.406±1.75**
10S	CIFAR-10H	2.459±0.21	0.311±0.02	**8.334±1.75**
	Ours (T2, Clamp)	**2.355±0.14**	**0.297±0.03**	8.405±1.59

Table 3.2 Training with de-aggregated labels; on each batch, a single humans' label is used as supervisory signal from a pool of M humans ($M = 51$ CIFAR-10H; $M = 6$ ours).

To test this idea, we construct labels using the *same* total number of annotators across both labeling approaches. Note, for compute-resource considerations, the remainder of the experiments in this Chapter, we run 5 seeds of the ResNet 34-A model used in (Uma et al., 2020). In Table 3.3, we subsample $M = 2$ of the annotators in CIFAR-10H from which to construct a label per batch, and compare the utility of learning with said labels against a similarly sub-sampled version over two of our per-annotators' soft labels per image. We find that our labels provide a substantial boost along nearly all metrics – and the gains of our method in the few annotator setting become even more apparent when considering access to only $M = 1$ human. While this is expected, as a single CIFAR-10H labeler is simply a hard label, we demonstrate that **if one has access to only a single annotator, our label**

method provides the best training signal. We depict a full comparison of performance when varying the number of humans employed to construct training labels in Appendix Fig. **??**, which suggests the benefits of our approach.

			M=2			M=1	
	Labels	CE	Calib	FGSM	CE	Calib	FGSM
10H	10H	1.71±0.15	0.25±0.02	14.16±0.67	2.17±0.11	0.29±0.01	19.20±0.86
	Ours	**1.49±0.08**	**0.22±0.01**	**11.57±0.44**	**1.58±0.08**	**0.23±0.01**	**12.57±1.29**
10S	10H	3.37±0.42	**0.34±0.08**	13.13±0.7	4.49±0.31	0.45±0.05	17.85±0.68
	Ours	**2.88±0.17**	0.36±0.04	**11.39±0.59**	**2.90±0.31**	**0.38±0.06**	**12.13±1.31**

Table 3.3 Investigating model performance when fewer M annotators are assumed to provide labels. Training labels are sampled per batch instead from a pool of $M = 2$ or $M = 1$ annotators.

3.3.4 Accounting for Total Annotation Time

In practice, however, the amount of *time* taken by annotators ought to be considered. Here, our method struggles. We follow the same annotator subsampling introduced above and compute the estimated annotation time $M * t_{\text{per}}$ to construct such labels, where t_{per} is the estimated amount of time taken by an individual annotator on a given image. CIFAR-10H annotations for the same examples that we query take approximately 2.5 seconds each to annotator, compared against 32 seconds for ours. Performance accounting for annotation time is visualized in Fig. 3.5. Our labels are somewhat comparable with respect to generalization and calibration in terms of cost efficiency, but struggle with adversarial robustness. Poor time efficiency may in part be due to a cumbersome interface design, given that we collect several bits of information per human, there will always be a trade-off in terms of the richness of the forms of human knowledge elicited and the number of humans one can query.

3.3.5 What Forms of Human Knowledge are Most Beneficial?

We therefore consider what aspects of our elicited labels are most essential. Can we get by with eliciting less information from annotators? We consider a “human knowledge ablation” in the labels used for learning. We train in the $M = 6$ setting over the label varieties introduced in Section 3.2.2. We estimate that each of the “bits” of human knowledge elicited each take 6.4 of the 32 seconds, as we do not have access to time information per bit. We recognize this is likely not accurate and could be considered in future studies. Table 3.4 reveals that the

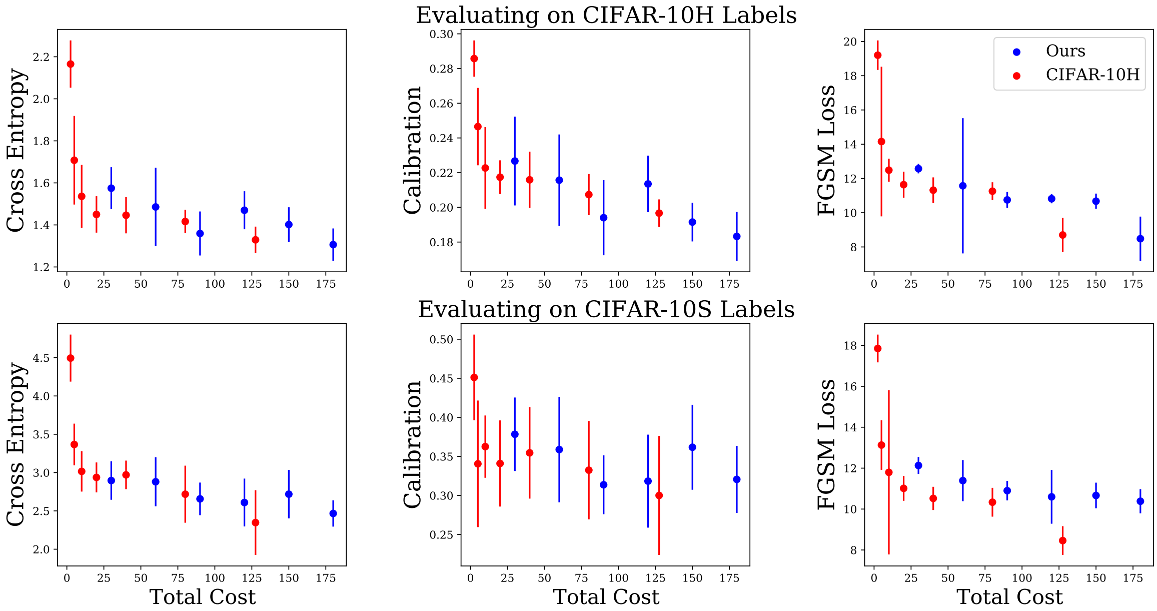


Fig. 3.5 Comparison of learner performance as a factor of estimated total cost of elicitation ($M * t_{\text{per}}$). Red dots depict performance when aggregating M CIFAR-10H annotators for $M \in \{1,2,4,8,16,32,51\}$. Blue dots indicate CIFAR-10S T2 Clamp soft labels, constructed from varying $M \in \{1,2,3,4,5,6\}$. Dots represent performance averaged over 5 seeds. Evaluation is conducted over CIFAR-10H (top) and CIFAR-10S (bottom). Bars indicate 95% confidence intervals over the 5 seeds. Lower is better for all metrics.

best performance is enjoyed by models trained on our full set of human knowledge; however, we likely could get by with just eliciting the most probable label with our clamp if annotation time is a factor in budgeting. This supports the elicitation and incorporation of less common forms of human knowledge like the “impossible label” set-up we introduce. We acknowledge that further investigation is needed into alternative elicitation set-ups – across a range of domains and task difficulty, before drawing broad conclusions.

3.3.6 Impact of Number of Training Examples

All studies have always compared performance with varying the form of subsamplings of 900 of the 1000 labeled examples, and labeling examples not in CIFAR-10S with hard labels. We now compare model performance when training *just* on CIFAR-10H vs. our soft labels. We vary the number of examples we train on N' run for 10 seeds each, and evaluate performance on CIFAR-10H. We do not include evaluation on CIFAR-10S here as the labels are closer to uniform (by virtue of their softness) which results in challenges in disambiguating models which predict uniform from slightly not in this few label setting. We recognize this setting

	Label Type	Time	CE	FGSM Loss	Calibration
10H	T1, Unif	**12.8s**	1.423±0.08	10.891±0.32	0.198±0.01
	T1, Clamp	19.2s	1.368±0.12	10.386±0.8	0.198±0.02
	T2, Unif	25.6s t	1.448±0.14	10.474±0.44	0.203±0.02
	T2, Clamp	32s	**1.306±0.19**	**8.482±3.95**	**0.183±0.03**
10S	T1, Unif	**12.8s**	2.531±0.15	11.25±0.4	**0.312±0.03**
	T1, Clamp	19.2s	2.637±0.31	10.423±1.17	0.328±0.06
	T2, Unif	25.6s	2.63±0.29	10.667±1.15	0.339±0.1
	T2, Clamp	32s	**2.466±0.32**	**10.382±1.0**	0.321±0.07

Table 3.4 Training models over labels constructed from subsets of the human knowledge we elicit. $M = 6$ humans' labels are used to form the pool sampled.

is not ideal given how few labels we have; indeed, adversarial robustness trend here is peculiar: performance gets *worse* when training on more labels; we reason this may be due to some poorly performing models also being robust (Tsipras et al., 2018). However, even in this regime, we note that training on few of our labels seems to offer benefits over few CIFAR-10H.

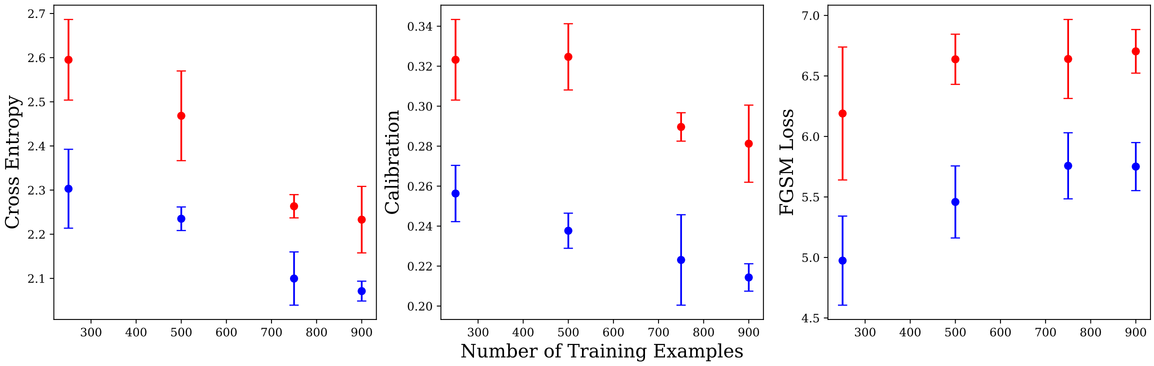


Fig. 3.6 Impact of varying the number of soft-labeled examples ($N' = 250, 500, 750, 900$) during training. Evaluation is run over heldout CIFAR-10H labels. Lower is better for all metrics.

Future work could consider how many examples are worth querying. For instance, Paul et al. (2021) find that substantially fewer CIFAR-10 examples can be used in training to achieve comparable performance. Alternatively, we could also consider *simulating* more of our labels by leveraging the statistics of the human knowledge we have collected. We explore this direction in Chapter 5 and Appendix Section B.4.

3.4 Takeaways

In this chapter, we have addressed a gap in the kinds of information elicited from humans: typically, annotators are only asked to indicate what they believe to be the most probable class of an example. We take a step towards filling this gap by collecting a new dataset, `CIFAR-10S` consisting of soft labels elicited from every annotator. We demonstrate the value of leveraging this form of human knowledge towards improving model generalization and calibration. We empirically find that our labels offer a path towards more annotator-efficient labeling. Such results could be incredibly important in situations **where few annotators are available**, e.g., in domains requiring specialty expertise like medicine or criminal justice, or in interactive personalization where there may only be a single annotator providing examples.

Next In the next chapter, we continue our quest of exploring the benefits of eliciting new forms of human knowledge to improve model reliability and general trustworthiness – turning now to the question of human perceptual judgments over *synthetic* examples.

Chapter 4

Inquiring for and Incorporating Human Knowledge over Synthetic Examples

We have just seen how the information we collect from humans to form our training dataset impacts a model's performance. However, thus far we have focused on eliciting and incorporating information from humans over the standard examples in datasets, i.e., the "typical" `CIFAR-10` images. A sensible next question is whether we there is value in extracting human knowledge over *synthetic* examples for use in training? We address this question in this chapter. While there are many kinds of synthetic data that we could inquire for human judgments about; here, we focus on convex combinations of images – the same kind of data used in *mixup* (Zhang et al., 2017).

Chapter Roadmap We first introduce *mixup* and appeal to why the method is worth comparing against human knowledge. We then formalize the kind of information that we aim to collect from humans and introduce the crowdsourced study designed to obtain such information. We address our hypothesis that human perceptual judgments over synthetic examples holds value by 1) showing that human percepts of mixed examples indeed differs from the kinds of labels *mixup* assumes, and 2) computationally verify that leveraging the information we elicit can improve model performance along a suite of metrics.

4.1 Background: *mixup*

4.1.1 Why Care?

As we reviewed in Chapter 2, there is a paucity of work addressing whether synthetically-generated data align with human perceptual knowledge, and if not, how human knowledge could be used to construct better synthetic examples. *mixup* (Zhang et al., 2017) is a training procedure wherein a neural network is trained on synthetic examples which are formed through a linear combination of pairs of training points. Given the simplicity of the *mixup* generative process (see Section 4.1.2), we consider it a reasonable case study for the value of eliciting and incorporating human knowledge over the generative process (i.e. here, in the mixing coefficient λ and the resulting synthetic label $\tilde{y}$).

Despite this simplicity, *mixup* is a powerful and popular training-time method which has been leveraged to ensure models have improved fairness (Chuang and Mroueh, 2020) and better gradient-based explanations (Kim et al., 2020b), and been proven to increase model robustness via implicitly regularizing the form of category boundaries learned (Zhang et al., 2020) and to improve calibration (Thulasidasan et al., 2019). Therefore, studying whether the form of labels align with human percepts is a worthy contribution in its own right, which has yet to be studied.

Prior work in human categorical perception demonstrating that humans show non-linear "warping" effects along category boundaries (Folstein et al., 2013; Goldstone and Hendrickson, 2010; Harnad, 2003) leads us to believe that humans will differ in their percepts from the linear category boundaries encouraged by *mixup*. And if humans do differ, it could inform the design of more effective mixing policies. Already, many alternative *mixup* input and target mixing functions have been proposed (Hendrycks et al., 2022; Kim et al., 2020a, 2021; Verma et al., 2018; Yun et al., 2019). Closest to our work, Sohn et al. 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., 2019; Cascante-Bonilla et al., 2020; Sohn et al., 2020). Similarly, Between-class (BC) learning (Tokozume et al., 2017a,b) proposes hand-crafted adjustments to the creation of augmented examples to better align with human perception; however, no human studies are actually conducted to verify alignment. To our knowledge, no previous works have *directly* considered incorporating humans in-the-loop for either the construction of *mixup* samples, or associated relabeling.

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, along with impact on model performance. Moreover, seeing as synthetic examples are sometimes "odd" and abnormal, we postulate this domain is a **prime candidate for utilizing human confidence** to inform downstream use of elicited human knowledge.

4.1.2 Constructing Synthetic Training Examples

mixup entails only training on synthetic examples $(\tilde{x}, \tilde{y})$, which are formed via convex combinations of pairs of the training observations $(x_i, y_i), (x_j, y_j$ for $i, j \in \{1, ..., N\}$.

$\tilde{x}$ can be viewed as the output of a *mixup* policy [1]. f, defined over the inputs ($f(x_i, x_j, \lambda_f) = \tilde{x}$), and $\tilde{y}$ is similarly the output of a policy g defined over the targets ($g(y_i, y_j, \lambda_g) = \tilde{y}$). Zhang et al. parameterize these policies via a mixing coefficient λ, sampled from a Beta distribution controlled by parameter α (e.g., $\lambda \sim \text{Beta}(\alpha, \alpha)$), and let $\lambda_f = \lambda_g = \lambda$. In the original *mixup*, f and g are pre-defined to be equivalent linear combinations of the corresponding observations:

$$f(x_i, x_j, \lambda_f) = \lambda_f x_i + (1 - \lambda_f) x_j = \tilde{x}$$

$$g(x_i, x_j, \lambda_g) = \lambda_g y_i + (1 - \lambda_g) y_j = \tilde{y}$$

In this work, we refer to λ_f as the **generating mixing coefficient** and the labels y_i, y_j of the mixed examples as the **endpoints**.

4.2 Human Percepts vs. Synthetic Labels

While Zhang et al. (2017) employ linear *mixup* policies f, g over both the input and targets of the original training examples, manual inspection of *mixup* samples in the image domain suggests that these synthetic points may not be consistent with human perception. Practitioners could optionally customize the *mixup* policy over the observations (f) or the targets (g). Here, we focus on the latter: utilizing human input to design a perceptually-aligned target *mixup* policy g_h. A schematic of our approach is depicted in Fig. 4.1.

We assume f is mixed to be the linear mixing policy over inputs employed in (Zhang et al., 2017). 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 notice that while we assume $\tilde{x}, y_i, y_j$ are fixed, nothing requires us to have λ used in g be the same as in f. We therefore introduce the notation that there are *two* separate mixing coefficients: one for the input policy λ_f and one for the target policy λ_g. We therefore 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

[1] We employ the nomenclature and notation around "*mixup* policies" from (Liu et al., 2021b).

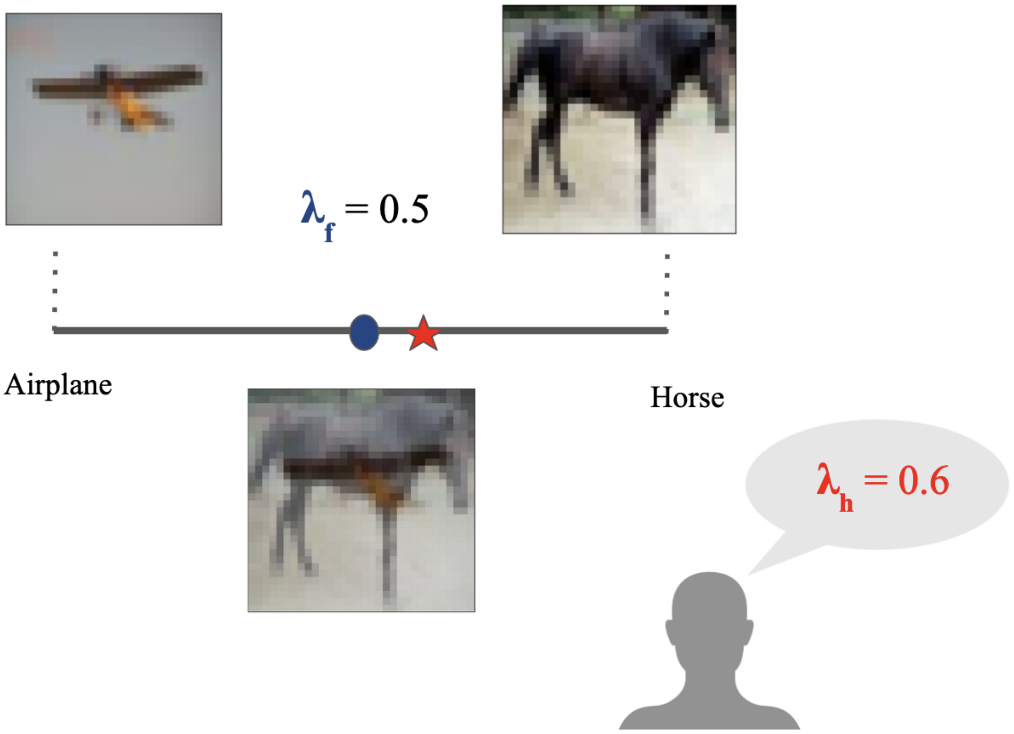


Fig. 4.1 Illustration of our elicitation paradigm. Images drawn from two different categories are mixed. The human is shown the mixed image and the category labels, and asked to provide an inference (red) for the generating mixing coefficient λ_f (blue).

underlying labels y_i, y_j. Individuals are 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 $\tilde{x}$.

4.2.1 Elicitation Paradigm

To elicit such information from humans, we design an interface wherein subjects infer the mixing coefficient between two given labels. We show each worker a mixed image $\tilde{x}$ and tell them the categories that were mixed to generate the image. Additionally, participants 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 We sample a random subset from the same 7,000 set of `CIFAR-10` examples discussed in Section 3.3. We sample 6 pairs of images per unique class combination (e.g., dog-cat, dog-truck, ship-airplane, etc). The `CIFAR-10` hard labels as the label endpoints. For each of these pairs of images, we generate an interpolated version from 3 mixing coefficients – 0.5, and one chosen randomly from each of the sets $\{0.1, 0.25\}$ and $\{0.75, 0.9\}$, respectively. As there are 45 such class combinations, we result in $N = 810$ total stimuli. Combined images are mixed in pixel space, and retain the 32x32 image resolution of the original `CIFAR-10` set.

Crowdsourcing We run our relabeling experiment on a total of 33 participants. Each participant sees $59-60$ images, where two images are repeated to measure raters' internal consistency. Repeats are placed at the end, and correspond to the images presented on trials 15 and 20, respectively. Images and endpoint label orderings are shuffled across participants. An example survey screen can be seen in Fig. A.2. At least two participants saw each image.

4.2.2 (Mis)-Alignment to Human-Inferred Mixing Coefficient

We compare the elicited λ_h against λ_f, and analyze participants' confidence ω in such inferences. We also conduct a preliminary exploration into the relationship between participants' predicted confidence and the label entropy.

Relationship between Generating Mixing Coefficient and Alignment Averaging over all images reveals a remarkable alignment with the underlying mixing coefficient. As depicted in Fig. 4.2, on aggregate, participants recover close to the generating coefficient when considering the median. This may suggest that the mixing coefficient is aligned with human perception. However, wide error bars and a closer look at how individual images would be relabeled (see Fig. 4.3) uncovers significant deviations; human perception is *not* consistently aligned with the mixing coefficient lead us to believe that such calibration is likely due to averaging effects which may cancel out differences in participants' percepts. Practically, this raises questions as to whether we should train on *de-aggregated data* to reflect the disagreement across individuals like we found beneficial in Chapter 3. We explore this idea in Section 3.3.2.

Moreover, inspecting the inferred mixing coefficient at a category level in Fig. 4.4 reveals that when broken down into class pairings, there are significant deviations from the expected

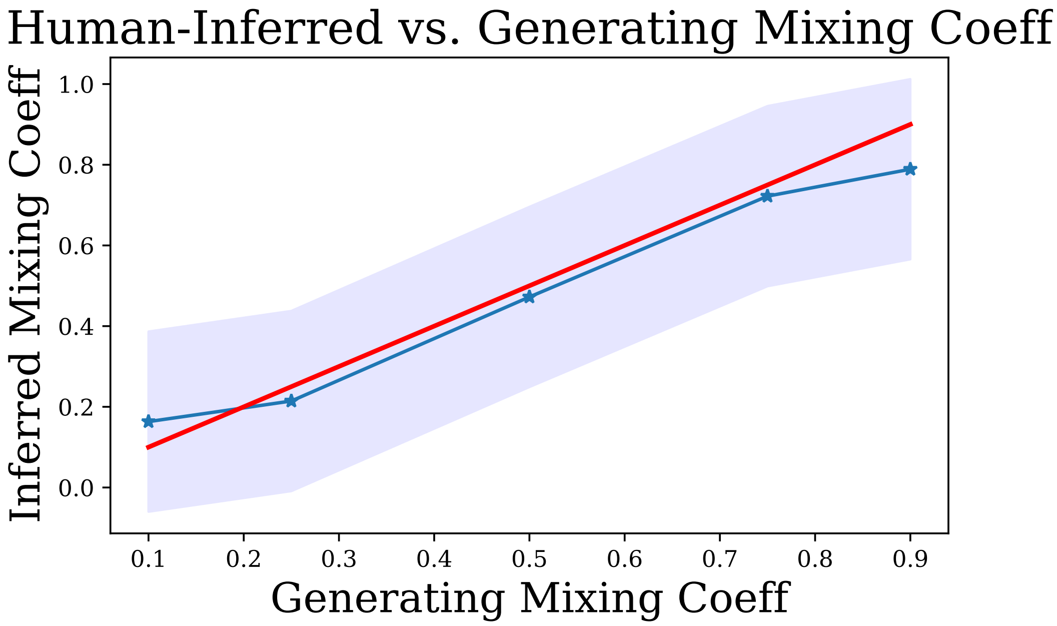
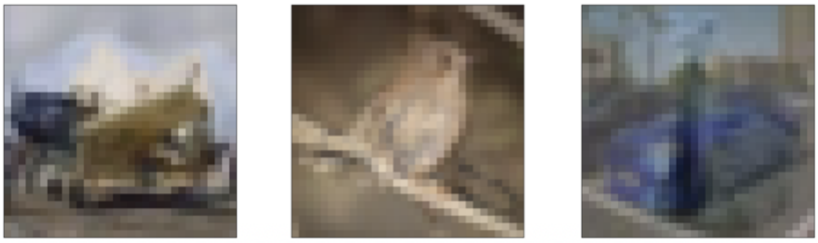


Fig. 4.2 All human-inferred mixing coefficients compared against the mixing coefficient used to generate the image (blue). Red line indicates what perfect alignment with the generating coefficient would look like. Although when mixed, these data look remarkably well-calibrated, we reason this is due to significant averaging effects.



	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

Fig. 4.3 Example actual average human relabelings of the generating mixing coefficient.

linearity. These findings corroborate non-linearities found in human categorical perception (Destler et al., 2019; Folstein et al., 2013; Goldstone and Hendrickson, 2010; Harnad, 2003). We also uncover that participants' confidence in the "correctness" of their inferred mixing coefficient is lowest at $\lambda_f = 0.5$.

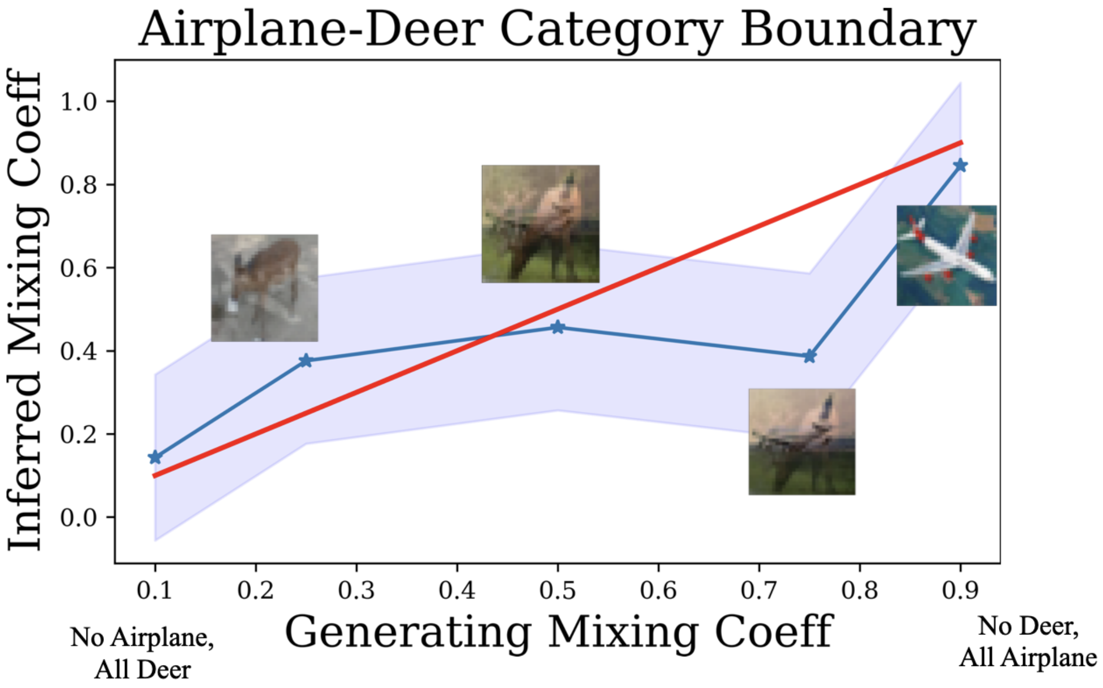


Fig. 4.4 An example "category boundary" extracted from elicited from people diverge from linearity in the generating coefficient. λ_f is depicted against $_h$ (where the mixing coefficient is varied between all deer and all airplane). The red line indicates what an exact parallel between inferred and generating mixing coefficient would look like (highlighting perceived human deviation).

Decomposing Human Confidence by Endpoint Entropy 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 compare our annotator 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[2]. We find in Fig. 4.5 that this is the case when considering confidence – if

[2] While we ought to use `CIFAR-10S` for reasons discussed in Chapter 3, we use `CIFAR-10H` given there is complete coverage of the labels we have relabeled.

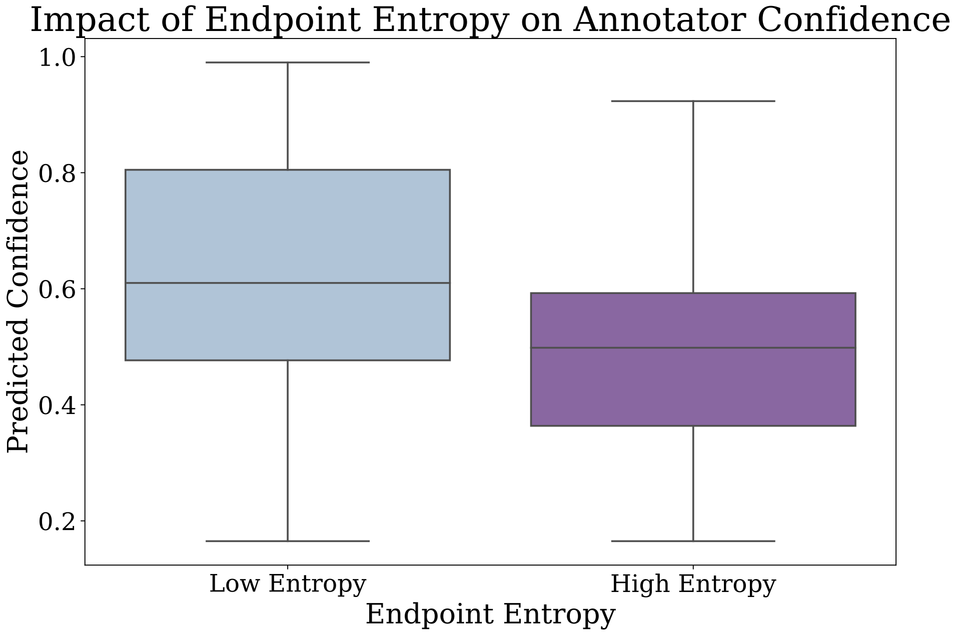


Fig. 4.5 Confidence reported by annotators in their inference of λ, as a factor of whether the combined labels y_i, y_j are high or low entropy. Entropy is measured over the CIFAR-10H human-derived labels.

both endpoints are very high entropy under CIFAR-10H (i.e., $H \geq 0.5$), participants report markedly lower confidence in their inference than if both endpoints have low entropy ($H \leq 0.1$). 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.

4.3 Learning with Human Perceptual Knowledge over Mixed Examples

Our crowdsourced study indicates that indeed, human perceptual judgment does not consistently correspond to the classical *mixup* target policy. The fact that there is a discrepancy **suggests that eliciting human knowledge over this kind of synthetic example indeed has**

value, at least in helping verify the alignment of human perception which is important for model trustworthiness (Nanda et al., 2021).

We now consider whether human knowledge over these kinds of examples also holds value in improving the supervisory signal used to train models, compared against those typically used in *mixup*. We hypothesize that constructing target label distributions from said human-elicited labelings could yield more perceptually consistent classifiers – which could enjoy better generalization, calibration, and robustness.

To test this idea, we compare the performance of classifiers trained on different labelings over the synthetic mixed examples. Ideally, we would compare using human relabelings for every synthetic image that would be generated when employing *mixup*, in light of our finite dataset, we investigate the impact of using our labels versus the traditional *mixup* labels over a *finite, augmenting set* of the combined images.

Despite the artificiality of this setting, which we acknowledge renders it challenging to draw concrete conclusions, we can still ask interesting questions, which are reminiscent of Chapter 3, namely: 1) do average human relabelings already improve performance in the augmenting set paradigm, 2) how best should we manage information from multiple different humans (is it actually best to keep labels separate?), and 3) what form of human knowledge (λ_h, ω) is most beneficial in this inference? We address each in turn after first discussing the training and evaluation set-up used.

4.3.1 Setup

Model and Data

We rely on the original *mixup* code provided by Zhang et al. (2017) and use the same PreAct ResNet-18 model (He et al., 2016) and training procedures, i.e., train for 200 epochs with a learning rate of 0.1, which is reduced by a factor of 0.1 after epochs 100 and 150, respectively. We train over the 7,000 images of the `CIFAR-10` test set split that was detailed above, with the optional inclusion of the 810 augmenting mixed images. While we could have trained on the entire `CIFAR-10` train set, we chose this design as it is small enough to let us more concretely elucidate the impact of various labelings over the finite augmenting set, and allows us to readily swap in `CIFAR-10H` or `CIFAR-10S` labels later (which, as noted in Chapter 3, are over the `CIFAR-10` test set). Note, we similarly do not apply any data augmentation, such as rotation or cropping during training that was used by (Zhang et al., 2017), as to best disentangle the impact of using human-derived labels over the raw mixed images participants had seen. We train each model over 5 random seeds.

Evaluation Suite

We consider the same suite of evaluation metrics discussed in Section 3.3.1. We evaluate performance on the held-out set of 3,000 images of CIFAR-10H. While as noted in Chapter 3, CIFAR-10S may be a more suitable test set, given insufficient amount of labels, to ensure stability we focus on CIFAR-10H. Extending these analyses to CIFAR-10S is grounds for future study.

Baselines

We compare our human-derived labels against the traditional *mixup* labels (those that use the generating mixing coefficient) on the augmenting set, as well as labeling the augmenting set with uniform labels over the 10 CIFAR-10 classes and random labels. We also provide performance trained for a model without using the augmenting set (e.g., just over the CIFAR-10 images and their associated hard labels without any mixing).

4.3.2 Comparing Averaged Human Perceptual Judgments Against *mixup* Labels

We first consider whether replacing λ_f with the *average* mixing coefficient $\bar{\lambda_h}$ improves performance:

$$\Psi(\{\lambda_h^m\}_{m=1}^M) = \frac{1}{M}\sum_{m=1}^{M}\lambda_h^m = \bar{\lambda_h}$$

Seeing as in Fig. 4.2, the amount of relabeling is somewhat comparable when considered in aggregate, we expect that there may be minimal difference in performance. We find in Table 4.1 that this is the case: calibration and FGSM are comparable, though surprisingly held-out cross entropy is worse – though within error bars. We see though that, as expected, using either synthetic mixing policy is better than "meaningless" alternatives (i.e., uniform or random labels); we acknowledge that the high performance of the baseline model, however, points to the insufficient size of our augmenting set and broader experimental paradigm. We therefore emphasize that insights drawn are then with respect to various ways incorporating human knowledge impacts performance *in this limited setting.*

4.3.3 Learning with Knowledge From Many Humans

Given the somewhat poor performance of using naive-average human relabelings for the synthetically mixed examples, we next investigate one of the central themes of this work:

Labeling Scheme	CE	Calibration	FGSM Loss
Regular (No Aug)	2.25±0.04	0.29±0.01	14.53±0.22
+ Uniform Labels	2.56±0.05	0.29±0.01	18.54±0.31
+ Random Labels	2.66±0.05	0.3±0.01	19.75±0.31
+ *mixup* Labels	**2.19±0.04**	**0.28±0.01**	**14.35±0.21**
Ours (Agg, $\bar{\lambda}_h$)	2.27±0.04	**0.28±0.01**	15.07±0.22

Table 4.1 Comparing performance of models trained on different labels for the N=810 augmented synthetic images ($\tilde{x}_n$). A baseline of training without augmentation ("No Aug") is considered. Evaluation is run over `CIFAR-10H`.

when we have information from many humans, is it worthwhile to aggregate for learning? Here, we reason that averaging annotator feedback in this case may be cancelling out important signal, or could be due to having too few annotations – in which case aggregation may be ill-advised according to Wei et al. (2022b). Alternatively, this could be due to people focusing on different aspects of the image, such that the average is not representative of how the human visual system overall processes the image. In Table **??**, we see that when de-aggreated, like the explorations we were conducting in Section 4.3.2, human percepts differ from linear mixing in ways that are advantageous for model performance.

Labeling Scheme	CE	Calibration	FGSM Loss
mixup Labels	2.19±0.04	0.28±0.01	14.35±0.21
Ours (Agg, $\bar{\lambda}_h$)	2.27±0.04	0.28±0.01	15.07±0.22
Ours (Sep, λ_h^m)	**2.01±0.0‘3**	**0.27±0.01**	**13.33±0.19**

Table 4.2 Training separately individual annotator (m) vs. average-aggregate human relabelings, vs. *mixup* labels vs. average human relabelings.

4.3.4 Leveraging Human Confidence

We have only focused on leveraging λ_h; we also elicited self-reported confidence ω in said inference. We hypothesize that modulating labels with human confidence could provide a more effective supervisory signal. We smooth the label based on an exponentially decaying transformation (i.e., smoothing = γ^{ω}) of the predicted confidence (ω). This is done over aggregate labels using the mean human confidence for any example $\bar{\omega}$, or separately applied per-annotator confidence (ω_m). Smoothing is designed such that if an annotator has zero confidence in their inference, a uniform label over all 10 `CIFAR-10` classes is used and the traditional two-class mass policy, with human-inferred mixing coefficient, is used if the annotator is 100% confident. In this case, we only apply significant smoothing if an annotator is very uncertain in their inference (achieved by setting $\gamma = 0.005$) [3].

Training instead using labels formed from both types of human knowledge leads to substantial boosts in generalization, calibration, and robustness over using the traditional linear mixing coefficient-based labels (see Table 4.3). Human confidence provides a powerful indicator as to whether the example is "confusing" and hence ought to have a more uniform label. This not only highlights the promise of eliciting human confidence in the construction of machine learning datasets to be used to craft training labels, but further underscores potential flaws in the traditional linear target *mixup* policy.

Labeling Scheme	CE	Calibration	FGSM Loss
mixup Labels	2.19±0.04	0.28±0.01	14.35±0.21
Ours (Agg, $\bar{\lambda}_h$)	2.27±0.04	0.28±0.01	15.07±0.22
Ours (Sep, λ_h^m)	2.01±0.03	0.27±0.01	13.33±0.19
Ours (Agg, $\bar{\lambda}_h$ with $\bar{\omega}$)	2.09±0.04	0.27±0.01	14.01±0.22
Ours (Sep, λ_h^m with ω_m)	**1.83±0.03**	**0.24±0.01**	**11.81±0.19**

Table 4.3 Leveraging individual (sep) and averaged confidence ω with inferred mixing coefficients.

4.4 Takeaways

In this chapter, we demonstrate that eliciting human knowledge over the synthetically constructed examples, namely those used in *mixup*, has value. This value comes from 1) providing human-grounded support for the development of alternative label mixing policies to *mixup* as human perceptual knowledge appears to systematically differ from what is typically assumed when synthesizing $\tilde{y}$, and, 2) empirically – at least within our constrained augmenting set paradigm – leveraging human perceptual knowledge to construct more suitable $\tilde{y}$ seems to hold promise towards improving model performance compared against a model trained with the traditional mixing coefficients. We find that *not* aggregating over humans' provided inferences is preferable, and uncover that human confidence ω is a particularly potent signal for designing effective training labels. These data illustrate that collecting additional human knowledge *over synthetic examples* has value, though in light of the limited amount of data collected thus far, scaling is needed before conclusive insights can be drawn.

[3] Alternative confidence-based smoothing measures could be considered in future work.

Next Given human participants are highly uncertain about the underlying mixing coefficient in a number of cases, we next 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 according to humans. We begin to address the challenges of scaling, considering alternative ways to encourage human-like softness in model predictions and demonstrating the potential value of simulating additional human knowledge.

Chapter 5

Synthesizing Human-Like Soft Target Distributions

We next bridge the forms of human knowledge that we have collected thus far: soft labels from every annotator (Chapter 3) and human perceptual judgments with associated confidence over synthetic examples (Chapter 4), to address more comprhensively whether the structure of the synthetic *mixed* examples is sensible. We then consider how to encourage a model to produce human-like labels, even in the absence of sufficient annotations.

Chapter Roadmap We first conduct a preliminary follow-up crowdsourced study to verify that indeed, humans perceive mixed examples to cover a broader spectrum of image categories than are traditionally used in *mixup*. We discuss how the human-derived soft labels we collect over $\tilde{x}$ validate various of extensions of *mixup* which are already being developed by the machine learning community. We then introduce our two proposed extensions of *mixup*: 1) a new entropy-weighted loss, inspired by the human knowledge we have collected, and 2) the use of soft label endpoints for mixing. We discuss how *simulating* human knowledge, grounded in the data we have collected thus far, can enable us to scale the benefits of incorporating more human-aligned information.

5.1 Humans Perceive "Softer" Category Compositions for Synthetic Mixed Images

We saw in Chapter 4 that not only do human perceptual judgments do not consistently align with the mixing coefficient used to form the typical mixing label $\lambda_f * y_i + (1 - \lambda_f) * y_j$, but humans are also not particularly confident in their inferences. We hypothesize that such discrepancies and confusion may indicate arise when humans perceive different categories than are used to form the synthetic image. Recall, we have thus far only told annotators the labels of the combined images; it could be the case that the result of combining a cat and a ship now looks like something else entirely. If so, it would be sensible to express the richness of these percepts in the supervisory targets.

To investigate, we consider whether human knowledge deviates from the traditional "two-hot" label formation approach[1]. We saw in Chapter 3 that soft labels can be effectively elicited from every annotator and offer ML performance boosts; here then, we explore the collection of soft $\tilde{y}$ directly from humans.

5.1.1 Elicitation Paradigm

We rely on a modified version of the soft label elicitation interface proposed in Chapter 3 (see Fig. A.3). We recruit 8 participants, yielding soft labels over a total of 100 mixed images. The images are drawn from the same set of stimuli created in Section 4.2.1; however, here, we only show images with a mixing coefficient $\in \{0.25, 0.5, 0.75\}$, as these yield the most interesting relabeling effects. 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 and 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.

5.1.2 Human-Inferred Category Composition

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*). We are encouraged that the task is reasonable as humans did tend to place probability mass on the generating endpoints that correlated with the mixing coefficient used (Pearson $r = 0.52$). However, we interestingly find that the humans' provided soft labels place 38.3% (±0.6%) of the probability mass of a label 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; and hence, alternative labelings may be preferred which are more aligned with human percepts. Examples of such labeled mixed images are shown in Fig. 5.1.

[1] We refer to labels which place all mass on only two labels as "two-hot" to parallel traditional "one-hot".

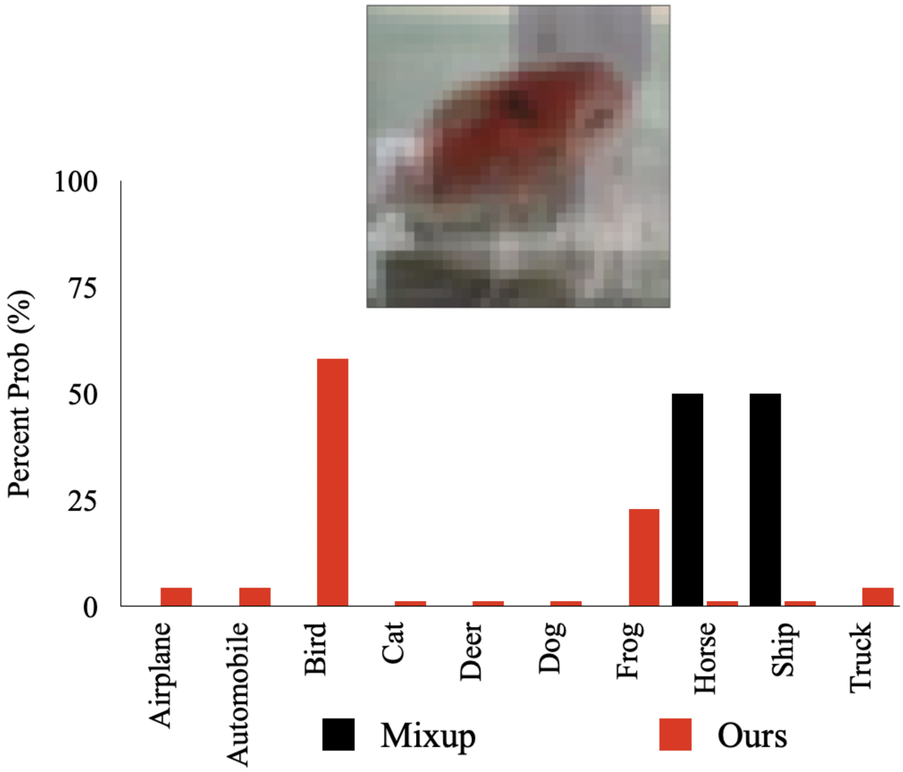


Fig. 5.1 Example combined image ($\lambda = 0.5$; horse/ship) which has been relabeled by humans (red) using our soft label elicitation. The label which would be used by *mixup* is shown in blue.

Takeaways The typical two-class 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.

5.2 Softening Synthetic Targets

We next consider how best to leverage the insights of our human study, i.e., human perceptual knowledge as soft labels indicate the traditional two-hot category composition for synthetic $\tilde{y}$ lacks human-like richness. It is therefore natural to investigate alternative training paradigms to those used in classical *mixup* that ascribe probabilities across a wider range of categories to those mixed.

In the face of limited data ($N = 100$ examples), rather than training directly on the elicited soft labels, we consider ways to *automate* the introduction of softness into the mixing policy. First, we connect the results of our user study to existing automated *mixup* approaches that leverage softness. We then introduce a new entropy-weighted loss inspired by the human knowledge elicited and empirically verify its utility towards improving generalization. We also consider a second way to incorporate softness in $\tilde{y}$ by mixing naturally soft endpoints, for instance, by mixing over (and simulating additional) `CIFAR-10S` examples.

5.2.1 Soft *mixup* Policies

Our preliminary elicited human knowledge suggests that encouraging the network to produce softer predictions when mixing inputs may be more sensible. Several works have already begun to address the softening of *mixup* labels, KD methods (Wang et al., 2020; Xu et al., 2020), GMM-based soft relabelers (Sohn et al., 2022), and entropy-based loss functions such as that proposed in Mix-MaxEnt (Pinto et al., 2021). Our elicited data in the Section above excitingly **provides the first human perceptual verification** of such approaches being (loosely) more aligned with human perceptual knowledge than classical the *mixup* label policy.

Here, we add to this family of soft *mixup*-based methods with two variants: a new clamped entropy-weighted loss (Ent-WC), and mixing over softer endpoints. As we do not employ an auxiliary relabeler like the KD methods, our work then most closely relates to Mix-MaxEnt. We next delve deeper into Mix-MaxEnt and its relation to our proposed method.

Overview of Mix-MaxEnt in Relation to Our Method

To address improve model uncertainty calibration, Pinto et al. (2021) propose training a model over both classical in-distribution x and synthetic examples $\tilde{x}$, where the model is optimized with a traditional cross-entropy loss over the in-distribution examples and encouraged to drive up the entropy over the synthetic examples. In their work, synthetic examples – like ours – are constructed via *mixup* (Zhang et al., 2017); however, the constructed examples hem close to the 50/50 point (i.e., $\lambda_f \sim 0.5$).

This setup is compelling as it captures the intuition that we have begun to confirm with human experiments: at least in the image domain we have considered, humans tend not to have confident muddied perceptual judgments over synthetically mixed examples. It is therefore reasonable to encourage spreading probability mass over a range of classes for these examples.

However, we argue there are several assumptions in the design of the set-up which could be address. First, such softening takes place primarily over the midpoints of the mixes. As we saw in Sections 4.2 and 5.1, not just the 50/50 point is “confusing” for humans. Second, in Mix-MaxEnt, the model is encouraged to drive the entropy to a fully uniform distribution – provided performance remains high on in-distribution examples. Yet, as we have also seen through our human experiments, there are several cases where humans are *not* confused. Therefore, it seems sensible to encourage softer, but not meaningless, targets over synthetic examples. Third, training with cross-entropy over the in-distribution examples assumes that they are "flawless." The authors consider experiments using only hard labels, and as we demonstrated in Chapter 3, individual human percepts are also more uncertain than can be captured with hard labels alone. It seems sensible then that entropy could be increased over the in-distribution examples as well.

5.2.2 Our Method: Clamped Entropy-Weighted Loss

As such, in this work, we clamp[2] the entropy adjustment to ensure the model is not incentivized to unrestrictedly predict uniform labels. We also consider this loss in a set-up which mimics *mixup* in that we never train directly on the in-distribution examples. We also consider this loss in a set-up which mimics *mixup* in that we never train directly on the in-distribution examples. We believe this allows us to better take advantage of the regularization benefits conferred from *mixup* (Zhang et al., 2017) that Mix-MaxEnt may miss. Our training setup is therefore constructed as follows: We weight the standard CE loss of each data point by $\log(2)/H_\theta(\tilde{x})$, where H_θ is the predictive entropy of the model trained on mixed datapoint $\tilde{x}$. The weight is capped at 1, and lower bounded by a threshold, $\kappa \geq 0$. We also sample λ_f from a broader space of mixing coefficients.

Setup

We employ a similar training and evaluation paradigm to that used in Chapter 4, e.g., training a PreAct ResNet-34 model. However, in the experiments in this Chapter, we run *full* input mixing over all $N = 7,000$ examples, rather than just over an augmenting set. That is, at each batch, we sample a new mixing coefficient λ_f, drawn from $\text{Beta}(\alpha, \alpha)$. Here, we let $\lambda_f = \lambda_g$, a simplification that does violate our findings of Chapter 4 (preliminarily addressed in Appendix C). To further disentagle model performance, we also consider an additional robustness check – to *natural* data shifts. We therefore also measure CE on corrupted images (corruption strength 3) from `CIFAR-10-C` Hendrycks and Dietterich (2019).

We compare our clamped, entropy-weighted loss against three baselines: 1) hard label, without any *mixup*, 2) Mix-MaxEnt, and 3) *mixup* with our new clamped entropy-weighted

[2] This is a different "clamp" than Chapter 3; this is not human-derived and pertains to constraining the loss adjustment.

loss. The Mix-MaxEnt authors (Pinto et al., 2021) found that a higher α was advantageous. To give the fairest comparison to the approach, we re-tuned for our specific problem, using the same validation method in Chapter 3. We select $\alpha = 10$ from $\{5, 10, 20, 50\}$, and run a similar process for our entropy-weighted loss finding best performance at $0.2 \in \{0.1, 0.2, 0.4, 1, 5, 10\}$.

For our method, we select clamp κ from {0.0, 0.2, 0.5, 0.7, 0.9 } using a held-out validation set of CIFAR-10. We find the best performance is conferred with $\kappa = 0.7$, indicating that a moderate amount of entropy is ideal. Too strong of an incentive to lower entropy (e.g., $\kappa = 0$) leads the model to predict uniform labels, which is not consistent with the human knowledge we have collected as discussed above.

Results

We find that our entropy-weighted loss significantly improves model generalization and robustness compared to traditional *mixup* label policy (Zhang et al., 2017); however, the approach yields somewhat poorer calibrated predictions, particularly with regard to Mix-MaxEnt (Pinto et al., 2021). This is not entirely unexpected as Mix-MaxEnt is designed for good calibration (Pinto et al., 2021), and as discussed throughout this work, we care about performance over a range of metrics. We do note that Mix-MaxEnt yields the best *natural* robustness, painting a nuanced picture of the advantages of our method. However, the success of both Mix-MaxEnt and our new entropy-weighted loss demonstrate the power of encouraging moderate levels of softness over *mixed* examples, across the full space of mixing coefficients. While we do not directly verify this model with human knowledge, we believe the approach holds promise as a potentially more perceptually-consistent training approach when using synthetic, *mixup* inputs.

5.2.3 Our Method: Mixing Over Human-Derived and Simulated Soft Labels

We can go further though and address another assumption common in *mixup* work: the endpoints being mixed are hard. That is, instead of construction $\tilde{y} = \lambda_f * y_1 + (1 - \lambda_f) * y_2$ (Zhang et al., 2017) from hard y_1, y_2, we assume y_1, y_2 are *soft*. Where do we get these soft labels from? Our human experiments in Chapter 3 provide an "answer," at least for CIFAR-10. Here then, we consider running *mixup* and our entropy-weighted loss to explore more directly incorporating huamn knowledge to form softer synthetic labels.

To the best of our knowledge, we are the first to explore the use of soft variants of CIFAR-10 labels as the endpoints for mixing. While Battleday et al. (2019) begin to investigate the use of their dataset `CIFAR-10` and *mixup* in their prior submission, here we extend these analyses across a wider range of metrics and consider how softness in the endpoints interacts with encouraging softness in the model predictions through an entropy-weighted loss.

Algorithm	CE	FGSM Loss	Calibration	Corruption CE
mixup	1.252±0.02	10.547±0.26	0.087±0.01	1.606±0.03
Mix-MaxEnt	1.070±0.04	6.436±0.59	**0.065±0.01**	**1.340±0.03**
Ours (Ent-WC)	**1.054±0.01**	**3.314±0.05**	0.102±0.0	1.346±0.03

Table 5.1 Comparing variants of the *mixup* algorithm.

Extending `CIFAR-10S`

A challenge arises though in exploring the use of `CIFAR-10S` in our set-up. To parallel the experiments of Chapters 4 and 5 thus far, we want to train over the set of 7,000 examples – however, we only have soft labels for 1,000. We now consider another utility of the human knowledge we have collected: we can *simulate* more of our labels over the 6,000 which we do not have per-human soft labels elicited for. We do this by softening the annotations from `CIFAR-10H` using the statistics of our collected data. We construct a generative model to mimic the human elicitation process. We detail said model in Appendix B.5. We hypothesize that the use of human-grounded simulated soft labels, can improve model generalization and calibration. We empirically explore this hypothesis next.

Setup

We compare training a model with the above algorithms when the endpoints are hard (traditional) versus soft (using our simulated expanded `CIFAR-10S`) in the same training paradigm.

Results and Discussion

Table 5.2 reveals that mixing over soft labels confers sizable improvements for traditional *mixup* and Mix-MaxEnt; particularly in calibration and corruption robustness. However, the soft endpoints yield inconclusive performance change when using our entropy-weighted loss. This is interesting, as it suggests that possibly the entropy-weighted loss is able to capture much of human softness already. Though the top-tier performance of Mix-MaxEnt when combined with our soft labels supports that this extension to classical *mixup* (incorporating human knowledge – and specifically human-grounded simulated soft labels) holds promise. Though as discussed in Sections 6.2 and B.5, the use of simulated soft labels is preliminary and further investigation is essential to ensure the fidelty of the resulting labels.

Algorithm	Endpoints	CE	FGSM Loss	Calibration	Corruption CE
mixup	Hard	1.252±0.02	10.547±0.26	**0.087±0.01**	1.606±0.03
mixup	Ours (Soft)	**1.18±0.02**	**8.893±0.24**	0.131±0.01	**1.48±0.03**
Mix-MaxEnt	Hard	1.07±0.04	6.436±0.59	0.065±0.01	1.34±0.03
Mix-MaxEnt	Ours (Soft)	***1.019±0.02***	**6.013±0.59**	***0.054±0.01***	***1.252±0.02***
Ours (Ent-WC)	Hard	1.054±0.01	***3.314±0.05***	**0.102±0.0**	1.346±0.03
Ours (Ent-WC)	Ours (Soft)	**1.027±0.02**	3.399±0.06	0.107±0.01	**1.30±0.02**

Table 5.2 Evaluating performance when algorithms are run instead over soft labels endpoints. Hard labels are regular `CIFAR-10` examples; soft labels are derived from a simulated set of `CIFAR-10S` labels. Bold indicates best within an algorithm and varying label base. Italices indicate best over all variants.

5.3 Takeaways

Overall, our crowdsourced study demonstrates that indeed, the structure of the synthetic labeling policy used in *mixup* does not match human perceptual knowledge, and that approaches aimed at increasing the entropy of synthetic labels implicitly push models towards this more human-aligned label space. One such method is Mix-MaxEnt (Pinto et al., 2021). We also introduce and verify two new methods of introducing softness to *mixup*: 1) training with a clamped entropy-weighted loss, and 2) mixing over soft label endpoints. While the first approach is inspired by human percepts, and has room to be grounded directly in human judgments, the latter *is* rooted in the statistics of human knowledge, thereby demonstrating the potential of leveraging a small collection of human knowledge to construct *scalable simulators*. However, as discussed, this work is in an early stage and ought to be explored more; for instance, here, we are always taking $\lambda_f = \lambda_g$. As uncovered in Chapter 4, this assumption is not human-aligned. A preliminary investigation into an additional alternative human-based *mixup* labeling is included in Appendix C.

Next We now will take a step back and consider the broader picture of the promises of eliciting and incorporating additional human knowledge: both to further improve machine learning models *and* consider improve human-machine teams. We also discuss some of the key challenges towards scaling the kinds of studies we have collected here, as well additional limitations of the work included in this thesis.

Chapter 6

Discussion

We demonstrated that eliciting and learning with additional forms of human knowledge holds great potential to improve model performance; however, our findings are nuanced – eliciting richer information from individuals may come at a cost of more annotation time, and scalability is a challenge.

Chapter Roadmap We highlight several aspects of our experimental paradigms that ought to be considered before extrapolating from our results. We discuss paths to address the thorn of scaling human knowledge elicitation. We conclude with exciting directions hinted at from this work around the utility of human uncertainty in ML systems.

6.1 Considerations

We acknowledge that the work of this dissertation has focused on a single domain – image classification – and a particular dataset in said domain, `CIFAR-10` (Krizhevsky, 2009). We have always trained models from scratch, explored only a handful of popular architectures, and relied primarily on a standard cross entropy loss (with the exception of Chapter 5). Alternative learning setups such as fine-tuning on a small amount of richly human-labeled examples, the exploration of a broader range of models like in (Peterson et al., 2019), and the consideration of other loss functions particularly those designed for noisy labels like peer loss (Peer et al., 2017) are sensible next steps to validate from trends suggested from this work (i.e., human knowledge is valuable towards improving model performance) generalize.

Additionally, the human data collected was itself of a moderate size and all participants were recruited from United States crowdsourced workers who speak English as their first language. Expanding dataset collection to other subpopulations, and exploring elicitation beyond crowdsourced workers (e.g., to domain experts such as doctors and lawyers) are prime targets for future work. Elicitation from specialist annotators offers a fertile testbed to explore whether our annotator-efficiency findings hold in practice.

6.2 Scaling

A related challenge that has permeated this work is the question of scaling human knowledge extraction. This is a particular hurdle when annotation is time-intensive (Chapter 3) or involves human judgments over synthetic examples, of which there are an infinite or near-infinite set which cannot all possible be annotated manually (Chapters 4 and 5). One approach to address scaling we consider is simulating additional human knowledge – directly from the statistics of collected annotations, and inspired by the user studies. A complementary and exciting direction to address scaling is the identification of which examples are most likely to benefit from human labeling for efficient querying (Charusaie et al., 2022; Liu et al., 2021a). We encourage more work in these directions, as well as in the design of more efficient elicitation schemes, with an eye on utility for downstream models.

6.3 Leveraging Human Uncertainty Information

While rigid conclusions cannot be drawn given the above considerations, all core chapters (2-5) of this work highlight that eliciting information about human uncertainty – which is not often collected – has tremendous potential. Individual annotators' reported personal label distribution can be used to form soft labels (P_m) which we have found provide rich, effective supervisory signals. Additionally, humans' confidence ω in their annotations (e.g., confidence in whether an inferred mixing coefficient is "correct") can also be used to shape training labels. Despite reasoning under uncertainty being a linchpin of human cognition (Lake et al., 2017) and has been shown to be a central component of "good" decision-making (Bhatt et al., 2021; Cox et al., 2021; Hall, 2002; Laidlaw and Russell, 2021; Platts-Mills et al., 2020), its elicitation and incorporation in ML datasets $\mathcal{D}$ has been underappreciated. We are excited by the prospects of this work illuminating the value such information can provide ML systems.

Additionally, as discussed in Chapter 3, richer label distributions elicited from every annotator could converge to different aggregate label distributions – which ought to be considered when forming "ground truth" labels (to which there may be done, e.g., in subjective toxicity classification).

Future directions could also explore the implications of leveraging human uncertainty information not just to improve a single model, but to boost the performance of a human-machine team (Steyvers et al., 2022; Wilder et al., 2020). For instance, in the selective prediction setting (Mozannar and Sontag, 2020), a model could be trained not just to defer to a human when the model is uncertain, but particularly allot capacity to learn to complement when the human is uncertain (Bondi et al., 2022).

Chapter 7

Conclusions

We therefore have identified gaps in the current space of information elicited from humans to be provided to ML models: namely, soft labels are rarely collected from individual annotators, and human judgments are infrequently collected over synthetically-generated data. We address these gaps with several crowdsourced studies, resulting in the creation of datasets which we release for the machine learning community. We highlight how each of these forms of **human knowledge, particularly representing annotator uncertainty, provides new information that is not often contained in the labels used to train models**. Through computational experiments, we validate the utility of leveraging these additional forms of human knowledge – from each and every annotator, and across a set of annotators – to improve performance across a suite of metrics, such as generalization, robustness, and calibration. We demonstrate that collecting and training on richer forms of human knowledge can be especially beneficial **when there are fewer human annotators available**. We then exhibit how one can simulate the generation of human judgments over more examples, thereby enabling us to form automated labeling schemes ground in and inspired by the additional kinds of human knowledge we collect to further enhance model performance.

The corresponding improvements enjoyed by models trained on our elicited human knowledge emphasize the **power and promise of expanding the kinds of information we collect from humans and incorporate into our training pipelines**. We hope this work inspires the ML community to consider what forms of human knowledge could be leveraged to design more effective supervisory signals and broadly bolster model performance. We acknowledge that there remain significant hurdles, such as annotation time, towards the broader applicability of the richer kinds of human knowledge considered here: we encourage the design of better, more efficient annotation interfaces and the exploration of alternative kinds of human information that we could collect to improve the trustworthiness and reliability of today's ML systems.

References

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. *CoRR*, abs/1908.02983, 2019. URL http://arxiv.org/abs/1908.02983.

Alexandry Augustin, Matteo Venanzi, Alex Rogers, and Nicholas R Jennings. Bayesian aggregation of categorical distributions with applications in crowdsourcing. In *IJCAI*, pages 1411–1417, 2017.

Shahar Avin, Haydn Belfield, Miles Brundage, Gretchen Krueger, Jasmine Wang, Adrian Weller, Markus Anderljung, Igor Krawczuk, David Krueger, Jonathan Lebensold, et al. Filling gaps in trustworthy development of ai. *Science*, 374(6573):1327–1329, 2021.

Varun Babbar, Umang Bhatt, and Adrian Weller. On the utility of prediction sets in human-ai teams. *arXiv preprint arXiv:2205.01411*, 2022.

Ruairidh M. Battleday, Joshua C. Peterson, and Thomas L. Griffiths. Improving machine classification using human uncertainty measurements. https://openreview.net/forum?id=rJl8BhRqF7, 2019.

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

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *CoRR*, abs/2006.07159, 2020. URL https://arxiv.org/abs/2006.07159.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413, 2021.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Elizabeth Bondi, Raphael Koster, Hannah Sheahan, Martin Chadwick, Yoram Bachrach, Taylan Cemgil, Ulrich Paquet, and Krishnamurthy Dvijotham. Role of human-ai interaction in selective prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5286–5294, 2022.

Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. Visual recognition with humans in the loop. In *European Conference on Computer Vision*, pages 438–451. Springer, 2010.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. *arXiv preprint arXiv:2001.06001*, 2020.

Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. Sample efficient learning of predictors that complement humans, 2022. URL https://arxiv.org/abs/2207.09584.

Quan Ze Chen, Daniel S Weld, and Amy X Zhang. Goldilocks: Consistent crowdsourced scalar annotations with relative uncertainty. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25, 2021.

Valerie Chen, Umang Bhatt, Hoda Heidari, Adrian Weller, and Ameet Talwalkar. Perspectives on incorporating expert feedback into model updates. *arXiv preprint arXiv:2205.06905*, 2022.

Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2020.

John Joon Young Chung, Jean Y. Song, Sindhu Kutty, Sungsoo (Ray) Hong, Juho Kim, and Walter S. Lasecki. Efficient elicitation approaches to estimate collective crowd answers. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019. doi: 10.1145/3359164. URL https://doi.org/10.1145/3359164.

Mark Collier, Rodolphe Jenatton, Effrosyni Kokiopoulou, and Jesse Berent. Transfer and marginalize: Explaining away label noise with privileged information. In *International Conference on Machine Learning*, pages 4219–4237. PMLR, 2022.

Katherine M Collins, Catherine Wong, Jiahai Feng, Megan Wei, and Joshua 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*, 2022.

Caitríona L Cox, Benjamin M Miller, Isla Kuhn, and Zoë Fritz. Diagnostic uncertainty in primary care: what is known about its communication, and what are the associated ethical issues? *Family practice*, 38(5):654–668, 2021.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

Josh de Leeuw. jspsych, 2016. URL http://docs.jspsych.org/.

Celso M. de Melo, Antonio Torralba, Leonidas Guibas, James DiCarlo, Rama Chellappa, and Jessica Hodgins. Next-generation deep learning based on simulators and synthetic data. *Trends in Cognitive Sciences*, 26(2):174–187, 2022. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2021.11.008. URL https://www.sciencedirect.com/science/article/pii/S136466132100293X.

Nathan Destler, Manish Singh, and Jacob Feldman. Shape discrimination along morphspaces. *Vision Research*, 158:189–199, 2019. ISSN 0042-6989. doi: https://doi.org/10.1016/j.visres.2019.03.002. URL https://www.sciencedirect.com/science/article/pii/S0042698919300677.

Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2342–2351, 2022.

Zeyad Emam, Andrew Kondrich, Sasha Harrison, Felix Lau, Yushi Wang, Aerin Kim, and Elliot Branson. On the state of data in computer vision: Human annotations remain indispensable for developing deep learning models. *CoRR*, abs/2108.00114, 2021. URL https://arxiv.org/abs/2108.00114.

Li Fei-Fei, Jia Deng, and Kai Li. Imagenet: Constructing a large-scale image database. *Journal of vision*, 9(8):1037–1037, 2009.

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

Vincent Fortuin. Priors in bayesian deep learning: A review. *International Statistical Review*, 2022.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521 (7553):452–459, 2015. URL http://dblp.uni-trier.de/db/journals/nature/nature521.html#Ghahramani15.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014a.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Katherine H Hall. Reviewing intuitive decision-making and uncertainty: the implications for medical education. *Medical education*, 36(3):216–224, 2002.

Stevan Harnad. Categorical perception. 2003.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16783–16792, 2022.

James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Divyansh Kaushik, Eduard H. Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *CoRR*, abs/1909.12434, 2019. URL http://arxiv.org/abs/1909.12434.

Divyansh Kaushik, Amrith Setlur, Eduard H. Hovy, and Zachary C. Lipton. Explaining the efficacy of counterfactually-augmented data. *CoRR*, abs/2010.02114, 2020. URL https://arxiv.org/abs/2010.02114.

Divyansh Kaushik, Zachary C Lipton, and Alex John London. Resolving the human subjects status of machine learning's crowdworkers. *arXiv preprint arXiv:2206.04039*, 2022.

Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. *CoRR*, abs/2009.06962, 2020a. URL https://arxiv.org/abs/2009.06962.

Jang-Hyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song. Co-mixup: Saliency guided joint mixup with supermodular diversity. *CoRR*, abs/2102.03065, 2021. URL https://arxiv.org/abs/2102.03065.

JangHyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song. Co-mixup: Saliency guided joint mixup with supermodular diversity. In *International Conference on Learning Representations*, 2020b.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

Christoph Koller, Göran Kauermann, and Xiao Xiang Zhu. Going beyond one-hot encoding in classification: Can human uncertainty improve model performance? *arXiv preprint arXiv:2205.15265*, 2022.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Cassidy Laidlaw and Stuart Russell. Uncertain decisions facilitate better preference learning. *Advances in Neural Information Processing Systems*, 34:15070–15083, 2021.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

Neil D. Lawrence. Data science and digital systems: The 3ds of machine learning systems design. *CoRR*, abs/1903.11241, 2019. URL http://arxiv.org/abs/1903.11241.

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

Christopher H. Lin, Mausam, and Daniel S. Weld. To re(label), or not to re(label). In Jeffrey P. Bigham and David C. Parkes, editors, *Proceedings of the Seconf AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2014, November 2-4, 2014, Pittsburgh, Pennsylvania, USA*. AAAI, 2014. URL http://www.aaai.org/ocs/index.php/HCOMP/HCOMP14/paper/view/8978.

Weiyang Liu, Zhen Liu, Hanchen Wang, Liam Paull, Bernhard Schölkopf, and Adrian Weller. Iterative teaching by label synthesis. In *NeurIPS*, 2021a.

Zicheng Liu, Siyuan Li, Di Wu, Zhiyuan Chen, Lirong Wu, Jianzhu Guo, and Stan Z. Li. Automix: Unveiling the power of mixup. *CoRR*, abs/2103.13027, 2021b. URL https://arxiv.org/abs/2103.13027.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2017. URL https://arxiv.org/abs/1706.06083.

Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021.

Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Edward Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. Orbit: A real-world few-shot dataset for teachable object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10818–10828, 2021.

Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Juan Ciro, Lora Aroyo, Bilge Acun, Sabri Eyuboglu, Amirata Ghorbani, Emmett Goodman, Tariq Kane, Christine R. Kirkpatrick, Tzu-Sheng Kuo, Jonas Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Ce Zhang, James Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi. Dataperf: Benchmarks for data-centric ai development, 2022. URL https://arxiv.org/abs/2207.10062.

Ana Elisa Méndez, Mark Cartwright, Juan Pablo Bello, and Oded Nov. Eliciting confidence for improving crowdsourced audio annotations. *Proc. ACM Hum.-Comput. Interact.*, 6 (CSCW1), apr 2022. doi: 10.1145/3512935. URL https://doi.org/10.1145/3512935.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Brent Miller and Mark Steyvers. The wisdom of crowds with communication. *Cognitive Science*, 33, 2011.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7076–7087. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Richard F Murray, Khushbu Patel, and Alan Yee. Posterior probability matching and human perceptual decision making. *PLoS computational biology*, 11(6):e1004342, 2015.

Vedant Nanda, Ayan Majumdar, Camila Kolling, John P. Dickerson, Krishna P. Gummadi, Bradley C. Love, and Adrian Weller. Exploring alignment of representations with human perception. *CoRR*, abs/2111.14726, 2021. URL https://arxiv.org/abs/2111.14726.

Quang Nguyen, Hamed Valizadegan, and Milos Hauskrecht. Learning classification models with soft-label information. *Journal of the American Medical Informatics Association : JAMIA*, 21, 11 2013. doi: 10.1136/amiajnl-2013-001964.

Quang Nguyen, Hamed Valizadegan, and Milos Hauskrecht. Learning classification models with soft-label information. *Journal of the American Medical Informatics Association*, 21 (3):501–508, 2014.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

Jeremy E Oakley and Anthony O'Hagan. Shelf: the sheffield elicitation framework (version 2.0). *School of Mathematics and Statistics, University of Sheffield, UK (http://tonyohagan. co. uk/shelf)*, 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. URL http://oro.open.ac.uk/17948/.

Anthony O'Hagan. Expert knowledge elicitation: Subjective but scientific. *The American Statistician*, 73(sup1):69–81, 2019. doi: 10.1080/00031305.2018.1518265. URL https://doi.org/10.1080/00031305.2018.1518265.

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

Rebecca J Passonneau and Bob Carpenter. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326, 2014.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34:20596–20607, 2021.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *CoRR*, abs/2012.05345, 2020. URL https://arxiv.org/abs/2012.05345.

Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626, 2019.

Francesco Pinto, Harry Yang, Ser-Nam Lim, Philip HS Torr, and Puneet K Dokania. Mixmaxent: improving accuracy and uncertainty estimates of deterministic neural networks. 2021.

Emmanouil Antonios Platanios, Maruan Al-Shedivat, Eric P. Xing, and Tom M. Mitchell. Learning from imperfect annotations. *CoRR*, abs/2004.03473, 2020. URL https://arxiv.org/abs/2004.03473.

Timothy F Platts-Mills, Justine M Nagurney, and Edward R Melnick. Tolerance of uncertainty and the practice of emergency medicine. *Annals of emergency medicine*, 75(6):715–720, 2020.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*, 2021.

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

Vikram V Ramaswamy, Sunnie SY Kim, Ruth Fong, and Olga Russakovsky. Overlooked factors in concept-based explanations: Dataset choice, concept salience, and human capability. *arXiv e-prints*, pages arXiv–2207, 2022.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *CoRR*, abs/1902.10811, 2019. URL http://arxiv.org/abs/1902.10811.

Lars Schmarje, Vasco Grossmann, Claudius Zelenka, Sabine Dippel, Rainer Kiko, Mariusz Oszust, Matti Pastell, Jenny Stracke, Anna Valros, Nina Volkmann, et al. Is one annotation enough? a data-centric image classification benchmark for noisy and ambiguous label estimation. *arXiv preprint arXiv:2207.06214*, 2022.

Viktoriia Sharmanska, Daniel Hernandez-Lobato, Jose Miguel Hernandez-Lobato, and Novi Quadrianto. Ambiguity helps: Classification with disagreements in crowdsourced annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

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

Yonadav Shavit, Divyansh Kaushik, Zachary C Lipton, Samuel R. Bowman, and Kira Goldner. Request for information (rfi) on implementing the initial findings and recommendations of the national artificial intelligence research resource task force: Response. *Federal Register Notice 87 FR 31914*, 2022.

Victor S Sheng, Jing Zhang, Bin Gu, and Xindong Wu. Majority voting and pairing with multiple noisy labeling. *IEEE Transactions on Knowledge and Data Engineering*, 31(7): 1355–1368, 2017.

Ayame Shimizu and Kei Wakabayashi. Examining effect of label redundancy for machine learning using crowdsourcing. In *The 23rd International Conference on Information Integration and Web Intelligence*, iiWAS2021, page 87–94, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450395564. doi: 10.1145/3487664.3487677. URL https://doi.org/10.1145/3487664.3487677.

Ben Shneiderman. . In *Human-Centered AI*. Oxford University Press, 01 2022. ISBN 9780192845290. doi: 10.1093/oso/9780192845290.001.0001.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587): 484–489, jan 2016. ISSN 0028-0836. doi: 10.1038/nature16961.

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

Padhraic Smyth, Michael C Burl, Usama M Fayyad, and Pietro Perona. Knowledge discovery in large image databases: Dealing with uncertainties in ground truth. In *KDD workshop*, pages 109–120, 1994.

Jy-yong Sohn, Liang Shang, Hongxu Chen, Jaekyun Moon, Dimitris S. Papailiopoulos, and Kangwook Lee. Genlabel: Mixup relabeling using generative models. *CoRR*, abs/2201.02354, 2022. URL https://arxiv.org/abs/2201.02354.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 596–608. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/06964dce9addb1c5cb5d6e3d9838f733-Paper.pdf.

Jinhua Song, Hao Wang, Yang Gao, and Bo An. Active learning with confidence-based answers for crowdsourcing labeling tasks. *Knowledge-Based Systems*, 159:244–258, 2018.

Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11): e2111547119, 2022.

Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. Provably improving expert predictions with conformal prediction. *arXiv preprint arXiv:2201.12006*, 2022.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. URL http://arxiv.org/abs/1512.00567.

Shakila Thangaratinam and Charles WE Redman. The delphi technique. *The obstetrician & gynaecologist*, 7(2):120–125, 2005.

Rachel L Thomas and David Uminsky. Reliance on metrics is a fundamental challenge for ai. *Patterns*, 3(5):100476, 2022.

Sunil Thulasidasan, Gopinath Chennupati, Jeff Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks, 2019. URL https://arxiv.org/abs/1905.11001.

Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Between-class learning for image classification. *CoRR*, abs/1711.10284, 2017a. URL http://arxiv.org/abs/1711.10284.

Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Learning from between-class examples for deep sound recognition. *CoRR*, abs/1711.10282, 2017b. URL http://arxiv.org/abs/1711.10282.

Dustin Tran, Jeremiah Liu, Michael W. Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, Neil Band, Tim G. J. Rudner, Karan Singhal, Zachary Nado, Joost van Amersfoort, Andreas Kirsch, Rodolphe Jenatton, Nithum Thain, Honglin Yuan, Kelly Buchanan, Kevin Murphy, D. Sculley, Yarin Gal, Zoubin Ghahramani, Jasper Snoek, and Balaji Lakshminarayanan. Plex: Towards reliability using pretrained large model extensions, 2022. URL https://arxiv.org/abs/2207.07411.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Alan Turing. Computing machinery and intelligence. *Mind*, 59(236):433, 1950.

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

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. A case for soft loss functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):173–177, Oct. 2020. URL https://ojs.aaai.org/index.php/HCOMP/article/view/7478.

Alexandra Uma, Dina Almanea, and Massimo Poesio. Scaling and disagreements: Bias, noise, and ambiguity. *Frontiers in Artificial Intelligence*, 5, 2022.

Vladimir Vapnik, Rauf Izmailov, et al. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 16(1):2023–2049, 2015.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, Aaron Courville, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states, 2018. URL https://arxiv.org/abs/1806.05236.

Dongdong Wang, Yandong Li, Liqiang Wang, and Boqing Gong. Neural networks are more productive teachers than human raters: Active mixup for data-efficient knowledge distillation from a blackbox model. *CoRR*, abs/2003.13960, 2020. URL https://arxiv.org/abs/2003.13960.

Jerry Wei, Lorenzo Torresani, Jason Wei, and Saeed Hassanpour. Calibrating histopathology image classifiers using label smoothing, 2022a. URL https://arxiv.org/abs/2201.11866.

Jiaheng Wei, Zhaowei Zhu, Tianyi Luo, Ehsan Amid, Abhishek Kumar, and Yang Liu. To aggregate or not? learning with separate noisy labels. *arXiv preprint arXiv:2206.07181*, 2022b.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22, 2009.

Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans, 2020.

Catherine Wong, Kevin M Ellis, Joshua Tenenbaum, and Jacob Andreas. Leveraging language to learn program abstractions and search heuristics. In *International Conference on Machine Learning*, pages 11193–11204. PMLR, 2021.

Guodong Xu, Ziwei Liu, and Chen Change Loy. Computation-efficient knowledge distillation via uncertainty-aware mixup. *arXiv preprint arXiv:2012.09413*, 2020.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *CoRR*, abs/1905.04899, 2019. URL http://arxiv.org/abs/1905.04899.

John Zerilli, Umang Bhatt, and Adrian Weller. How transparency modulates trust in artificial intelligence. *Patterns*, 3(4):100455, 2022. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2022.100455. URL https://www.sciencedirect.com/science/article/pii/S2666389922000289.

Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. Delving deep into label smoothing. *IEEE Transactions on Image Processing*, 30:5984–5996, 2021a.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412, 2017. URL http://arxiv.org/abs/1710.09412.

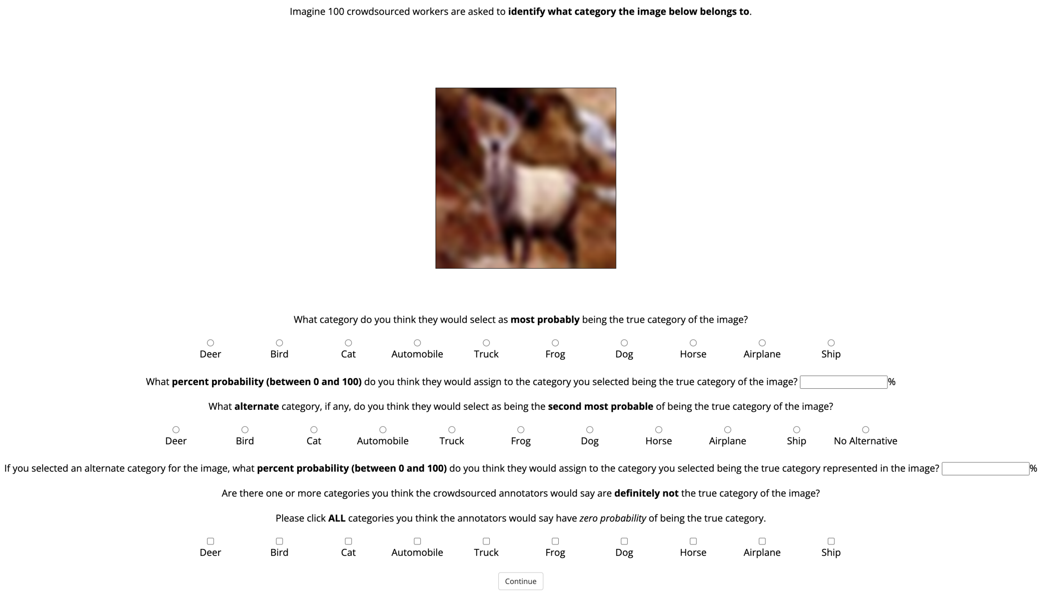
Jing Zhang and Xindong Wu. Multi-label inference for crowdsourcing. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2738–2747, 2018.

Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Y. Zou. How does mixup help with robustness and generalization? *CoRR*, abs/2010.04819, 2020. URL https://arxiv.org/abs/2010.04819.

Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10145–10155, June 2021b.

Appendix A

Human Subject Experiment Interfaces



Imagine 100 crowdsourced workers are asked to **identify what category the image below belongs to**.

What category do you think they would select as **most probably** being the true category of the image?

○ Deer ○ Bird ○ Cat ○ Automobile ○ Truck ○ Frog ○ Dog ○ Horse ○ Airplane ○ Ship

What **percent probability (between 0 and 100)** do you think they would assign to the category you selected being the true category of the image? ______ %

What **alternate** category, if any, do you think they would select as being the **second most probable** of being the true category of the image?

○ Deer ○ Bird ○ Cat ○ Automobile ○ Truck ○ Frog ○ Dog ○ Horse ○ Airplane ○ Ship ○ No Alternative

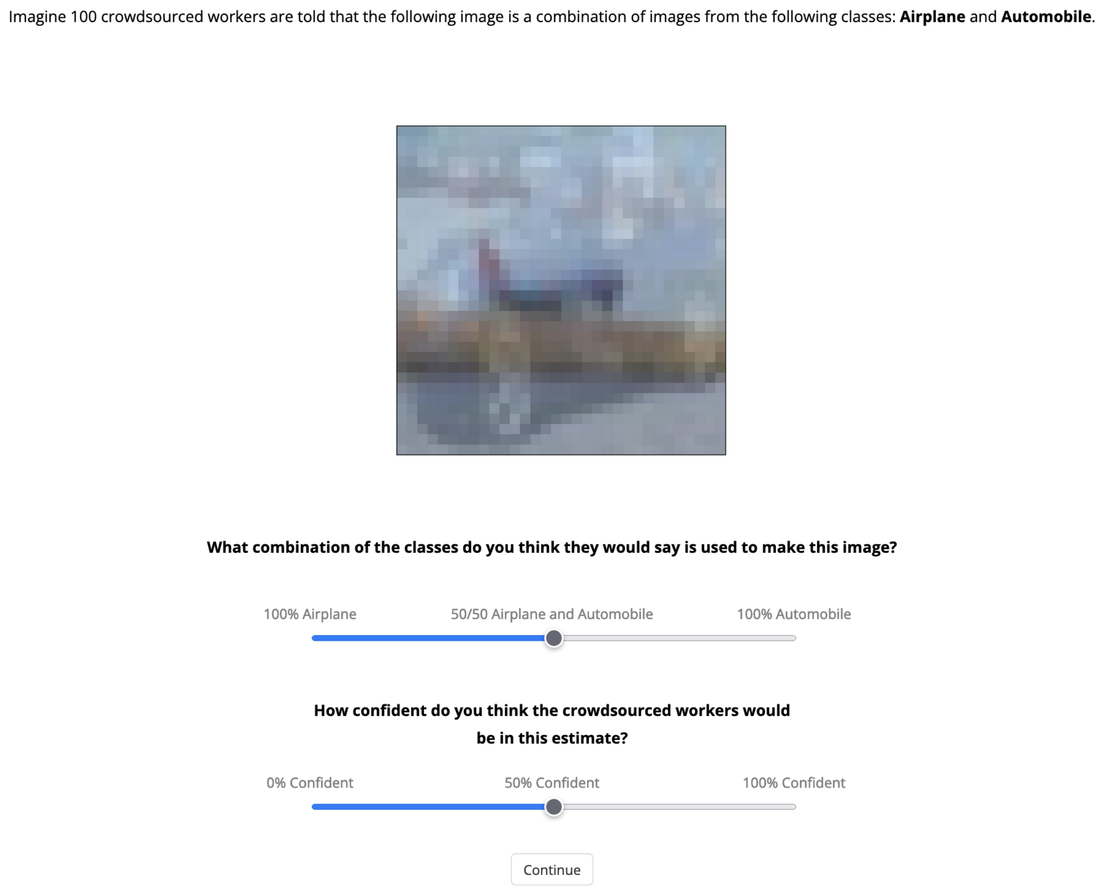
If you selected an alternate category for the image, what **percent probability (between 0 and 100)** do you think they would assign to the category you selected being the true category represented in the image? ______ %

Are there one or more categories you think the crowdsourced annotators would say are **definitely not** the true category of the image?

Please click **ALL** categories you think the annotators would say have *zero probability* of being the true category.

□ Deer □ Bird □ Cat □ Automobile □ Truck □ Frog □ Dog □ Horse □ Airplane □ Ship

Fig. A.1 Depiction of our soft label elicitation interface for classical observational examples.

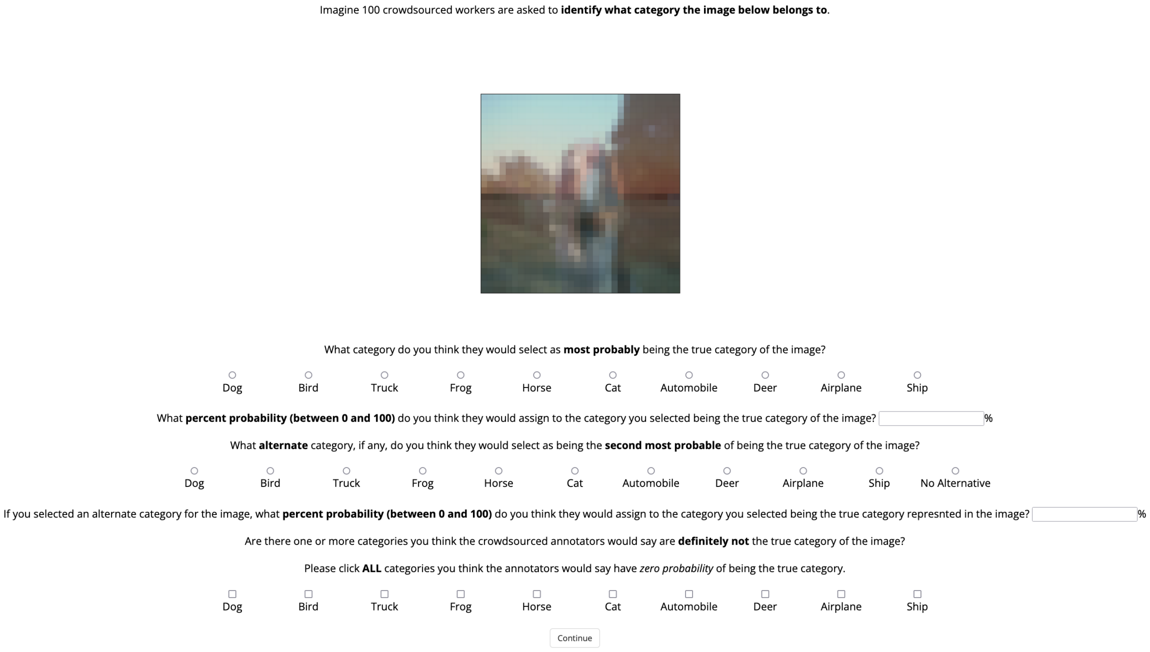


Imagine 100 crowdsourced workers are told that the following image is a combination of images from the following classes: **Airplane** and **Automobile**.

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?

Fig. A.2 Example relabeling elicitation task shown to each crowdsourced worker.



Imagine 100 crowdsourced workers are asked to **identify what category the image below belongs to**.

What category do you think they would select as **most probably** being the true category of the image?

○ Dog ○ Bird ○ Truck ○ Frog ○ Horse ○ Cat ○ Automobile ○ Deer ○ Airplane ○ Ship

What **percent probability (between 0 and 100)** do you think they would assign to the category you selected being the true category of the image? %

What **alternate** category, if any, do you think they would select as being the **second most probable** of being the true category of the image?

○ Dog ○ Bird ○ Truck ○ Frog ○ Horse ○ Cat ○ Automobile ○ Deer ○ Airplane ○ Ship ○ No Alternative

If you selected an alternate category for the image, what **percent probability (between 0 and 100)** do you think they would assign to the category you selected being the true category represnted in the image? %

Are there one or more categories you think the crowdsourced annotators would say are **definitely not** the true category of the image?

Please click **ALL** categories you think the annotators would say have *zero probability* of being the true category.

□ Dog □ Bird □ Truck □ Frog □ Horse □ Cat □ Automobile □ Deer □ Airplane □ Ship

Fig. A.3 Example synthetic soft label elicitation interface.

Appendix B

Additional CIFAR-10S Investigation

We include additional experiments, analyses, and investigations into the value of collecting soft labels from individual humans.

B.1 Additional Examples

Further comparison between CIFAR-10H and our CIFAR-10S are shown in Figs. B.1 and B.2.

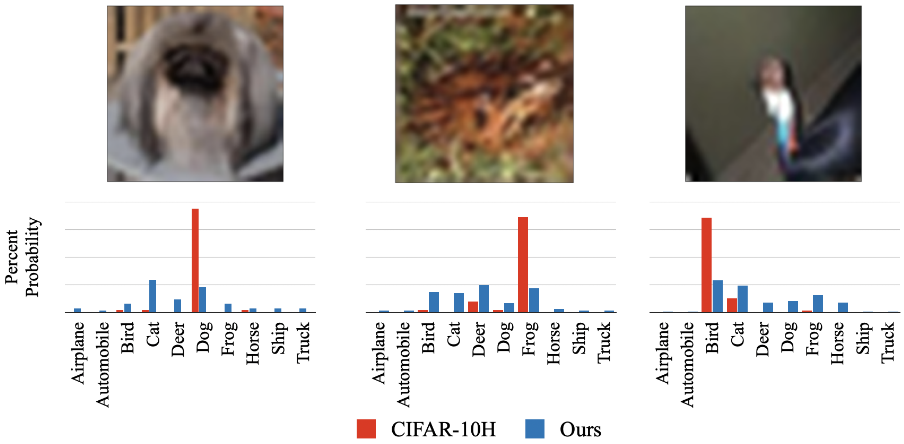


Fig. B.1 Top three highest Wasserstein distance examples between our CIFAR-10S labels (blue) and CIFAR-10H (red). The hard labels in CIFAR-10 are: dog, frog, and bird.

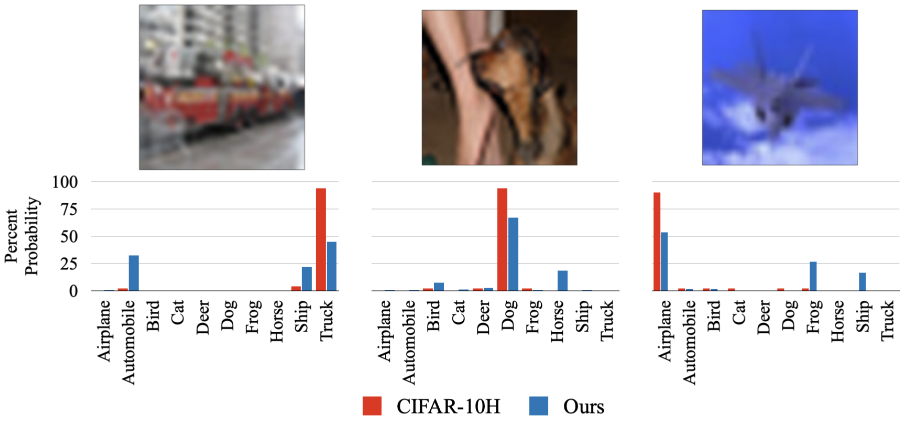


Fig. B.2 Additional examples demonstrating the "softening" of CIFAR-10H labels with our elicitation. The hard labels in CIFAR-10 are: truck, dog, and airplane.

B.2 Comparison to Classical Label Smoothing

When considering the value of eliciting and incorporating additional human knowledge in ML systems, one may ask why not just use traditional label smoothing (LS)?

$$y_n^{\text{LS}} = y_n^{\text{hard}} * (1-\beta) + \beta * y^{\text{smoother}}$$

y^{smoother} is a uniform of length K (each "probability" $= \frac{1}{K}$ (applied to all N examples), y_n^{hard} is the conventional one-hot label, and α is the smoothing factor $\in [0,1]$.

We train the three models considered in Section 3.3 over labels on which LS ($\beta = 0.05$) was applied. We tune β ($\in \{0, 0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4\}$ with the same validation method discussed (Section 3.3.1). We find in Table B.1 that LS does outperform both CIFAR-10H and CIFAR-10S in terms of calibration and roubstness over CIFAR-10H; however, this does not hold when evaluated on CIFAR-10S. While these results do shed light on a nuance of human knowledge elicitation (alternative, automated ML approaches are powerful already); we argue that human knowledge still contributes valuable insights. Importantly, LS does not capture *meaningful softness*, i.e., an image most likely to be a deer has equal probability of alternatively being a dog as a ship. This lack of human-sensible alternatives may prohibit effective generalization to our richer, harder CIFAR-10S evaluation set; a result of which has been shown on other datasets Zhang et al. (2021a). Furthermore, the blanket mass spread over all $K-1$ alternatives also has been found to lead to oversimplified clusters Müller et al. (2019). It is worth investigating the kinds of latent spaces that result from training over our human-derived soft labels instead.

	Label Type	CE	Calibration	FGSM Loss
10H	Label Smoothing	1.368±0.19	**0.175±0.05**	**6.965±1.7**
	CIFAR-10H	1.293±0.08	0.194±0.01	8.577±1.91
	Ours (T2, Clamp)	**1.281±0.06**	0.184±0.01	8.406±1.75
10S	Label Smoothing	2.674±0.33	0.299±0.05	9.375±3.4
	CIFAR-10H	2.459±0.21	0.311±0.02	**8.334±1.75**
	Ours (T2, Clamp)	**2.355±0.14**	**0.297±0.03**	8.405±1.59

Table B.1 Comparing de-aggregated human-derived soft labels against label smoothing.

B.3 Full Annotator Efficiency

We study the performance of training with each M, run for 5 seeds.

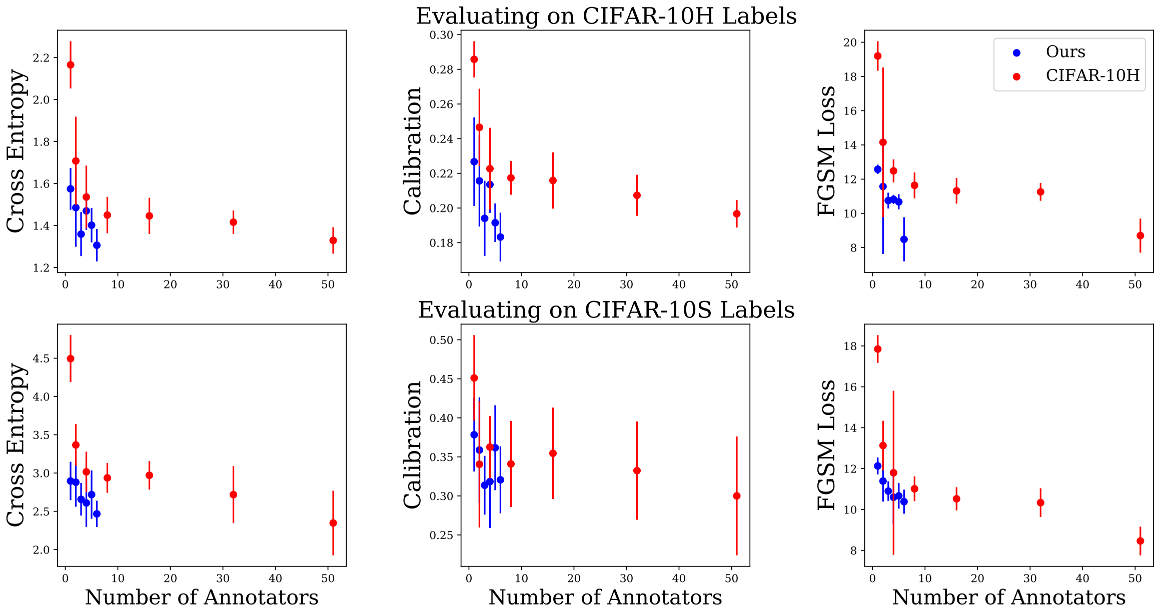


Fig. B.3 Comparison of learner performance based on number of annotators used to create the training labels. Same depiction as Fig 3.5 with only M.

B.4 Alternative Semantic-Based Redistribution

In Section 3., we introduced redistribution of probability mass when we have partial probability distributions from humans. As discussed in Section **??**, uniform redistribution does not capture meaningful alternative categories which humans do seem to perceive (see 3.2.3 and

B.1). As such, we consider redistributing mass based on the semantic distance of alternative classes to that selected in Top 1. We approximate semantic distance with cosine similarity of GloVe embeddings (Pennington et al., 2014) converted to Word2Vec formatMikolov et al. (2013). We find mixed results for whether the redistribution, in this setting, improves over Uniform. It appears that semantic redistribution is more effective when we use it in concert with our T2 setting (shown in Table B.3) compared to just using the T1 information (Table B.2). Study into why this may be, and whether other experimental paradigms reveal the benefits of semantic-based smoothing are warranted.

	Redist	CE	Calibration	FGSM Loss
10H	Uniform	**1.423 +/- 0.08**	**0.198 +/- 0.01**	10.891 +/- 0.32
	Semantic	1.439 +/- 0.05	**0.198 +/- 0.02**	**10.369 +/- 1.23**
10S	Uniform	**2.531 +/- 0.15**	**0.312 +/- 0.03**	11.25 +/- 0.4
	Semantic	2.568 +/- 0.97	0.32 +/- 0.17	**10.284 +/- 1.35**

Table B.2 Uniform vs. semantic redistribution methods to construct labels from the T1 variety. The redistribution method used to spread leftover mass. 5 seeds are run for ResNet-34A.

	Redist	CE	Calibration	FGSM Loss
10H	Uniform	1.448 +/- 0.14	0.203 +/- 0.02	10.474 +/- 0.44
	Semantic	**1.382 +/- 0.19**	**0.195 +/- 0.04**	**10.378 +/- 1.11**
10S	Uniform	**2.63 +/- 0.29**	0.339 +/- 0.1	10.667 +/- 1.15
	Semantic	2.64 +/- 0.73	**0.335 +/- 0.18**	**10.326 +/- 1.72**

Table B.3 Uniform vs. semantic redistribution methods to construct labels using both Top 1 and Top 2 elicited information (T2). Setup follows B.2.

B.5 Simulating Label Construction

We hand-craft a generative model based on our Top-2 Clamp label formulation to convert hard labels into soft labels which approximate those we elicit from individual humans. The model is as follows:

1. We are provided a hard label. We let this be the selected Top 1.
2. Sample from a Beta distribution the probability an annotator may have assigned to that hard label. Where Beta is fit to the Top-1 probabilities assigned by our humans.

3. Sort all alternative classes based on their semantic distance (using the definition in Section B.3).

4. Sample a number of alternative labels K_p which would have been deemed possible based on a histogram of the number of labels not selected as "impossible." Keep only K_p of the ranked list.

5. Sample from another Beta distribution fit to human-provided Top-2 probabilities. Spread this amount of mass on the closest alternative class, if $K_p \geq 1$.

6. Spread any remaining mass uniformly over any other possible categories, mimicking our Clamp.

Note, we fit a different Beta and use a separate "possible-class" histogram based on whether the example was high or low entropy (defined against CIFAR-10H as discussed in Section 3.2.1). We route to the proper distribution based on CIFAR-10H entropy of the example being converted, and sample the hard label from the CIFAR-10H label. This means that our approach is somewhat constrained by the data collected by CIFAR-10H. As such, we emphasize this is a *first pass* at leveraging grounding more (but not fully) scalable human knowledge simulators. Future work could explore alternative ways to estimate image entropy, such as predicted model entropy, when deciding which distribution to use in generating a label, as well as full learned generative models of such soft labels.

Appendix C

Scaling with Human-Aligned Logistic Functions

We explore the value of leveraging the human knowledge that we have collected to define automated, scalable *mixup* policies. We choose a logistic function to parameterize λ_g based on Fig. 4.4 and cognitive neuroscience literature (Destler et al., 2019; Folstein et al., 2013; Goldstone and Hendrickson, 2010; Harnad, 2003).

We fit a logistic function $\tau_{i,j} \in [0,1]$, using the `scipy.curve_fit` function, based on the human data we have collected for every class pair (y_i, y_j) in `CIFAR-10`. At each batch, and for each pair, then we let $\lambda_g = \tau_{i,j}(\lambda_f, y_i, y_j)$. However, the utility of this approach is not yet clear as: 1) though combining with the entropy-weighted loss of Chapter 5 achieves the best CE, without said loss, the method is worse than linear *mixup* alone, 2) comparison against a baseline against non-human-fitted logistic is necessary, and 3) this set-up assumes the endpoints are hard so that we can index to the right logistic; as seen in Section 5.2.3, mixing softer endpoints may be beneficial. Investigative work into these points is ongoing.

Algorithm	Ent-WC	CE	FGSM Loss	Calibration
mixup	No	1.252±0.02	10.547±0.26	**0.087±0.01**
τ mixing	No	1.28 +/- 0.04	6.502 +/- 1.61	0.172 +/- 0.11
mixup	Yes	1.054±0.01	**3.314±0.05**	0.102±0.0
τ mixing	Yes	**1.039 +/- 0.02**	4.171 +/- 0.06	0.109 +/- 0.01

Table C.1 Mixing with transformed λ_g using functions fit to human data (Ours, Class-Pair Transform; i.e. τ). Varying whether entropy-weighting (Ent-WC) is applied to loss. Same setting as 5.1.