Toward a Perspectivist Turn in Ground Truthing for Predictive Computing

Federico Cabitza[1, 2], **Andrea Campagner**[2], **Valerio Basile**[3]

[1] Department of Informatics, Systems and Communication, University of Milano-Bicocca, v.le Sarca 336 – 20126 Milan, Italy
[2] IRCCS Istituto Ortopedico Galeazzi, Milan, Italy
[3] University of Turin, C.so Svizzera 185 – 10149 Turin, Italy
federico.cabitza@unimib.it, a.campagner@campus.unimib.it, valerio.basile@unito.it

Abstract

Most current Artificial Intelligence applications are based on supervised Machine Learning (ML), which ultimately grounds on data annotated by small teams of experts or large ensemble of volunteers. The annotation process is often performed in terms of a majority vote, however this has been proved to be often problematic by recent evaluation studies. In this article, we describe and advocate for a different paradigm, which we call perspectivism: this counters the removal of disagreement and, consequently, the assumption of correctness of traditionally aggregated gold-standard datasets, and proposes the adoption of methods that preserve divergence of opinions and integrate multiple perspectives in the ground truthing process of ML development. Drawing on previous works which inspired it, mainly from the crowdsourcing and multi-rater labeling settings, we survey the state-of-the-art and describe the potential of our proposal for not only the more subjective tasks (e.g. those related to human language) but also those tasks commonly understood as objective (e.g. medical decision making). We present the main benefits of adopting a perspectivist stance in ML, as well as possible disadvantages, and various ways in which such a stance can be implemented in practice. Finally, we share a set of recommendations and outline a research agenda to advance the perspectivist stance in ML.

Motivations and Background

Data annotation is the practice of (manually or automatically) labelling a set of digital representations of objects. The common pipeline for this kind of data work includes (Muller et al. 2021): *data collection*; the top-down definition of the pertinent classification schema, i.e., the eligible labels by which to annotate the collected data; the manual *data annotation*[1] by some domain experts or larger groups of volunteers (as in case of crowdsourced annotation); and, crucially to our aims, *label aggregation*, that is producing one (or, in the case of multi-label learning, a set of) representative label for each object out of the multiple ones given by the annotators. This latter step is critical because disagreement among the annotators is anything but rare, especially in case of ambiguous phenomena to classify, such as texts, social media content (Chandrasekharan et al. 2017), or medical cases (Cabitza et al. 2019).

How to deal with disagreement in collaborative data annotation is a topic that has attracted some interest in the ML community, especially in case of crowdsourcing, e.g. (Sheng and Zhang 2019); that notwithstanding, the most common approach is one: to get rid of disagreement. This can be accomplished in a number of ways: better training of annotators, so as to improve their alignment and compliance with the classification criteria (Gadiraju, Fetahu, and Kawase 2015); conflict reconciliation and adjudication after collective discussion, followed by "some kind of consensus voting" (Muller et al. 2021); or simply by some variation of *majority voting*, even without the direct involvement of the annotators. This latter post-hoc technique, as drastic and trivial as it might look (or perhaps because of that and because it is the fastest and cheapest method), is also the most frequently performed in the process of ML *ground truthing*, that is the construction of the reference truth to be "learned" by the predictive system.

The main idea behind disagreement removal grounds on the ideal of truth for which a "higher-quality ground truth is one in which multiple humans provide the same annotation for the same examples" (Aroyo and Welty 2015), and on the corresponding idea that different (and contrasting) labels for the same cases indicate that some errors occurred, committed by the raters who do not agree with the majority of coworkers: non-perfect performance necessarily introduces a sort of bias – usually called *label noise* (Kahneman et al. 2016), *label bias* (Freeman et al. 2021) or *annotation bias* (Hube, Fetahu, and Gadiraju 2019). This "noise" affecting annotations could be possibly due to cognitive biases (Draws et al. 2021); or to differences in the raters' background or experience (Gurari and Grauman 2017); or to the quality of task instructions and how these latter are followed (Kairam and Heer 2016). The idea of disagreement as a sign of bias has inspired specialized academic initiatives (Draws et al. 2021) and various bias mitigation and removal strategies (Sheng and Zhang 2019), whose common element is trust in the tenet of "collective intelligence": the judgment of many is usually more correct than that of individuals.

However, the above idea of noise could be wrong and the common wariness of disagreement be based on the two fal-

[1] A note on terminology: in this paper we speak about *objects* to be classified according to a *phenomenon* by *raters*.

lacies pointed out by (Aroyo and Welty 2015): first of all, that there is always only "one correct interpretation for every input example"; and second, that "disagreement is bad" and should be always avoided or reduced.

In fact, real-world settings show that disagreement is unavoidable and essentially irreducible, especially when the objects to classify are so complex that most of the raters can actually get them wrong, and the real experts are a minority (Basile 2021; Cabitza et al. 2019); or when the objects are so ambiguous, as it often happens in Natural Language Processing (NLP) (Artstein and Poesio 2008), emotion recognition (Cabitza, Campagner, and Mattioli 2022) or Computer Vision (Yun et al. 2021), that disagreement between annotators may embed valuable nuances challenging the very idea of clear-cut classification (Aroyo and Welty 2015). Moreover, the ambiguity and complexity of objects and cases to be interpreted can lead to high disagreement among raters not only in the notoriously subjective domains mentioned above, but also in seemingly objective disciplines like medicine or engineering: for instance, as considered in (Chernova and Veloso 2010), training a self-driving vehicle may involve states in which multiple actions are perfectly reasonable; Schaekermann et al. (2019) reported a disagreement rate of over 50% in the identification of Parkinson, which could not be completely eliminated even after Delphi-like group deliberation; similarly, Cabitza et al. (2019) reported poor agreement between clinicians even in merely descriptive tasks, when they were called to describe electrocardiograms they had just read or surgical operations they had attended in presence. This entails concerns regarding the potential lack of fairness or representativity (Balayn, Bozzon, and Szlavik 2019) of the majority judgment, or about the embedding of biases in the ML models developed on these data.

As discussed in a recent survey (Paullada et al. 2021), the "data" aspect, and the previously described associated issues, has always been a critical aspect of the ML development but it remains overlooked, both extensively mishandled in practice and ignored in theory. To address this gap, a number of similar initiatives have recently focused, and invited scholars to focus, on the data annotation processes: the Data Nutrition Project (Holland et al. 2018); the Model Cards (Mitchell et al. 2019); the Data Statements (Bender and Friedman 2018); or the Datasheets for Datasets (Gebru et al. 2021) initiative, followed by IBM with their AI FactSheets (Richards et al. 2021) and Andrew Ng with his recent proposal of *Data-Centric AI*.

These initiatives are all aimed at raising awareness for the need of greater attention and transparency in regard to the data production process and the annotation tasks, including the need to document in which (technical, social, economical and political) context the data were collected and how annotation was actually performed. However, they are limited to this. To complement but also exceed these approaches, we propose and argue for a paradigm shift: moving away from monolithic, majority-aggregated gold standard datasets, and adopting methods that more comprehensively and inclusively integrate the opinions and perspectives of the human subjects involved in the knowledge representation step of ground truthing.

As we will see in what follows, our proposal comes with important and still-to-investigate implications: first, supervised models equipped with full, non-aggregated annotations have been reported to exhibit a better prediction capability (Akhtar, Basile, and Patti 2020), in virtue of a better representation of the phenomena of interest; secondly, new techniques for AI explainability can be devised that describe the classifications of the model in terms of multiple and alternative (if not complementary) perspectives (Noble 2012); finally, we should consider the ethical implications of the above mentioned shift and its impact on cognitive computing, whereas the new generation of models can give voice to, and express, a diversity of perspectives, rather than being a mere reflection of the majority (Noble 2012).

In short, our main contributions are as follow[2]: we review existing state-of-the-art approaches in crowdsourcing, multi-rater labeling and similar settings in ML; we discuss a novel conceptual framework (*perspectivism*) which unifies and extends previous work and outline our proposal for a novel *perspectivist* research program in ML that goes beyond just learning from disaggregated labels, e.g. by including perspectivism in the evaluation step; we describe future directions to exploit the richness of multi-rater labelling and turning disagreement into a valuable resource for more comprehensive and representative ground truthing in supervised ML.

Strong and Weak Perspectivism

As anticipated above, in this paper we propose what we denote as a *perspectivist* approach to *ground truthing* and *learning*, that is in producing ground truths and learning from them models for supervised classification tasks based on Machine Learning (ML) methods and techniques. This general stance can be articulated in two main versions, which could be connoted as either a *weak* or *strong* approach (see Figure 1).

A *weak perspectivist* approach is adopted when researchers involved in ground truthing are not content to collect a single label for each object to be classified, that is to produce a *gold standard* label set; but rather aim to collect as many raters *and* annotations as possible, i.e., to build what in (Campagner et al. 2021) has been called *diamond standard* (see Figure 1). We emphasize the distinction between raters and annotations because, as simply as it can be, raters could express more than one label for a given object to classify (Dumitrache, Aroyo, and Welty 2018) (also as a way to express their indecision in case of strictly alternative labels), or they could expressly be asked to rank available labels in terms of pertinence or to associate each class with a confidence/probability level.

One could rightly wonder why ML researchers would want to collect such redundant information about the phenomena for which they design and develop decision support systems which are usually aimed at improving human decision making by proposing the one best label for each object to classify (Aroyo and Welty 2015). For now, we are not

[2]For an extended version, see https://arxiv.org/abs/2109.04270.

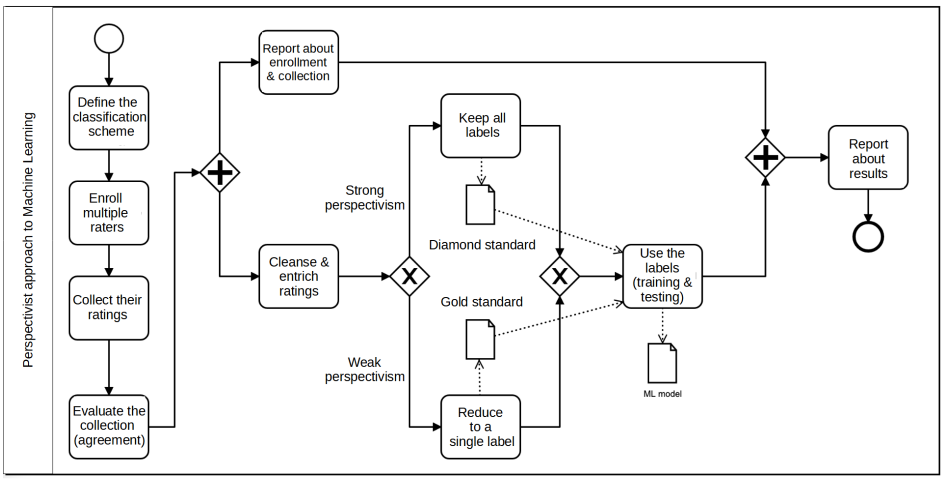

Figure 1: A BPMN (Business Process Model and Notation) diagram of the ground truthing process in a perspectivist setting. Many tasks are common with more traditional ML pipelines, but the core distinction with these latter ones lies in the exclusive gateway (X) in the centre of the diagram. Parallel gateways (+) indicate opportunities for parallel activities, which we made explicit to emphasize the importance of the comprehensive reporting of the ground truthing process.

going to dispute the assumption that conceiving the output of such systems as single labels is the best way to improve human decision making[3]; even when the output must be a single label, ML researchers could want to collect multiple labels for single objects to classify when they deem relying on single judgments either too *limiting* (like in multi-label tasks, where labelling categories are not disjoint and they overlap to some extent) or too *error-prone* (Vaughan 2017).

The above mentioned situations could arise for two main reasons: 1) distrust in the raters involved, as it can be the case in crowdsourcing initiatives or questionnaire-driven studies where researchers cannot oversee the annotation task or ensure its accuracy (Vaughan 2017); 2) the recognized variability of the phenomenon of interest (Aroyo and Welty 2015), that is the recognition that different raters could classify the same object differently, and not necessarily because they are wrong.

Indeed, and most relevantly, ratings can differ not only because the raters are fallible, but also for a number of other factors, among which we recall: the intrinsic ambiguity and complexity of the phenomenon, including the so-called *cumulative mess*, that is the condition in which the same object can legitimately be classified as many things at the same time (Bechmann and Bowker 2019); the (in)stability of the phenomenon over time; the complexity of the task, also in terms of number of distinct states or configurations of the phenomenon to detect and of the focus and attention that is requested to the raters, and their necessary proficiency to detect and understand the phenomenon; the raters' susceptibility to somehow systematic cognitive biases, both at individual and team level, like overconfidence, confirmation and availability bias, anchoring and halo effects (Eickhoff 2018), as well as to more contingent and context-driven external factors (what Kahneman et al. (2016) have denoted as *chance variability of judgments*, or "decision noise").

Thus, collecting multiple labels from a sample of individuals can help ML researcher get a sample of perceptions, opinions and judgments that could be maximally *representative* of the population of interest (in terms of both the annotators *and* the annotated objects). We note that representativeness cannot be given for granted with case-wise majority voting. For instance, in a crowdsourcing study (Cabitza, Campagner, and Mattioli 2022) in the emotion recognition field, we observed thatthe involved raters agreed with the majority judgement only less than once in two times (average alpha: .44) on average. A similar agreement was observed also in a radiological study (Cabitza et al. 2020), where the authors involved 13 experts to diagnose 427 images: there the average agreement was higher (alpha: .76) but no radiologist agreed with the majority decision more than 89% of cases. In both these cases (not at all extraordinary), majority ground truth poorly represented everyone involved in ground truthing, or it properly expressed the judgments of only very few raters.

Weak perspectivism would then advocate the production of gold standards by considering and taking into account multiple perspectives, ideally the judgments of all, equally. This does not necessarily require to make each perspective symbolically explicit, in terms of multiple labels: a meeting where multiple experts are invited to share their opinion about an object would be perspectivist as long as all people involved could express their opinions and views in a discussion, which could then be summarized into single positions. Likewise, a weak perspectivist approach would require to

[3] As a matter of fact, alternative ways, like in conformal prediction, could improve it more, especially if we consider improvement also beyond the mere dimension of error rate

collect multiple labels from a corresponding number of observers or raters but then would combine these labels and select one single label for each object to be annotated, mostly by some kind of majority voting (e.g., weighted voting). This process has been called *perspective reduction* in (Campagner et al. 2021) and entails the observation that collecting multiple labels could also allow to draw the right labels on the basis of *qualified* majorities for the sake of higher accuracy[4].

On the other hand, we would speak of *strong perspectivism* whenever the researchers' aim is to collect multiple labels, or multiple data about each class, about a specific object, and *keep them* all in the subsequent phases of training or benchmarking of the classification models. Doing so certainly impacts model training and evaluation, but can be realized in several ways, of varying complexity (Campagner et al. 2021; Sudre et al. 2019; Uma et al. 2020). The easiest way, that is the most backward-compatible way that does not require ad-hoc implementations, is to apply data augmentation schemes (Nishi et al. 2021), such as replicating each object in the training set to reflect the number of times this object has been associated with a certain label by the raters; nonetheless, also other methods have been proposed in the literature, that we will better describe in the following sections.

Review of Perspectivist Approaches in AI

As we extensively discussed in the introduction, the scientific communities interested in ML, like NLP, Computer Vision (CV) and medical informatics, have traditionally relied on gold standard datasets to design, develop and evaluate supervised models: these datasets have usually been obtained by the annotation of a single rater or by means of the majority aggregation of few raters. By contrast, the reliability of these gold standard, and their representativeness for the real task under consideration, have scarcely been questioned.

In the recent years, however, due to the huge increase in raw data availability, the increasing reliance on crowdsourcing and similar annotation protocols has highlighted the issue of observer variability in Machine Learning tasks (Cabitza et al. 2019; Schaekermann et al. 2019), an issue which was already well known in certain settings such as the computational linguistics (Artstein and Poesio 2008) and medical ones (Cabitza et al. 2019). Most relevantly, Aroyo and Welty (2015) posited seven "myths" regarding human annotation, a set of principles that traditionally guide data annotation — including the assumption of single possible ground value per case (that is, the *one-truth assumption*), and the overarching goal of avoiding disagreement at all cost — but ultimately hinder the creation of rich datasets that account for human subjectivity. Yun et al. (2021) showed that the majority-aggregated labels in the original ImageNet dataset are not representative of the images in the dataset, due to observer variability and the un-reliability of the annotation process; Svensson, Figge, and Hübler (2015) noted the influence of observer variability on the performance of Machine Learning in a task of cancer detection and proposed ways to measure model performance in settings affected by variability; Schaekermann et al. (2018) noted the large observer variability in tasks as different as sarcasm detection and Parkinson diagnosis, showing how weak perspectivism solutions (in particular group deliberation) could reduce some sources of disagreement and improve label annotations.

While many works, and specifically those focusing on the crowdsourced learning setting, have adopted a weak perspectivist stance for the development of ML methods able to account for this observer variability (Sheng and Zhang 2019); the need for ML methods explicitly taking into account a strong perspectivist approach, however, has only recently started to become a focus of research. Akhtar, Basile, and Patti (2020) showed that a strong perspectivist approach to model training may also lead to performance improvements. Similarly, Kocoń et al. (2021) proposed to leverage non-aggregated data to train models adapted to different users, in what they call "human-centered approach"; Sudre et al. (2019), Gordon et al. (2022) and Guan et al. (2018) proposed multi-task approaches to deal with observer variability and dissenting voices, showing how jointly learning the consensus process and the individual raters' labels improves classification accuracy and representation; Sachdeva et al. (2022a) and Kralj Novak et al. (2022) showed how accounting for disagreements among raters may more accurately represent performance of ML models in hate speech detection and also improve the identification of target groups; similarly, Rodrigues and Pereira (2018) proposed a novel deep learning model that by internally capturing the reliability and biases of different annotators achieves state-of-the-art results for various crowdsourced datasets; Peterson et al. (2019) showed that accounting for raters' disagreement and uncertainty may lead to generalizability and performance improvements in CV tasks; Uma et al. (2020) proposed the use of soft losses as a perspectivist approach for the training of ML models in NLP tasks, while Campagner et al. (2021) proposed a soft loss ensemble learning method, inspired by possibility theory and three-way decisions, for the training of ML models in perspectivist settings; similarly, Washington et al. (2021) showed how the use of soft-labels, that is distributions over labels obtained by means of crowdsourcing, could be useful to better account for the subjectivity of human interpretation in emotion recognition tasks.

In a similar direction, recent work explicitly explored the impact of strong perspectivism on the development and evaluation of supervised models, also from a more conceptual perspective. In particular, in Basile (2021) experiments are presented in support of the thesis that disagreement in annotation may come from the subjectivity of a task to varying extent, and therefore it should not be cast away as noise in the data, but rather it should be systematically accounted for at evaluation time. Sommerauer, Fokkens, and Vossen (2020) note how, for some tasks, "disagreement is not only valid but desired" for the information it carries,

[4]With qualified majority we intend either a *statistically significant* majority (with respect to some hypothesis testing procedure), or an *overwhelming* majority, regardless of how this may be defined, and thus irrespective of what perspectives are considered and how they are distributed within the sample

and proposes an alternative evaluation metric for annotation quality based on answer coherence. Similarly, *disagreement convolution* (Gordon et al. 2021) incorporates disagreement fully into the evaluation pipeline, showing that on natural language tasks related to toxicity and misinformation the performance of traditional ML models is often overstated. Furthermore, Basile (2021), Rizos and Schuller (2020) and Sachdeva et al. (2022b) showed an additional advantage of strong perspectivism in supervised learning, namely its potential impact on the interpretability and fairness of the models. In experiments on real data (annotated corpora of hate speech), Basile (2021) showed how individual labels can be used to cluster the raters by affinity, leading to the emergence of patterns that helps identifying sociodemographic aspects of the raters themselves, which are in principle opaque, especially in a crowdsourcing scenario, while Sachdeva et al. (2022b) showed how the same information could be useful to assess annotator identity sensitivity and thus identify biases in annotation patterns; also Rizos and Schuller (2020) described how a similar approach could be used to detect biases in the data and labels provided by raters. Far from being an exhausted topic, the discussion over perspectivism has recently been fostered, among other venues, at international workshops such as *Investigating and Mitigating Biases in Crowdsourced Data*[5] (ACM CSCW 2021) and the *1st Workshop on Perspectivist Approaches to NLP*[6].

Looking at the Two Sides of the Same Coin

As anticipated above, a perspectivist approach to ground truthing requires to preserve the classification multiplicity instead of getting rid of it by majority voting (if original labels have been produced) or consensus surveys (if original labels do not exist). Obviously, as we previously highlighted when discussing the related works setting the background for our proposal, this comes with some advantages and also some shortcomings, which we discuss in what follows.

The main benefits of the perspectivist approach are:

1. To provide a theoretical backbone that recognizes and accepts the categorical irreducibility of some phenomena. This is especially relevant to those phenomena which exhibit a natural ambiguity, such as many tasks in NLP (Artstein and Poesio 2008), or seemingly inconsistent clinical manifestations (Cabitza et al. 2019);
2. To extract valuable knowledge from what it is usually discarded as noise (cf. label noise (Kahneman et al. 2016)), i.e., disagreement. Such extra information is valuable for a decision support to be more useful in border-line and complex cases;
3. To avoid to ratify and legitimize the opinion of raters belonging to a majority group, re-iterating their truth in seemingly objective advice. Instead, the perspectivist approach aims at giving voice to the few who hold a minority view (Noble 2012), or to those who are intimidated in collective debates;
4. To be able to build models that learn typical human error patterns (if it is plausible to define "errors" on the basis of minority stances) and use this information as a form of decision support;
5. To be able to develop models that can leverage label-uncertain (Uma et al. 2020), fuzzy (Campagner et al. 2021) or soft (Washington et al. 2021) data to improve performance, generalizability and robustness;
6. To allow for cautious methods that represent the uncertainty in the considered phenomena, and provide decision makers with useful advice, that is methods that can improve trust, enhance user experience, and possibly mitigate the risk of automation bias and deskilling.

Since there is no rose without thorns, here we enumerate the main shortcomings that are associated with a perspectivist approach to supervised ML.

1. Need to involve multiple raters: this may represent an important bottleneck in terms of costs or time, and thus result to be impractical or expensive in some domain (e.g., medicine) or when dealing with large datasets;
2. Incompatibility with standard ML approaches, which are usually not designed to take into account multiple perspectives or annotation, and need to design ad-hoc ML methods. While certain classes of learning algorithms (e.g. multi-label ones) may be able to handle multiple labels, it is not clear if these methods can be proficuously applied in the perspectivist ground truthing setting;
3. More complex validation, due to the absence of a uniquely defined ground truth. While in some cases majority labels can be used as a benchmark (Svensson, Figge, and Hübler 2015), these may not be appropriate in strongly subjective or ambiguous settings.

Recommendations and a Research Agenda

This article aims to disseminate a renovated interest for an alternative approach to ground truthing with respect to the "reductionist" one where multiple ratings collected about a single object are reduced into single labels. As we saw in the previous section, this approach entails both advantages and challenges, posed by the will to cope with information richness and manage complexity (also in terms of redundancy, uncertainty and inconsistency) instead of getting rid of it, in light of the research that we presented in our review. In this section, in lieu of a conclusion, we proceed with two sections that shed light on the future: we present a set of agile recommendations for those willing to adopt a perspectivist approach to their ground truthing tasks; and then we propose a number of possible proposals calling for further research and contributions.

Recommendations

In what follows, we share some recommendations to embrace a perspectivist stance in ground truthing. While the impact of some of these practices must still be soundly evaluated, we also mention some of the main studies providing preliminary evidence supporting the recommendations, when available.

[5]https://sites.google.com/view/biases-in-crowdsourced-data
[6]https://nlperspectives.di.unito.it/

- Design annotation schemes that allow raters to associate objects with multiple labels, or also with a 'none of these' label, to account for multiple perspectives directly acknowledged by the single raters. Moreover, allow the raters to express a judgment of inadequacy of the available label set (see (Aroyo and Welty 2015));
- Involve *enough* raters. This can mean different things depending on the application domain: a number that allows for statistically significant majorities to emerge (e.g., at least 12 raters for dichotomous tasks) or a number of raters that allows to create a majority of superhuman accuracy, according to some estimate of the average accuracy of the raters (see (Cabitza et al. 2020));
- Involve heterogeneous raters, both in regard to their origin and culture as well as to their expertise and skills: different opinions are not always a source of noise, as asserted in (Kahneman et al. 2016), but rather of richness (Aroyo and Welty 2015);
- Evaluate and validate ML models also with respect to *robustness*, or their capability to adequately perform also on out-of-distribution data, that is data coming from settings other than where the training data were produced. If performance degrades on external datasets, this could suggest to adopt a perspectivist approach as a way to mitigate the risk that the models "overfit" to a non-representative sample of users;
- Report about the rater enrollment process and about the quality of their ratings in detail, describing the process of ground truthing in terms of: a) Number of raters involved to produce the labels; b) The raters' profession and expertise (e.g., years from specialization or graduation); c) The incentive provided, if any; d) Particular instructions given to raters for quality control (e.g., which data were discarded and why); e) how long the labelling process took and, in the case of critical domains such as medicine and law, where and under what conditions it took place (e.g., controlled conditions, real-world interruptions); f) Any chance-adjusted measure of inter-rater agreement (e.g., Rho (Cabitza et al. 2020), Krippendorff's Alpha, Fleiss' Kappa); g) The Labelling technique (e.g., majority voting, Delphi method, structured adjudication (Schaekermann et al. 2019), consensus iteration or swarm intelligence). Other relevant information such as those indicated in the previously mentioned Data Statements (Bender and Friedman 2018) or Datasheets for Datasets (Gebru et al. 2021), and regarding for example the data collection and sharing details or the intended uses for the collected data, should also be reported;
- Collect additional information from the raters involved, so as to take into account of their perspective and way of seeing the objects at hand as comprehensively as possible. Examples of such side-information pieces that could be collected include the confidence expressed by the raters about each of their annotations in terms of an ordinal score, as well as other dimensions regarding the objects to classify (e.g., complexity, difficulty, rarity, relevance) along similar scales. These annotation metadata can be useful for multiple purposes: for instance, confidence scores can be used to weight the raters' annotations in aggregating procedures (Campagner et al. 2021), or to detect the cases that were the most difficult ones to decide about (e.g. those associated with the lowest raters' confidence); likewise, relevance or complexity scores can be used to detect those cases that it is most important that the ML model gets right (i.e., with a sufficient average accuracy) not to mislead its users.
- Interpret the concept of majority flexibly: as said above, majority voting can be weighted by either the raters' confidence, or accuracy (which can be evaluated by either profiling, preliminary testing or even with respect to the majority of the others' judgments taken as gold standard). However, even the idea of general case-wise majority could be challenged, e.g., by considering the largest coherent minorities, that is the majority opinion within the groups of raters who agree more on specific sub-partitions of the whole dataset.

These recommendations are complementary to those proposed by the *Perspectivist Data Manifesto*[7], a collaborative initiative to promote a perspectivist research agenda in the AI community. In particular, the signatories of the above manifesto also recommend to create and distribute non-aggregated datasets, in order to foster the discussion on the principles of perspectivism among the research community and to facilitate experimental research in this direction. Finally, at a more general level, our proposal suggests to rethink any theoretical and experimental research work in artificial intelligence and machine learning under the perspectivist lens, asking questions such as *whose opinion is the model relying on in its prediction?* and their implications towards the ethics and responsibility of choices made with the support of predictive models.

Open Problems for a Possible Research Agenda

Finally, in what follows we delineate some possible open problems and research directions that we deem relevant to advance the perspectivist stance in artificial intelligence and machine learning:

1. First, and in connection with the manifesto mentioned in the previous section, we recommend the creation and dissemination of benchmark datasets that could be used to evaluate perspectivist ML models, possibly with respect to different data types (like images, texts, structured data) and for different classification tasks (e.g., detection, risk stratification, forecasting). Such datasets are obviously necessary for the evaluation of novel algorithmic proposals, but could be used also as benchmarks for setting up challenges, which have recently proven to be important drivers for the development of novel methodologies;
2. Disagreement (and possibly errors) in the multi-rater labels are part and parcel of the perspectivist stance to ground truthing. It is of interest to develop techniques and approaches that are able to exploit the multiple labels to understand and model how the raters err, or on which types of objects they disagree more, so as to develop

[7] https://pdai.info

learning models with the ability to predict the chance that the raters would err on a new object;

3. The perspectivist stance in ground truthing amounts to associating multiple labels to each object. However, especially in low-intersubjectivity tasks, a certain amount of divergent labels is expected in, and may be intrinsic to, the task, and a part of these could be due to errors, inattention or negligence. Thus, it would be interesting to develop ML models that are effectively able to disentangle subjectivity from error, and also to characterize (from a learning theoretic perspective) when this disentanglement is possible at all;
4. In the literature, different algorithmic approaches able to account for a perspectivist ground truth have been proposed: these include data augmentation strategies based on the replication of objects associated with multiple labels (Nishi et al. 2021), ensemble learning methods (Campagner et al. 2021), or soft loss approaches (Uma et al. 2020). It would be interesting to assess, on real-world problems and applications, the performance of these (and other) approaches, so as to understand their properties and the appropriateness of different algorithms for specific tasks;
5. While some ways to evaluate perspectivist ML models have been proposed in the literature (e.g., by evaluating the performance of ML models on objects associated with "overwhelming majorities" (Campagner et al. 2021; Svensson, Figge, and Hübler 2015), or by adopting soft loss metrics (Uma et al. 2020)), these proposals should be unified to develop novel metrics that can better take into account the (chance-adjusted) agreement rate between the raters, or the presence of chance effects and label noise;
6. Similarly to the above point, further research is needed on the impact of observer variability on the predictive performance of ML models. A particularly promising direction regards the definition of generative models for simulating this kind of variability: such models would allow to test for the robustness of perspectivist ML models and get an estimate of the extent any model is "overfitting" on either opinions that are not representative of the user population or on partial aspects of the phenomenon of interest.
7. A perspectivist ground truth could be seen as rich as it is noisy. While standard measures of inter-rater agreement (or similar measures, like entropy) can capture this "noise" aspect within a *diamond standard*, these measures are not fully capable to represent the informational content and value of such multi-faceted information (Campagner et al. 2021). Therefore, this approach needs the definition of a theoretical framework by which to evaluate the compromise between richness (in that the more different perspectives, the better) and reliability (in that multiplicity does not indicate confusion but complementarity).In doing so, learning-theoretic characterizations of perspectivist ML can be developed, which could enable to understand when a given dataset can be reliable "ground" for sound decision support; or, conversely, when its quality needs to be improved with some intervention of *decision hygiene* (Kahneman et al. 2016), like, e.g., aggregating judgements with methods that leverage professional expertise;
8. As previously described in (Basile 2021; Rizos and Schuller 2020), the perspectivist stance could also be potentially applied to the aim of increasing model interpretability and algorithmic fairness, e.g. by enabling the detection of biases and discrimination induced by (some groups of) raters. Further research should thus be devoted toward the investigation of possible applications of perspectivist ML in eXplainable and fair AI;
9. Finally, we believe it would be interesting to consider settings in which raters are able to express more information than a single label (Aroyo and Welty 2015) – for example, by providing a ranking of the possible labels or expressing their confidence (Cabitza et al. 2020). Given the similarity of these settings to the problems typically investigated in the field of social choice, further research should investigate the approaches proposed in that context and how they could be applied to design perspectivist ML methods that could deal with more general, structured label representations.

To conclude this contribution and review of the ideas and promising lines of research that can be traced back to the tenets of *perspectivism*, which we outlined it in this proposal, we summarize the following points: the perspectivist approach is essentially aimed at *caring about* the representativeness and reliability of the ground truth of ML systems. More specifically, and programmatically, perspectivism fosters wariness in aggregated gold standards: these reference datasets express single-truth assumptions (Aroyo and Welty 2015) that can fall short of capturing the necessary complexity of the phenomena for which we want to have support from computational means (Bechmann and Bowker 2019). After all, ML systems are complicated machines that essentially reiterate past judgments and legitimate them (Hildebrandt 2021) by putting the perceptions of very few (relatively speaking) raters to the attention of countless users and decision makers. There is no guarantee that new objects will be comparable to (or equatable with) those with which such models had been trained, not to speak of the contingent context. We believe that adapting the single-truth assumption of ML to the perspectivist paradigm is not only more fair towards minority opinions, but it also has the potential to yield a more accurate and explainable quantitative evaluation of the trained models, as shown by the recent works that are exploring this research direction (Uma et al. 2020).

Thus, such a stance creates the necessary room for asking questions such as *whose opinion is the model relying on in its prediction?*, and *what opinions do we want to project into the human interpretation of the unexpected new?*; and for reflecting on the implications of these matters on the ethical nature and impact of the decisions made with the support of predictive computing means.

References

Akhtar, S.; Basile, V.; and Patti, V. 2020. Modeling Annotator Perspective and Polarized Opinions to Improve Hate Speech Detection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, 151–154.

Aroyo, L.; and Welty, C. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1): 15–24.

Artstein, R.; and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4): 555–596.

Balayn, A.; Bozzon, A.; and Szlavik, Z. 2019. Unfairness towards subjective opinions in Machine Learning. *arXiv preprint arXiv:1911.02455*.

Basile, V. 2021. It's the End of the Gold Standard as we Know it. Leveraging Non-aggregated Data for Better Evaluation and Explanation of Subjective Tasks. In Baldoni, M.; and Bandini, S., eds., *AIxIA 2020: Advances in Artificial Intelligence. XIXth International Conference of the Italian Association for Artificial Intelligence, Virtual Event, November 24-27, 2020, Revised and Selected papers*, volume 12414. Springer Nature Switzerland AG.

Bechmann, A.; and Bowker, G. C. 2019. Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society*, 6(1).

Bender, E. M.; and Friedman, B. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6: 587–604.

Cabitza, F.; Campagner, A.; Albano, D.; Aliprandi, A.; Bruno, A.; Chianca, V.; Corazza, A.; Di Pietto, F.; Gambino, A.; Gitto, S.; et al. 2020. The elephant in the machine: Proposing a new metric of data reliability and its application to a medical case to assess classification reliability. *Applied Sciences*, 10(11): 4014.

Cabitza, F.; Campagner, A.; and Mattioli, M. 2022. The unbearable (technical) unreliability of automated facial emotion recognition. *Big Data & Society*, 9(2): 20539517221129549.

Cabitza, F.; Locoro, A.; Alderighi, C.; Rasoini, R.; Compagnone, D.; and Berjano, P. 2019. The elephant in the record: on the multiplicity of data recording work. *Health informatics journal*, 25(3): 475–490.

Campagner, A.; Ciucci, D.; Svensson, C.-M.; et al. 2021. Ground truthing from multi-rater labeling with three-way decision and possibility theory. *Information Sciences*, 545: 771–790.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW): 1–22.

Chernova, S.; and Veloso, M. 2010. Confidence-based multi-robot learning from demonstration. *International Journal of Social Robotics*, 2(2): 195–215.

Draws, T.; Rieger, A.; Inel, O.; Gadiraju, U.; and Tintarev, N. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 48–59.

Dumitrache, A.; Aroyo, L.; and Welty, C. 2018. Capturing ambiguity in crowdsourcing frame disambiguation. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6.

Eickhoff, C. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 162–170.

Freeman, B.; Hammel, N.; Phene, S.; Huang, A.; Ackermann, R.; Kanzheleva, O.; Hutson, M.; Taggart, C.; Duong, Q.; and Sayres, R. 2021. Iterative Quality Control Strategies for Expert Medical Image Labeling. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 60–71.

Gadiraju, U.; Fetahu, B.; and Kawase, R. 2015. Training workers for improving performance in crowdsourcing microtasks. In *European Conference on Technology Enhanced Learning*, 100–114. Springer.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gordon, M. L.; Lam, M. S.; Park, J. S.; Patel, K.; Hancock, J.; Hashimoto, T.; and Bernstein, M. S. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*, 1–19.

Gordon, M. L.; Zhou, K.; Patel, K.; Hashimoto, T.; and Bernstein, M. S. 2021. 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*, 1–14.

Guan, M.; Gulshan, V.; Dai, A.; and Hinton, G. 2018. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Gurari, D.; and Grauman, K. 2017. Crowdverge: Predicting if people will agree on the answer to a visual question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3511–3522.

Hildebrandt, M. 2021. The Issue of Bias. The Framing Powers of Ml. In editor, T., ed., *Machine Learning and Society: Impact, Trust, Transparency*. MIT Press.

Holland, S.; Hosny, A.; Newman, S.; Joseph, J.; and Chmielinski, K. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*.

Hube, C.; Fetahu, B.; and Gadiraju, U. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Kahneman, D.; Rosenfield, A.; Gandhi, L.; and Blaser, T. 2016. Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making. *Harvard Business Review*, 94: 38–46.

Kairam, S.; and Heer, J. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1637–1648.

Kocoń, J.; Figas, A.; Gruza, M.; Puchalska, D.; Kajdanowicz, T.; and Kazienko, P. 2021. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5): 102643.

Kralj Novak, P.; Scantamburlo, T.; Pelicon, A.; Cinelli, M.; Mozetič, I.; and Zollo, F. 2022. Handling Disagreement in Hate Speech Modelling. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 681–695. Springer.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

Muller, M.; Wolf, C. T.; Andres, J.; Desmond, M.; Joshi, N. N.; Ashktorab, Z.; Sharma, A.; Brimijoin, K.; Pan, Q.; Duesterwald, E.; et al. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Nishi, K.; Ding, Y.; Rich, A.; and Höllerer, T. 2021. Improving Label Noise Robustness with Data Augmentation and Semi-Supervised Learning (Student Abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 15855–15856.

Noble, J. A. 2012. Minority voices of crowdsourcing: Why we should pay attention to every member of the crowd. In *proceedings of the ACM 2012 conference on computer supported cooperative work companion*, 179–182.

Paullada, A.; Raji, I. D.; Bender, E. M.; Denton, E.; and Hanna, A. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11): 100336.

Peterson, J. C.; Battleday, R. M.; Griffiths, T. L.; and Russakovsky, O. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9617–9626.

Richards, J.; Piorkowski, D.; Hind, M.; Houde, S.; Mojsilovic, A.; and Varshney, K. R. 2021. A Human-Centered Methodology for Creating AI FactSheets. *Data Engineering*, 47.

Rizos, G.; and Schuller, B. W. 2020. Average Jane, Where Art Thou?–Recent Avenues in Efficient Machine Learning Under Subjectivity Uncertainty. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 42–55. Springer.

Rodrigues, F.; and Pereira, F. 2018. Deep learning from crowds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Sachdeva, P.; Barreto, R.; Von Vacano, C.; and Kennedy, C. 2022a. Targeted Identity Group Prediction in Hate Speech Corpora. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, 231–244.

Sachdeva, P. S.; Barreto, R.; von Vacano, C.; and Kennedy, C. J. 2022b. Assessing Annotator Identity Sensitivity via Item Response Theory: A Case Study in a Hate Speech Corpus. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1585–1603.

Schaekermann, M.; Beaton, G.; Habib, M.; Lim, A.; Larson, K.; and Law, E. 2019. Understanding expert disagreement in medical data analysis through structured adjudication. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–23.

Schaekermann, M.; Goh, J.; Larson, K.; and Law, E. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 1–19.

Sheng, V. S.; and Zhang, J. 2019. Machine learning with crowdsourcing: A brief summary of the past research and future directions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9837–9843.

Sommerauer, P.; Fokkens, A.; and Vossen, P. 2020. Would you describe a leopard as yellow? Evaluating crowd-annotations with justified and informative disagreement. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4798–4809. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Sudre, C. H.; Anson, B. G.; Ingala, S.; Lane, C. D.; Jimenez, D.; Haider, L.; Varsavsky, T.; Tanno, R.; Smith, L.; Ourselin, S.; et al. 2019. Let's agree to disagree: Learning highly debatable multirater labelling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 665–673. Springer.

Svensson, C.-M.; Figge, M.-T.; and Hübler, R. 2015. Automated Classification of Circulating Tumor Cells and the Impact of Interobsever Variability on Classifier Training and Performance. *Journal of Immunology Research*, 2015.

Uma, A.; Fornaciari, T.; Hovy, D.; Paun, S.; Plank, B.; and Poesio, M. 2020. A Case for Soft Loss Functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, 173–177.

Vaughan, J. W. 2017. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *J. Mach. Learn. Res.*, 18(1): 7026–7071.

Washington, P.; Kalantarian, H.; Kent, J.; Husic, A.; Kline, A.; Leblanc, E.; Hou, C.; Mutlu, C.; Dunlap, K.; Penev, Y.; et al. 2021. Training Affective Computer Vision Models by Crowdsourcing Soft-Target Labels. *Cognitive Computation*, 13(5): 1363–1373.

Yun, S.; Oh, S. J.; Heo, B.; Han, D.; Choe, J.; and Chun, S. 2021. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2340–2350.