New Frontiers in Explainable AI: Understanding the GI to Interpret the GO

Federico Cabitza, Andrea Campagner, Davide Ciucci

New Frontiers in Explainable AI: Understanding the GI to Interpret the GO

Federico Cabitza[1,2], Andrea Campagner[1], and Davide Ciucci[1]0000000280837809

[1] Universitá degli Studi di Milano-Bicocca, Milan, Italy
federico.cabitza@unimib.it
[2] IRCCS Istituto Ortopedico Galeazzi, Milan, Italy

Abstract. In this paper we focus on the importance of interpreting the quality of the input of predictive models (potentially a GI, i.e., a Garbage In) to make sense of the reliability of their output (potentially a GO, a Garbage Out) in support of human decision making, especially in critical domains, like medicine. To this aim, we propose a framework where we distinguish between the Gold Standard (or Ground Truth) and the set of annotations from which this is derived, and a set of quality dimensions that help to assess and interpret the AI advice: fineness, trueness, representativeness, conformity, dryness. We then discuss implications for obtaining more informative training sets and for the design of more usable Decision Support Systems.

Keywords: Ground Truth, Explainable AI, Reliability, Usable AI

1 Introduction

In the specialist literature around the topics of Fairness, Accountability, and Transparency in Machine Learning (FAT-ML), many approaches to make *AI explainable* (XAI) are proposed and discussed. A XAI system can be *intrinsically interpretable*, when it adopts a model whose internal functioning is immediately accessible to the decision maker, like in the case of linear or rule-based models (e.g., decision trees); or it can be made interpretable by focusing on two aspects: the model itself; or its output on one or more given cases. The former case of interpretability (also called *understandability* or *intelligibility*) regards "how the model works": this kind of model interpretability is pursued by providing the decision makers, i.e., the users of XAI systems, with indications about how the model produced a certain prediction, e.g., by plotting the loss function, or by visualizing the boundary region on a PCA-reduced space, or by telling what feature the model based more on to produce its prediction, as represented by feature relevance scores or saliency maps. In the latter case, when authors speak also of *post-hoc interpretations*, the focus is on output data, and the aim is "to explain the predictions without elucidating the mechanisms by which models work" [31]. In this case, decision makers can be given counterfactual outputs

(that is alternative outputs if the input case were different) or the rules or functional relationships that locally apply for the output of surrogate (and more interpretable, in the sense mentioned above) models. These models are intended to locally "simulate" the black-box model "at the terminals", and explain the original relationship between the prediction and the input instance more intuitively. This approach is also the basis for the only proposal, to our knowledge, to make the concept of interpretability fully formalized [29].

In this paper we want discuss a third, and still neglected, general approach: instead of focusing on either the model or its outputs, we aim to discuss *input explainability*, that is on ways to have the decision makers to get an idea of how much they should trust the single output prediction on the basis of the "quality" of the *ground truth* on which the model has been trained, that is on the basis of the input of the learning process that yielded the model.

2 First things first: the importance of input

Ground Truth, or Gold Standard (as the reference data are commonly called in medicine, our reference domain), is assumed to be *true*, by definition: the ML model is then supposed to "learn" from it the hidden patterns actually lying in the complex and manifold relationships between the phenomenon's predictors (variables that express the phenomenon symbolically) and the target variable (seen as a sort of interpretation or further measure of the phenomenon). However, any data is but an approximation of reality, a mere representation of it: as obvious as it sounds, maps are not the territory, likewise, also our "truths" are more "map truth" rather than ground truths. However, scholars in the Machine Learning and AI communities seldom address the question of *how good their ground truth actually is*, that is how much "golden" their Gold Standard is (or, to adopt the jewellery jargon, what its *fineness* is).

Most works that compare machine and human performance in delicate tasks, like diagnostic ones in medical practice, assume ground truth good enough to yield reliable results but, at the same time, understand that relying on the interpretation of a single source or interpreter would be over-optimistic, hence lead to too inaccurate performance. For this reason, Gold Standard sets are usually built by gathering a number of observers (or raters, annotators) and asking them to observe a phenomenon of interest (i.e., a unit of observation, or case), judge it, rate it and annotate the sets of data that describe it with a value from a scale of measurement, which can be either scalar, ordinal or nominal in nature. In this later case, the raters annotate the case with a code, class or category, which best describes the case. The ML model is then aimed at associating the one best class with any new case extracted from the same reference population.

The multiplicity of ratings at the origin on the Gold Standard does not result only from multi-rater settings, but also when there is the necessity to "sample" a complex phenomenon with multiple measurements. For instance, a Gold Standard could regard the outcome of a medical intervention as it is perceived 3 months after the intervention; this outcome could be represented in terms of

PROs (i.e., Patient Reported Outcome Measures), by asking the patient to report how they feel on an ordinal scale a number of times in the week occurring approximately a dozen of weeks after the intervention, and then averaging these measures [6].

Both in multi-rater and in single-rater settings, it is seldom considered whether the Gold Standard built from a set of annotations is reliable or not, i.e., whether each case were described by a sufficient number of rating, or whether the raters involved were expert or adequately committed to the task. For instance (to limit ourselves to some of the most relevant works in medical AI), the authors of [14] report to have used Gold Standard diagnoses "based on expert opinion (including dermatologists and dermatopathologists)" from open-source repositories, where yet the details on the number and expertise of the raters involved are not available. Also the supplementary materials related to the work by Haenssle and colleagues [21] do not provide any detail on the number of dermatologists involved. The dataset used in [22] was annotated by just three dermatologists. The data set used in [37] for the task to detect tumor cells was annotated by non-specialists. One of the studies that has involved more raters to date, i.e., the study mentioned in [19], involved 54 raters, and these were all either US-licensed ophthalmologists or ophthalmology trainees; however, we do not know the proportions of trainees, and inter-rater reliability was assessed for less than a third of the sample, as only 16 raters had graded a sufficient volume of repeat images; furthermore, agreement proportions were not adjusted for chance effects. Although these are only anecdotal mentions, we argue that current debate on accuracy (and explainability) of AI focuses primarily on the technology (i.e., the model), and not on the underlying data, whose production and validation still lies in the background.

Notwithstanding this relative lack of transparency on the number and skills of the original annotators involved in ground truthing, in various ambits – and especially in medicine – the phenomenon of observer variability has been known (and studied) since decades [3]: this phenomenon regards how different observers, who are called to annotate data can simply differ and disagree with each others. This observer variability affects the reliability of the resulting data set, what we call Diamond Standard (as it represents a multi-perspective view on, and a multi-facet record of, reality). In the context of observer variability assessment reliability is defined as the concordance of repeated measurements (the annotations of the multiple annotators) and is usually calculated by the intraclass correlation coefficient (ICC) [33], which estimates the average correlation among all possible orderings of data pairs. As ICC is sensitive to data range also standard error of measurement SEM is proposed as a measure of variability in case of scalar values.

To this regard we will focus on questions such as: how much *true* is the ground truth? To this respect, we will introduce the concept of *fineness* of the Gold Standard. How much *reliable* is the ground truth? We relate ground truth reliability to the extent the single "measuring instruments", often human annotators, are *accurate* in their measure (i.e., in mapping a property of the object

of interest to a value) and *precise* with respect to each other, i.e., how much their measures/annotations vary (or agree upon each other) for a single object of observation. How much *informative* (or *representative*) is the ground truth with respect to the reference population? This is also related to its *conformity*, that is the degree of resemblance between the available data and the reference population from which they have been drawn. How much *uncertain* is the set (which we call Diamond Standard) of all of the observations from which the ground truth is derived? To try to address the above questions, we propose a general framework to circumscribe the main concepts regarding the quality of data feeding the learning process of Machine Learning. With reference to Figure 1, we call *Gold Standard* the training set, that is the data set where each case is annotated with a unique "true" value for the target feature. We distinguish it from the *Diamond Standard*, that is the data set where multiple (m) annotators (also called raters or observers), have associated the description of the cases to the target class. We call *reductions*, the data transformations that produce the Gold Standard from the Diamond Standard: reductions necessarily entail some information loss, because they allow to pass from a multi-rater labelling to "the one best" labelling by a "collective" rater. Obviously, if $m = 1$, the Gold Standard and the Diamond Standard coincide. On the basis of the number and interpretative skills of the annotators the Diamond Standard represents a more or less approximate representation of the truth (still yet, a symbolic and *datafied* expression of the truth), that is, of an *unknown* (and *unknowable*) data set that we call the *UR-SET* (Ultimately Realistic Symbolic Expression of Truth [3]).

In the next sections, we will consider methods to assess the quality of the input of the learning process, that is the data with which it has been trained to produce an accurate output when applied to new instances of data: we will illustrate the common cases of *reliability* and *representativeness*, and will introduce three original dimensions, by distinguishing between the *Fineness* and *Dryness* (of the Gold Standard), and the *Trueness* of the Diamond Set.

3 Reliability

The intuitive notion of reliability is straightforward: how much can we *rely upon* our ground truth to make decisions? How much can a ML model rely on its training set to make realistic (beside accurate[4]) predictions? More technically, the reliability of a dataset regards the *precision* of the measures it contains, for each case that it represents. This allows us to speak of reliability of a Gold Standard only in terms of the reliability of the Diamond Standard from which it has been derived. The reliability of a Diamond Standard regards the extent

[3] Notably, the UR-SET could be annotated with a different alphabet than the Gold Standard and the Diamond Standard. For instance, while the Gold Standard uses a binary symbol set (e.g., positive/negative) the UR-SET could be annotated with a set encompassing a symbol expressing that a case is 25% positive and 75% negative.

[4] Accuracy is historically defined in terms of *closeness* between the prediction and the Gold Standard, not with respect to the reality.

this set expresses a *unitary* interpretation of the single cases observed, despite the multiplicity of views entailed by the different raters involved in interpreting each case. If all of the raters agree upon each and every case (or the single raters agree with themselves, as in the case of the PRO meaasures mentioned above), that is if no disagreement among the case's annotations has been observed, both the reliability (and the trueness, as we will see) are maximum[5].

Over time, many measures of inter-rater agreement, and hence reliability, have been proposed, like the Fleiss's Kappa, the Cohen's Kappa, or the Krippendorff's Alpha. These indices aim to go beyond the simple proportion of matched pairs (a score called Proportion of Agreement, and usually denoted as P_o). This aim is motivated for the important, and often neglected, limitation of the P_o: it includes the amount of agreement that could be due to chance, and hence it produces an overly optimistic measure of the real agreement. All of the proposed metrics present some limitations, for instance in regard to the presence of missing values, or to the nature of ratings (e.g., categorical or ordinal), and all of them are subject to a number of paradoxes, e.g., when the cases to be rated are not well-distributed across the rating categories [34].

Unfortunately, scholars interested in assessing the reliability of annotated data still often rely on one of the indices presenting the most severe methodological problems [15], i.e., the Kappa; and, what is worst, they still usually adopt the range divisions proposed by Landis and Koch in 1977 [30] to interpret the scores, i.e., a scale that is obsolete, related to the first formulations of the Kappa, is "clearly arbitrary" (as frankly admitted by the first proponents), and mani-

[5] That notwithstanding, it would be inaccurate to say that the Diamond Standard coincides with the UR-SET, which is unknowable.

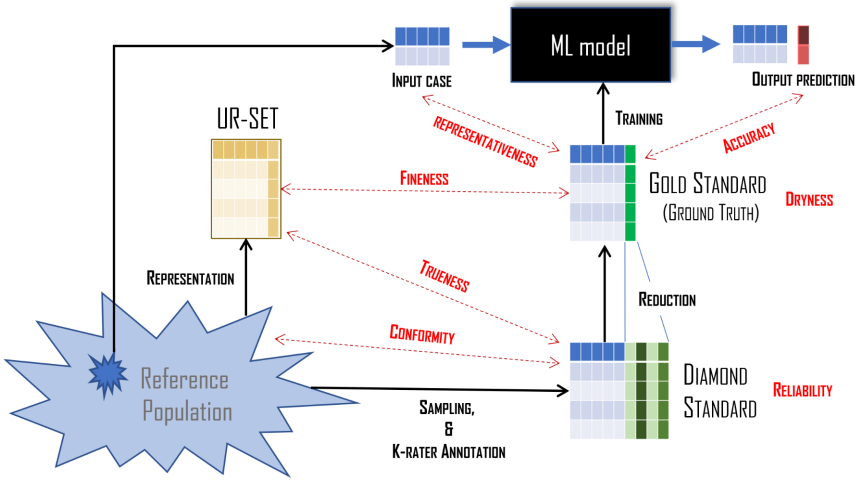

Fig. 1. The general framework and the main concepts illustrated in this contribution.

festly inflating the degree of agreement (e.g., for the agreement to be considered "fair" it is sufficient that only 20% of times raters agree beyond the effect due to chance), likely one of the reasons for its fortune[6]. For these reasons we propose the adoption of more robust reliability measures, like the Krippendorf's Alpha, and to follow the indications for its interpretation given by Krippendorf [28]: he proposed to consider as sufficiently reliable for critical applications, like in the case of medical interpretation and prediction, collective annotations that would be associated with an Alpha of .8, or above, only. Krippendorff also considers two more robust criteria for acceptable reliability, both considering the distribution of the α (computed via bootstrapping): the first considers computing the confidence interval $[\alpha_{min}, \alpha_{max}]$ and then establishing acceptable reliability if the established threshold $\alpha_{required}$ (at least 0.8, as previously specified) is lower than α_{min}; the second approach, on the other hand, consists of computing the probability q that $\alpha \leq \alpha_{required}$ and then confronting this probability q with an a priori confidence threshold. These demanding requirements are seldom verified in the ML literature and, when they are, even less frequently met. We raised awareness on the issue of low reliability of the ground truth used to train medical AI in [5] and [7]. In this latter study we reported the low agreement between multiple raters in two settings from different medical specialties: cardiology and spine surgery. It is important to notice, yet, that disagreements do not occur only because some rater is less skilled than the others, and hence commits an interpretation error (due to what is called *label bias* [25]); in fact, this is seldom the case. More often, it is the *intrinsic ambiguity* of the *interpretand* phenomenon that brings raters to different, yet equally plausible, interpretations [5]. Other factors that could undermine the potential for agreement between raters, and hence the reliability of the Diamond Standard (and then the Gold Standard as mentioned above), are related to differences in how the raters react to the experimental conditions in which their opinions and interpretations are collected (since ground truthing tasks occur often in controlled experimental settings), and more generally, to the fact of being involved in an experiment. These phenomena are generally known as "Hawthorne effect" [36], but it is not clear whether the "awareness of being observed or involved in an experiment" affects the ratings more in terms of increasing the accuracy (up to levels that in real-world settings would not be tenable, mainly for conditions of uninterrupted concentration and focused commitment), or rather in terms of its reduction (an effect known as "laboratory effect" [20], which is mainly due to lack of real motivations, engagement or just of the fear of consequences in case of errors).

4 Representativeness and Conformity

"Representative" is a term that equally applies to individuals, with respect to a group from which they are ideally drawn; and to groups, with respect to

[6] To date, this single contribution has been cited almost 50,000 times, but likely more often by habit and imitation, than by the deliberate adoption of the assumptions therein discussed.

wider groups, or populations, from which these groups are drawn as samples. To consider both these kinds of representativeness and, at the same time, avoid potential ambiguities, we distinguish between the *representativeness* of the Gold Standard, with respect to the single new case to predict; and the *conformity* of the Diamond Standard, with respect to the reference population. This analysis requires to focus on the *moments* of the probability distribution of our data: what we call *representativeness* regards the first moment, i.e., the centroid of the distribution, while *conformity* regards other higher-order moments, like variance, skewness, and kurtosis (the "shape" of the multi-dimensional distribution).

The simplest way to assess the *conformity* of the Diamond Standard is to consider, when available, the reference distributions of the single features, considering them separately: we call this basic type of population representativeness $conformity_u$, and this is based on the strong assumptions to know how the multivariate distribution of the population really is (e.g., from census information or other random sampling surveys), and that its change rate (or time constant) is negligible with respect to the sampling procedure.

Suppose that f is a categorical feature with k possible values, then we can test the *conformity* using the χ^2 goodness–of–fit test or the *G-test*; if f is an ordinal or continuous feature instead, we can apply the *Kolmogorov–Smirnov test*. In both cases, the obtained statistic (or the related p–value) represents a degree of the extent the Diamond Standard is similarly shaped with respect to the reference population.

When having access to the full joint distribution for the reference population we can extend the approach above described to define a multivariate definition of conformity, that we denote as $conformity_m$, using the multivariate versions of the respective statistical tests (see, as an example, [26] for a multivariate extension of the Kolmogorov–Smirnov test).

If we also have access to an analytic or model–based representation $\mathcal{M}$ of the reference population distribution we can give a third measure of conformity, that we term $conformity_p$, by directly computing the probability of the Diamond Standard D given $\mathcal{M}$, $P(D|\mathcal{M})$ and then sample (e.g., using Markov Chain Monte Carlo simulation techniques) the model in order to compute the probability q to obtain a probability $P \leq P(D|\mathcal{M})$ which can be taken as a measure of *conformity* (i.e., the greater q the greater our belief that D is indeed a fair representation of the reference population), because large values of q would imply that the Diamond Standard D is indeed "more probable" than most datasets generated according to the reference population distribution.

On the other hand, *representativeness* is defined between a given *input case* for the ML model, drawn from the reference population, and the Gold Standard: the Gold Standard is said to be representative of the input case if the input case resembles a "typical member" of the Gold Standard. This concept, while not usually evaluated, is important in checking whether the prediction that we would obtain from our model is meaningful; indeed, one of the major assumption of ML methodologies is that all the cases (the ones given as training examples as well as those which we are interested in making predictions on) come from the

same distribution, that is are independent and identically distributed (IID). The most basic approach is to consider a case x representative of the Gold Standard G if it is "close" to its center, as described by the following algorithm:

Algorithm 1: Centroid–based Representativeness

Data: Gold Standard G, input case x
Result: Representativeness $r_c(G, x)$ of x

1 $c = \frac{1}{|G|} \sum_{p \in G} p$;
2 $dist(x, c) = \sqrt{\sum_{f \in F} (\frac{v_f^x - v_f^c}{v_f^{max} - v_f^{min}})^2}$;
3 $r_c(G, x) = 1 - \frac{dist(x,c)}{max\{dist(p,c) | \text{p is not an outlier}\}}$;

where $\sum_{p \in G} p$, assuming that the instances belong to a vector space, is simply defined as the vector sum, F is the set of all features and the outlierness of a case is established via any outlier–detection algorithm.

The centroid–based representativeness r_c assumes values in $(-\infty, 1]$, with maximum value when x is exactly equal to the centroid of the Gold Standard. This basic technique, while simple also from a computational point of view, has various limitations: the most relevant one is that the whole distribution of G is not taken into account: the centroid in itself could be a non–representative point of G; x, while being quite distant from the center, could be in a region of the feature space which is actually homogeneous with respect to the distance and so on. A more valid approach would be to consider locality–based outlier–detection algorithm, such as the *Local Outlier Factor* [4], as described in the algorithm 2 which is based on the statistical transformation defined in [27].

Algorithm 2: Locality–based Representativeness

Data: Gold Standard G, input case x, number of neighbors k
Result: Representativeness $r_l(G, x, k)$ of x

1 $k - distance(x) = d(x, p_k)$ where $p_k$is the k-th nearest neighbor of x;
2 $\mathcal{N}_k(x) = \{p \in G | dist(x, p) \leq k - distance(x)\}$;
3 $S(x) =$ Locality-Based-Outlier-Scoring$(x, N_k(x))$;
4 $R(x) = max\{0, S(x) - 1\}$;
5 $r_l(G, x, k) = max\{0, erf(\frac{R(x) - \mu_R}{\sigma_R * \sqrt{2}})\}$

In the algorithm erf is the *Gaussian error function*, $S(x) \in [0, +\infty)$ and *Locality-Based-Outlier-Scoring* refers to any locality–based outlier detection algorithm. The locality–based representativeness can be understood as the probability of obtaining, from the Gold Standard G, a point similar to x, considering its connectivity degree (how much is it near to its nearest points) with respect to that of its neighbors. We also notice that algorithm 2 can be used also in case of nominal attributes and missing values, by means of a suitable distance [40].

A last approach to define a measure of representativeness, which we denote as $r_p(\mathcal{M}_G, x)$, can be given when we have access to a generative ML model $\mathcal{M}_G$

for G. In this case, using a procedure analogous to the one used for defining $conformity_p$, we can compute the probability of x given the model $P(x|\mathcal{M}_G)$ and then sample the model to evaluate the probability q of getting a probability value as extreme as $P(x|\mathcal{M}_G)$. In the case of representativeness we could also refine this approach in order to define a local version of $r_p(\mathcal{M}_G, x)$ by limiting the sampled cases to ones belonging to a neighborhood of x.

An open question that these reflections invite to consider regards the feedback loop that could be established between the model's predictions (which affect the human decision making) and the reference population. If the decisions affected by the AI's advice can have an impact on the population from which new cases are to be extracted (like in case of prognostic models, where the model suggests how much an intervention could improve the health conditions of a patient, and hence also suggest who should receive a treatment, or an intensive one, and who should not), then it should be considered that the representativeness of the Gold Standard could change accordingly, usually for the worse. This would urge us for a continuous update of both the Diamond and the Gold Standard, or for the need to stratify the past interventions by distinguishing those who were likely impacted by the decision aid (directly or indirectly) and those who were not, and extract new cases for the ground truthing process from this latter portion of the reference population, using techniques akin to active learning [32].

5 Fineness of the Gold Standard

The $fineness(G, O)$ is the *probability* that the Gold Standard G, obtained from the Diamond Standard D by means of a *reduction* – e.g., taking the majority vote over a set O of observers (i.e., the mode for each case) – is equal to the true (unknowable) annotation (i.e., interpretation) of the portion of the reality of interest, what we call the $UR{-}SET$. For this reason, we consider $fineness(G, O)$ as a first measure of quality of the dataset which is fed into the ML model as a training set.

Let $O = \{o_1,...,o_m\}$ be m raters independently labeling the cases in dataset D; let also assume that each o_i has a constant error rate η_i. Assume that, in order to obtain the Gold Standard (G), for each case x we select the mode (i.e. the label who received the vote of the majority among the o_is) $\bar{o}$: thus, what is the probability that $\bar{o}(x)$ is a false label for x? This amounts to the probability that at least $\frac{m+1}{2}$ raters made an error, this probability can be computed via the *Poisson binomial distribution*:

$$P(error) = \sum_{k=\frac{m+1}{2}}^{m} \sum_{A \in F_k} \Pi_{i \in A} \eta_i \Pi_{j \notin A} (1 - \eta_j) \tag{1}$$

where F_k is the family of sets in which exactly k observers gave the wrong labeling.
Then, the probability to obtain a Gold Standard without errors is:

$$fineness(G, O) = (1 - P(error))^{|G|} \tag{2}$$

An interesting aspect of this is that the fineness of a Gold Standard (and thus the probability of no errors) is exponentially decreasing with the size of the Gold Standard itself. Via the Chernoff bound, and omitting some terms, we can upper bound $P(error)$ as:

$$P(error) \leq e^{-\frac{m+1}{2} log \frac{m+1}{2\mu}} \tag{3}$$

where $\mu = \sum_i \eta_i$, thus the probability of an error decreases exponentially with both *increasing number of raters* and *decreasing expected errors*. By directly inserting this estimate into the bound for PAC learnability given in [1] we obtain that the true (but unknown) target is learnable, with probability $1 - \delta$ over samples and maximum error ϵ, when given at least:

$$\mathcal{O}\left(\frac{d \cdot log\frac{1}{\delta}}{\epsilon(1 - 2e^{-\frac{m+1}{2} log \frac{m+1}{2\mu}})^2}\right) \tag{4}$$

samples whose target is obtained by taking the majority vote as previously specified, where d is the Vapnik-Chervonenkis dimension [38] of the class of models adopted.

The inverse problem of determining the minimum number of raters needed to obtain a certain level of fineness can be solved via the method proposed in [23]. Then, to obtain a desired level of $fineness = 1 - \delta$ for each case $x \in D$ we should involve

$$\mathcal{O}\left(\frac{log\frac{|D|}{\delta}}{(1 - 2\eta_O)^2}\right) \tag{5}$$

raters, where η_O is the average error rate among O.

6 Trueness of the Diamond Standard

Where *fineness* is a propriety of the Gold Standard (with respect to the UR-SET), *trueness* is a propriety of the original Diamond Standard (always with respect to the UR-SET). The total trueness of the Diamond Standard is defined on the basis of the case-wise trueness. Basically, the trueness of a labeling $\langle o_1(x), ..., o_m(x)\rangle$ for a given case x is a measure of how much this labelling could be taken as a representation of the underlying (and unknown) true labeling, that is the corresponding case in the $UR - SET$. In other words, the trueness is the probability that this *diamond* (i.e., multi-facet, multi-rater) labeling actually corresponds to the true one. Basically we would assume that this probability is maximum when $o_1(x) = ... = o_m(x)$, that is, all of the raters agree with each other upon the labeling, while it is minimum when all the possible outcomes are equi–frequent. The trueness of the Diamond Standard as a whole can be computed in various ways starting from the the trueness of its units/cases, among which the simplest way is to take the average trueness for all its cases, with its %95 Confidence Interval (CI).

To quantify the *trueness* of a case x with its diamond labeling $\mathbf{o}(x)$ we propose two approaches. Without loss of generalization, we will focus on the

binary case (i.e. where the target can assume only values $\{0, 1\}$). Let $k \in [0, 1]$ be a threshold value, above which we get vote proportions that can be denoted as an *overwhelming majority* (usually proportions higher than 0.9 or even 0.95 and above according to the application domain); and let p be the observed probability of the majority labeling, taken as a rough estimate of the *trueness*.

Then the 95% confidence interval of p can be computed as

$$trueness'_c(\mathbf{o}(x)) = p \pm 1.96\sqrt{\frac{p(1-p)}{m}} \tag{6}$$

where m is the number of observers, and we say that $\mathbf{o}(x)$ has *acceptable trueness* if $inf(trueness_c(\mathbf{o}(x))) \geq k$.

In the second approach, we know that the maximum number of *disagreements* is $M_d = \frac{m^2-1}{4}$ and we expect the $trueness(\mathbf{o}(x))$ to decrease as the number of observed disagreements approaches M_d.

Thus if O_d is the number of observed disagreements, then

$$trueness''_c(\mathbf{o}(x)) = 1 - \frac{O_d + \epsilon}{M_d + \epsilon} \tag{7}$$

where the ϵ acts as a smoothing factor (avoiding a value of 1 when $O_d = 0$ that could be misleading, since even in that the case there is a non–zero probability that the, unique, Diamond labeling is distinct from the true one). The two approaches have the following properties:

1. With fixed m, $trueness''_c$ has minimum value when $p = \frac{m+1}{2m}$ and maximum value when $p = 1$;
2. With fixed m, $trueness'_c$ has maximum width when $p = \frac{m+1}{2m}$ and minimum width when $p = 1$;
3. Increasing m the width of $trueness'_c$ decreases monotonically, this means that, fixing k and p, it is easier to obtain *acceptable trueness*;
4. If $p \in o(m^2)$ then $lim_{m\to+\infty} trueness''_c(\mathbf{o}(x)) = 1$.

As suggested above, in order to extend this two case-wise definitions of trueness to the Diamond Standard trueness $trueness_D$, we can take different approaches. The most simple approach to extend the $trueness'_c$ definition is to say that the Diamond Standard D has *strong acceptable trueness* if $\forall_{x\in D}$ x has *acceptable trueness*. However, since this criterion of trueness is very restrictive, we can define other Diamond Standard–level measures of trueness, by making two assumptions: for the $trueness''_c$ definition we can assume that the trueness of the cases are distributed as indipendent Bernoullis; for both $trueness'_c$ and $trueness''_c$ we can assume an underlying distribution of the values of p (resp. $trueness_c(\mathbf{o}(x))''$), which could be seen as a distribution of the *difficulty degrees* of assigning the correct labeling to the cases. Under the first approach we obtain the following expression of $trueness^{ind}_D = \Pi_{x\in D} trueness''_c(\mathbf{o}(x))$

Under the second approach we first compute the average proportion (resp. trueness) and we can thus provide an interval estimate, about the average, of the value of trueness in two ways: assuming an underlying model distribution

(with expected value equal to the computed average) and then compute (analytically or numerically) the 95% confidence interval; in a non–parametric way via a bootstrap–based estimate of the confidence interval (i.e. drawing a large number of samples with replacements of the original Diamond Standard and then computing the average proportion of trueness for each of these samples). In both cases we obtain an interval estimate $trueness_D^{int} = [trueness_D^{inf}, trueness_D^{sup}]$ and we say that the Diamond Standard D has *weak acceptable trueness* if $trueness_D^{inf} \geq k$.

It is noteworthy that the concepts of trueness and fineness are, obviously, related with each other: in particular, the greater the trueness of the Diamond Standard, the greater the fineness of the resulting Gold Standard. Most significantly, we could take $trueness_c'$ or $trueness_c''$ as estimates of $1 - P(error)$ to have an approximation of the degree of fineness of the resulting Gold Standard. If we assume that the error rates η_i (i.e., the probability of the annotation of the observer to be in perfect disagreement with the true symbolic representation, cf. accuracy in metrology) are constant for all the cases x, then given that $P(error)$ is also constant, we could simply average p (respectively, $trueness_c''$) over the whole dataset to obtain an estimate $\widetilde{P}(error)$ that we can connect to the fineness bounds obtained in Section 5. Moreover we could also assume that for each case x each observer o_i has a distinct error rate $\eta_i(x)$ (approximated by p or $trueness_c''(x)$), this setting is known as *Constant Partition Classification Noise* (CPCN) which, as shown in [35], is equivalent (in terms of learning complexity) to the setting described in 5.

7 Dryness of the Gold Standard

Dryness regards how much the information content of the Diamond Standard has dried off, or "shrinked", in the *reduction* of this latter into the Gold Standard. The reference is an homage to the seminal idea by Goguen of dry and wet information [17]: the more multiple, collaborative, social, and even ambiguous, the information, the "wetter" (that is "impregnated" with information) it is. Therefore, the higher the *information loss* implied by the reduction, the higher the dryness. Since the reduction implies that the information contained in m columns is reduced in the content of a single column, assessing the dryness of the resulting set can be useful to understand if some reduction is more information-preserving than others, and hence preferable.

In the following we will assume a nominal valued target, thus the target of the Diamond Standard is expressed in terms of a m–dimensional vector over a set Y (i.e. $\mathbf{o}(x) \in Y^m$) and we suppose that the target of the Gold Standard is generated from $\mathbf{o}(x)$ via a reduction $T : Y^m \mapsto \mathcal{C}(Y)$ where $\mathcal{C}(Y)$ is a set of structures, in a general sense, over Y (e.g. the set of probability distributions over Y). In general, the reduction T involves an information loss (or an increase in dryness) given by the fact that only observing $T(\mathbf{o}(x))$ it is impossible to (perfectly) recover $\mathbf{o}(x)$ (assuming that $\mathcal{C}(Y) \neq Y^m$ and $T \neq id_{Y^m}$); this means

that T implicitly defines an inverse set–valued map $L : \mathcal{C}(Y) \mapsto \mathcal{P}(Y^m)$ allowing us to define a measure of dryness in both *quantitative* and *qualitative* terms.

In quantitative terms we will define the dryness of $T(\mathbf{o}(x))$ as:

$$dryness(\mathbf{o}(x), T) = \frac{|L(T(\mathbf{o}(x)))| - 1}{|Y|^m - 1} \tag{8}$$

which can be understood as the ratio of the *information contents* of $\mathbf{o}(x)$, in the denumerator, and $T(\mathbf{o}(x))$, in the numerator: in particular, the numerator is the number of objects in $|Y|^m$ satisfying the constraints imposed by L. From the values of $dryness(\mathbf{o}(x))$, for each case x, we can obtain the value of the dryness, under reduction T, for the whole Gold Standard as:

$$dryness(G, T) = \frac{1}{|G|} \sum_{x \in G} dryness(\mathbf{o}(x)) \tag{9}$$

Usually, in the nominal case, the reduction T is taken as the mode, that is $T(\mathbf{o}(x)) = mode(\mathbf{o}(x))$; in this case the numerator is given by all possible diamond labelings in which $mode(\mathbf{o}(x))$ is in fact the most frequent label, which can be approximated via the following bound:

$$dryness(\mathbf{o}(x), mode) = O(\frac{\sum_{\pi} \sum_{\pi_{|Y|} \leq \ldots \leq m^*} \binom{m}{m^*, \ldots, \pi_{|Y|}}}{|Y|^m - 1}) \tag{10}$$

where π is any assignment of $m - 1$ least frequent classes, π_i is the i-th least frequeny class in the assignment π and m^* is the frequency of the mode. However, other reductions could be defined, a first such example is the transformation $freq$ defined as:

$$freq(\mathbf{o}(x)) = \langle \frac{m_1}{m}, ..., \frac{m_{|Y|}}{m} \rangle \tag{11}$$

where m_i is the frequency of class $c_i \in Y$ in $\mathbf{o}(x)$. The dryness of $freq$ is defined as:

$$dryness(\mathbf{o}(x), freq) = \frac{\binom{m}{m_1, \ldots, m_{|Y|}} - 1}{|Y|^m - 1} \tag{12}$$

in which the numerator is given exactly by the number of diamond labelings in which the labels occur with exactly the frequency given by $freq(\mathbf{o}(x))$. Evidently, $dryness(\mathbf{o}(x), freq) \leq dryness(\mathbf{o}(x), mode)$ and, $freq$ is the reduction with minimal dryness among the ones that are order–irrelevant. However, besides the quantitative part of the dryness, there is also a qualitative part: each reduction defines which information is deemed relevant (and thus conserved), and which information is instead discarded. The *mode* reduction maintains only the most frequent label and discards every other information; on the other hand, the $freq$ reduction keeps the proportions of each possible alternative and only "forgets" the order–part of the vector (i.e. which option each observer selected).

Another qualitative aspect of the dryness is given by the fact that we can provide two different interpretations of each reduction T:

1. The *epistemic* view, according to which we suppose that the true labeling of x in the UR–SET is a single label in Y and $T(\mathbf{o}(x))$ represents our degree of belief assigned to the alternatives for that label (e.g. $freq$ represents our subjective posterior probability of which it is the real labeling);
2. The *ontic* view, according to which we suppose that in fact the true labeling of x in the UR-SET is not a label from Y but one from $im(T) = \mathcal{C}(Y)$ and reduction T allows us to estimate this label from the information given by $\mathbf{o}(x)$ (e.g. the ontic view associated with the $freq$ reduction is that our phenomenon is indeed a non–deterministic one and $freq(\mathbf{o}(x))$ is an estimation of the propensities of the system to be in one of the alternative states).

If we look at the quantitative component of the dryness, the $freq$ reduction is manifestly the optimal choice to construct the Gold Standard. However, the qualitative approach suggests that it may retain "too much" information: the exact proportions may be observed only "by accident" or they could be irrelevant. In the following, we will suggest two alternative reductions that are mid–way between *mode* and $freq$ in terms of dryness.

7.1 Fuzzy–Possibilistic Reductions

Let $m^* = max_i(m_i)$ be the index of the most frequent labels in $\mathbf{o}(x)$, then we define the *possibilistic reduction* as:

$$poss(\mathbf{o}(x)) = \langle \frac{m_1}{m^*}, ..., \frac{m_{|Y|}}{m^*} \rangle \tag{13}$$

Under the qualitative point of view, the *poss* reduction preserves the preference ordering among the possible alternatives and also a "relative" indication of degrees of preference of an alternative compared to the others: thus, the numerator of the dryness is the number of diamond labelings in which the proportions between the most frequent label and the other ones are determined by m and the values of $poss(\mathbf{o}(x))$. If we denote by σ the ordering of the labels in Y in order of decreasing value of *poss*, then we can bound the dryness as:

$$dryness(\mathbf{o}(x), poss) = o(\frac{\sum_{m_{\sigma_1}=\frac{1}{\rho_{|Y|}}}^{m} \binom{m}{m_{\sigma_1}, m_{\sigma_1} \cdot \rho_2, ..., m_{\sigma_1} \cdot \rho_{|Y|}} - 1}{|Y|^m - 1}) \tag{14}$$

Under the *epistemic* interpretation, the *poss* reduction models our degree of belief in terms of a *possibility distribution* [42], which could be taken as representing an *imprecise probability distribution* [12] representing our belief in the relative preferences and their proportions but not the exact counts.

Under the *ontic* intepretation, on the other hand, the *poss* reduction represents a *fuzzy set*, that is, we assume that the different labelings given by the observers are not due to errors but due to the fact that the phenomenon itself is multi–faceted and, in some sense, *vaguely defined* and the labelings reported more frequently are *prototypical* for the observed instance of the phenomenon.

7.2 Three–Way Reduction

Three–way decision theory [41] refers to an extension of standard decision–theory in which the "decision maker" (in a general sense, including also an algorithm) has the ability to abstain (totally or partially) instead of expressing a decision.

We will describe two approaches to perform a three-way transformation. Let $\epsilon \in [0,1]$, $freq(\mathbf{o}(x))$ be the frequencies of the labels in Y and σ the ordering of the labels in decreasing frequency order. Then we say that $\mathbf{o}(x)$ is $(m,\epsilon) - ambiguous$ if

$$\forall i \in \{1,...,m\}.|\sigma_1 - \sigma_i| \leq \epsilon \tag{15}$$

Let m^* be the greatest m such that $\mathbf{o}(x)$ is $(m,\epsilon) - ambiguous$, then we define the tw_a reduction as:

$$tw_a(\mathbf{o}(x),\epsilon) = \{\sigma_1,...,\sigma_{m^*}\} \tag{16}$$

In this case the numerator of equation (8) is given by the number of diamond labelings for which the labels in $tw_a(\mathbf{o}(x),\epsilon)$ are the most frequent ones and their distance is at most ϵ.

The second approach, that we term *decision–cost theoretic*, descends from our previous work on three–way classification [9,10]. Let ϵ be an error cost, α be an abstention cost and $freq(\mathbf{o}(x))$, σ defined as above. Then we define the tw_d reduction as:

$$tw_d(\mathbf{o}(x),\epsilon,\alpha) = \begin{cases} \{\sigma_1,...,\sigma_j\} & \alpha \cdot \sum_{i=1}^{j} \sigma_i + \epsilon \cdot \sum_{i=j+1}^{k} \sigma_i < \epsilon * (1-\sigma_1) \\ \sigma_1 & \text{the inequality has no solution} \end{cases} \tag{17}$$

where j is the optimal index satisfying the inequality.

Future work will be devoted to understand how knowledge about the raters' skills, and confidence (even self-perceived) in the raters' interpretation, can be integrated in the reduction to make the Gold Standard finer (and reduce the information loss in the transformation from the Diamond Standard).

Example 1 *Let D be a Diamond Standard of 3 cases and*

$$\mathbf{o}(D) = \begin{bmatrix} 0\ 1\ 0\ 1\ 0 \\ 1\ 0\ 1\ 1\ 1 \\ 1\ 0\ 0\ 0\ 0 \end{bmatrix}$$

the respective labeling given by 5 observers.

Applying the mode reduction we obtain, $mode(\mathbf{o}(D)) = \begin{bmatrix} 0\ 1\ 1 \end{bmatrix}$ for which the dryness $dryness(\mathbf{o}(D), mode) = \begin{bmatrix} 15/31\ 15/31\ 15/31 \end{bmatrix}$. The total dryness of G (reduced from D in this way) is then the average, 0.48.

On the other hand, for transformation $freq$ we obtain

$$freq(\mathbf{o}(D)) = \begin{bmatrix} (0:3/5, 1:2/5)\ (0:1/5, 1:4/5)\ (0:4/5, 1:1/5) \end{bmatrix}$$

for which the dryness is $dryness(\mathbf{o}(D), freq) = \begin{bmatrix} 10/31\ 5/31\ 5/31 \end{bmatrix}$. The total dryness of G (reduced from D in this second way) is then the average, 0.22.

For the tw_a reduction, setting $\epsilon = 0.4$ we have that $tw_a(\mathbf{o}(D), \epsilon) = \big[\{0,1\}\ 1\ 0\big]$ for which the dryness is $dryness(\mathbf{o}(D), tw_a) = \big[19/31\ 6/31\ 6/31\big]$. The total dryness of G (reduced from D in this third way) is then the average, 0.33.

Finally for the poss reduction we have that

$$poss(\mathbf{o}(D)) = \big[(0:1, 1:2/3)\ (0:1/4, 1:1)\ (0:1, 1:1/4)\big]$$

for which the dryness is $dryness(\mathbf{o}(D), poss) = \big[10/31\ 5/31\ 5/31\big]$. Thus, the total dryness of G (reduced from D in this last way) is then the average, 0.22. Summarizing, when computing the values of $dryness(G)$ obtained with different reductions, we get that:

$$dryness(D, poss) = dryness(D, freq) < dryness(D, tw_a) < dryness(D, mode)$$

We remind that the higher the dryness, the higher the information loss, and hence the informatively "poorer" the Gold Standard.

8 Some more idle reflections

Explainable AI (XAI) has recently been set forth as a necessary component of human agencies where decision making is supported by computational means[7].

Apart from a "XAI paradox", which we will mention at the end of this contribution, we agree that some form of XAI is necessary for human decision makers who use some kind of AI decision support to reach more informed decisions and be rightly held fully accountable for these decisions.

In the context of the XAI discourse, human decision makers must be able *to interpret* the AI system output, that is make sense of it in terms of *why* the system proposed a specific output for the provided input [24] and, to some extent, of *how* the system yielded this output, so as to take its advice into due consideration in making their decision. In this line, *interpretability* is often tightly related to *explainability* (so much that these two terms are often used interchangeably) and both are usually articulated in terms of the capability of the AI system "to explain its reasoning" [11]. Thus, the lack of a formal, or at least unique and non-ambiguous definition of explanation (and hence explainability), which is lamented by many observers (e.g., [31,11]), should not make us overlook the fact that the ability to interpret the system behavior by the humans, so that they can make an informed use of the system output, is often translated into *a property of the system*, that is its capability to provide human decision makers with resolving clues about its functioning and "reasons" for a prediction. However, while this property can be linked to the presence or absence of specific functions that make some information available to the decision makers (e.g., what aspects of the phenomenon at hand, i.e., predictor variables, were

[7] To this respect, here we are covering different cases than those covered by the GDPR article no. 22, which regards decisions that are solely based on automated processing, without human intervention [39].

more important for proposing a specific advice), self-explanations tell nothing about their suitability of being understood and hence of their potential to contribute to the interpretation of the system. This allows us to relate the notion of interpretability/explainability to the notion of *usability* of the system. A focus on usability suggests to assess AI not only in regard to task efficiency (e.g., time to completion) and effectiveness (e.g., error rate) but also in terms of user satisfaction. In the context of human decision making this regards the extent decision makers are satisfied by their interaction with the system; feel to be in control of the situation; believe to have got a sufficient number of indications to formulate an *informed decision*; feel to be able to account for it; are confident that the system supported them in considering all of the aspects that were due; and that it did not misled them. However, usability, as widely known, is not a property of the system, but rather of the coupling between the system and the human users; in other words, usability emerges in the interaction between the AI and its users, in the fit between system functionalities and the user skills. So does the XAI. In the light of seeing interpretability as a kind of usability (or better yet, as a way in which the usability of AI-driven decision support is manifested), we also advocate an *interpretable* and *explainable* AI [18] as a necessary condition for the embedding of AI in human agencies that are called to make critical decisions significantly affecting other people's life. Even more than this, we emphasize the importance to design for an *interpretand AI*, that is an *AI that must be interpreted* by the decision makers, so that that they build a *local narrative* to convince themselves, as well as the others, of the soundness and *reasonability* of the resulting decision. Thus, in the human-AI interaction, it is important to distinguish between a *right to explanation*, that is for the users to receive indications by the AI system that satisfactorily bring them to believe to have understood why the decision support gave them a certain advice; and the *obligation to interpretation*, that is for the users to have to adopt an active attitude to collect and interpret these indications: advocacy for explainable AI should not diminish responsibility for decision makers. This duty to active interpretation can be promoted, and even afforded, by the decision support system itself: to this aim we are testing a decision support system that is currently adopted in a large teaching hospital specialized in musculoskeletal disorders and surgery and is endowed with *programmed inefficiencies*, that is features aimed at purposely increase the "decision friction" (cf. [13]), by requiring an active stance by the users so as to minimize the risk of automation bias and deskilling [8].

9 Conclusion

In this paper, we focused on the importance of letting the decision makers know and understand the quality of the data used to train the models by which an AI can provide its predictions and advice. In fact, no model can bring meaningful output if the input data are not reliable: the notorious phrase "Garbage In, Garbage Out" here applies, and is the central tenet of our contribution, as the tongue-in-cheek title suggests.

To make the AI system more transparent, we propose to focus on the ground truth by which the AI has been trained. To make the ground truth more interpretable, we proposed a framework that distinguishes between Gold Standards and Diamond Standards, and encompasses some common (but relevant) quality dimensions, like representativeness and reliability, and some novel quality dimensions, like *fineness*, *trueness* and *dryness*, which we discuss and for which provide a preliminary yet formal specification.

These metrics are given for a twofold aim. First, their definition and application invite AI researchers to devise alternative ways to produce the ground truth from the observations and interpretations available (what we call alternative *reductions*), other than the simple majority vote, so that the quality of the training set could improve along multiple dimensions. However, this is still a technicality, although of no little importance. More importantly: since we usually *assume* that our ground truth is perfect, reflecting on its quality necessarily entails growing an informed *prudence* in regard to its reliability and adequacy for the task of supporting decision making in delicate domains. Thus, our ultimate main aim is to contribute to raising awareness of the impact of our assumptions, models, and representations in intensive cognitive tasks. The dimensions we started to envision are aimed at facilitating people to reflect on these aspects, rather than focus on model details and misleading performance metrics, like accuracy, which only regards the match between the AI predictions and the Gold Standard (see Figure 1). From the design point of view, we should ask what an actually *useful* support from AI looks like. We hold that a useful AI is a *usable* AI, but not necessarily an AI providing decision makers with simple and clear-cut predictions, nor the system that combines its output with a plenty of indications and explanations. In the light of the research on the use of *computers as persuasive technologies* [2] (evocatively called *captology* by Fogg [16]), we should be aware of a potential conundrum on effective XAI, what we could call a *captological XAI paradox*: "AI can give us a wrong advice, and yet also in that case accompany it with plausible reasons that *prime* our interpretation and convince us. The more imperscrutable AI is, the more likely we can doubt it, and make sense of the available data with less interference". Obviously, awareness of this paradox should not convince us to stop pursuing a better XAI. All the opposite, it urges us to consider new and more effective ways by which technology itself can promote a reflective stance in the decision makers and a stronger will and commitment to take full responsibility of the vigilant use of that technology.

References

1. Angluin, D., Laird, P.: Learning from noisy examples. Machine Learning 2(4), 343–370 (1988)
2. Atkinson, B.M.: Captology: A critical review. In: International conference on persuasive technology. pp. 171–182. Springer (2006)
3. Brennan, P., Silman, A.: Statistical methods for assessing observer variability in clinical measures. BMJ: British Medical Journal 304(6840), 1491 (1992)

4. Breunig, M.M., Kriegel, H.P., Ng, R.T. et al.: Identifying density-based local outliers. SIGMOD Rec. 29(2), 93–104 (2000)
5. Cabitza, F., Ciucci, D., Rasoini, R.: A giant with feet of clay: on the validity of the data that feed machine learning in medicine. In: Organizing for the Digital World, pp. 121–136. Springer (2019)
6. Cabitza, F., Dui, L.G., Banfi, G.: Pros in the wild: Assessing the validity of patient reported outcomes in an electronic registry. Computer methods and programs in biomedicine (2019)
7. Cabitza, F., Locoro, A., Alderighi, C., Rasoini, R., Compagnone, D., Berjano, P.: The elephant in the record: on the multiplicity of data recording work. Health informatics journal, SAGE Publications Sage UK: London, England (2019)
8. Cabitza F., Campagner A., Ciucci D., Seveso A.: Programmed Inefficiencies in DSS-supported Human Decision Making. To appear, Proceedings of 16th MDAI International Conference (2019)
9. Campagner, A., Cabitza, F., Ciucci, D.: Exploring Medical Data Classification with Three-Way Decision Trees. Proceedings of the 12th BIOSTEC International Joint Conference - Volume 5: HEALTHINF, 147-158 (2019)
10. Campagner A., Cabitza F., Ciucci D.: Three–Way Classification: Ambiguity and Abstention in Machine Learning. Rough Sets - International Joint Conference, IJCRS 2019. LNCS, vol 11499, Springer, 280-294 (2019)
11. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
12. Dubois, D., Prade, H.: Possibility Theory and Its Applications: Where Do We Stand?, pp. 31–60. Springer, Berlin, Heidelberg (2015)
13. Edwards, P.N., Mayernik, M.S., Batcheller, A.L., et al.: Science friction: Data, metadata, and collaboration. Social Studies of Science 41(5), 667–690 (2011)
14. Esteva, A., Kuprel, B., Novoa, R.A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639), 115 (2017)
15. Feinstein, A.R., Cicchetti, D.V.: High agreement but low kappa: I. the problems of two paradoxes. Journal of clinical epidemiology 43(6), 543–549 (1990)
16. Fogg, B.J.: Persuasive computers: perspectives and research directions. In: CHI'98, 225–232. ACM Press (1998)
17. Goguen, J.: The dry and the wet. In: Proceedings of the IFIP TC8/WG8.1 Working Conference on Information System Concepts: Improving the Understanding, 1–17 (1992)
18. Goebel, R., Chander, A., Holzinger, K., et al: Explainable AI: the new 42?. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction. pp. 295-303. Springer (2018)
19. Gulshan, V., Peng, L., Coram, M., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama 316(22), 2402–2410 (2016)
20. Gur, D., Bandos, A.I., Cohen, C.S., et al.: The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. Radiology 249(1), 47–53 (2008)
21. Haenssle, H., Fink, C., Schneiderbauer, R., et al.: Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology 29(8), 1836–1842 (2018)
22. Han, S.S., Park, G.H., Lim, W., et al.: Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis. PloS one 13(1), e0191493 (2018)

23. Heinecke, S., Reyzin, L.: Crowdsourced pac learning under classification noise, arXiv preprint arXiv:1902.04629 (2019)
24. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Mueller, H.: Causability and Explainability of AI in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9(4), (2019)
25. Jiang, H., Nachum, O.: Identifying and correcting label bias in machine learning. arXiv preprint arXiv:1901.04966 (2019)
26. Justel, A., Peña, D., Zamar, R.: A multivariate kolmogorov-smirnov test of goodness of fit. Statistics & Probability Letters 35(3), 251 – 259 (1997)
27. Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A.: Interpreting and unifying outlier scores. pp. 13–24 (2011)
28. Krippendorff, K.: Content analysis: An introduction to its methodology. Sage publications (2018)
29. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Interpretable & explorable approximations of black box models. arXiv preprint arXiv:1707.01154 (2017)
30. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977)
31. Lipton, Z.C.: The mythos of model interpretability. arXiv preprint arXiv:1606.03490 (2016)
32. MacKay, D.J.C.: Bayesian methods for adaptive models. Phd thesis, California Institute of Technology (1992)
33. Popović, Z.B., Thomas, J.D.: Assessing observer variability: a user's guide. Cardiovascular diagnosis and therapy 7(3), 317 (2017)
34. Quarfoot, D., Levine, R.A.: How robust are multirater interrater reliability indices to changes in frequency distribution? The American Statistician 70(4), 373–384 (2016)
35. Ralaivola, L., Denis, F., Magnan, C.N.: Cn = cpcn. In: ICML '06, ACM, (2006)
36. Stand, J.: The hawthorne effect - what did the original Hawthorne studies actually show. Scand J Work Environ Health 26(4), 363–367 (2000)
37. Svensson, C.M., Krusekopf, S., Lücke, J. et al.: Automated detection of circulating tumor cells with naive bayesian classifiers. Cytometry Part A 85(6), 501–511 (2014)
38. N. Vapnik, V., Ya. Chervonenkis, A.: On the uniform convergence of relative frequencies of events to their probabilities. Theoretical Probability and its Applicactions 17, 264–280 (1971)
39. Wachter, S., Mittelstadt, B., Floridi, L.: Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law 7(2), 76–99 (2017)
40. Wishart, D.: k-means clustering with outlier detection, mixed variables and missing values, in: M. Schwaiger, O. Opitz (Eds.), Exploratory Data Analysis in Empirical Research, Springer Berlin Heidelberg, 216–226 (2003)
41. Yao, Y.: An outline of a theory of three-way decisions. Lecture Notes in Computer Science, vol. 7413. Springer Berlin (2012)
42. Zadeh, L.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems 100, 9 – 34 (1999)