The Relevance of Non-Human Errors in Machine Learning

Ricardo Baeza-Yates[2,1,*], Marina Estévez-Almenzar[1,*]

[1]DTIC, Pompeu Fabra University, Barcelona, Spain
[2]Institute for Experiential AI, Northeastern University, USA

Abstract

The current practice of focusing the evaluation of a machine learning model on the accuracy of validation has been lately questioned, and has been declared as a systematic habit that is ignoring some important aspects when developing a possible solution to a problem. This lack of diversity in evaluation procedures reinforces the difference between human and machine perception on the relevance of data features, and reinforces the lack of alignment between the fidelity of current benchmarks and human-centered tasks. Hence, we argue that there is an urgent need to start paying more attention to the search for metrics that, given a task, take into account the most humanly relevant aspects. We propose to base this search on the errors made by the machine and the consequent risks involved in moving human logic away from that of the machine. If we work on identifying these errors and organize them hierarchically according to this logic, we can use this information to provide a reliable evaluation of machine learning models, and improve the alignment between training processes and the different considerations humans make when solving a problem and analyzing outcomes. In this context we define the concept of non-human errors, exemplifying it with an image classification task and discussing its implications.

Keywords

Machine Learning, Responsible AI, Evaluation, Error Analysis, Non-Human Errors

1. Introduction

Imagine that you enter a skyscraper and the elevator has a sign that says: "Works 99% of the time". Would you take the elevator? Most people would not. However, if the sign says "Does not work 1% of the time and when that happens, stops", you probably would use it, because you perceive that you will be safe thanks to the explanation of the error and the possibility to evaluate the consequences: "The elevator may fail, but when it does, the fail consists of stopping". Today, Machine Learning (ML) models are evaluated primarily on the basis of success rather than failure. Worse, this evaluation does not take into account the potential harm of its mistakes, like is done in the pharmaceutical or the food industry.

Along the same lines as this example, the current benchmarks fidelity to human-centered tasks has recently been called into question [1, 2, 3]. The practice of centering the model evaluation on the validation accuracy has been stated as a dangerous habit [4, 5] that is ignoring some important aspects of the human perception when developing a solution for a problem, such as carefully studying the risks of the solution and its different points of operation. This lack of diversity in evaluation procedures reinforces the difference between human and machine perception of the relevance of data features [6]. In fact, in many cases, the best operation point of a model is not the one of maximal accuracy.

We can state that the benchmark-task misalignment can be directly explained by the misalignment between human and machine perceptual mechanisms, and we propose a simple taxonomy to bridge those differences that are potentially harmful to humans, in order to achieve more reliable model training and evaluation procedures, even if it implies a decrease of the validation accuracy.

In order to do that, it is essential to drive a decentralization of the evaluation process in ML models, which is mainly focused on maximizing accuracy without paying attention to other parameters that could be of great relevance. Hence, our main objective will be to highlight the need to define new methodologies and metrics that represent the mechanisms of human perception in a more realistic way. Obviously, these metrics do not have to fully represent human perception but should, at least, cover the most humanly relevant aspects of the task at hand. We propose to base the search of these metrics on the different types of errors done by the model. More specifically, we want to focus on how they differ from those errors that a human might make. If we work on identifying these errors and organize them hierarchically according to these differences, we can use this information to provide a more meaningful evaluation of ML models, and improve the alignment between ML training processes and the different considerations humans make when solving a problem.

Therefore, after discussing the state of the art in Section 2, we introduce the concept of *non-human errors* in Section 3. This concept allow us to build our error tax-

EBeM'22: Workshop on AI Evaluation Beyond Metrics, July 25, 2022, Vienna, Austria
rbaeza@acm.org (R. Baeza-Yates); marina.estevez@upf.edu (M. Estévez-Almenzar)
https://https://www.baeza.cl/ (R. Baeza-Yates); https://ealmenzar.github.io/ (M. Estévez-Almenzar)
*The authors contributed equally to this work as first authors.

onomy. Then we use an image classification task to do a proof of concept that uses our taxonomy in Section 4, ending with a discussion of its consequences in Section 5. A simple notebook to illustrate this work is available in Github.[1]

2. Related Work

The practice of centering the model evaluation on accuracy has been and still is being questioned. For example, [6] warns that the only-use of accuracy to measure machine performance works as a limitation to humans when analyzing machines, and states that adversarial vulnerability results from the susceptibility of models to data features that are potentially candidates for generalization. They recall that the fact that we train machines to solely maximize accuracy is making its learning system use any available signal to achieve this goal, even those signals that look incomprehensible to humans.

This idea is also supported by [1], who states that training a model in a robust way leads to a reduction of accuracy. They argue that this trade-off between the accuracy of a model and its robustness to adversarial perturbations is a consequence of robust classifiers learning fundamentally different feature representations than standard classifiers. These differences, in particular, seem to result in unexpected benefits: the representations learned by robust models tend to align better with salient data characteristics and human perception. Other trade-offs of this kind were recently addressed for language models and their intrinsic risks in [3], where authors state that researchers are extending the state of the art on a wide array of tasks as measured by classification scores on some benchmarks, following the methodology of using some pre-trained models and then fine-tuning them for specific tasks. In this scenario, they take a step back and pay attention to the possible risks associated with this technology in terms of environmental and financial costs.

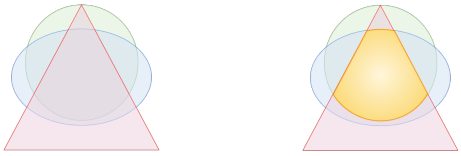
Another research that calls into question the results obtained with current state-of-the-art benchmarks was done by [4], that highlights the fact that some data sets contain errors in their labels, and they expose a subsequent study about the potential for these label errors to affect benchmark results. Surprisingly, they find that lower capacity models may be practically more useful than higher capacity models in real-world data sets with high proportions of erroneously labeled data. They conclude that ML practitioners must be careful when choosing which model to deploy based on validation or test accuracy.

In an attempt to overcome the limitation that entails the only-use of accuracy in ML evaluation, other metrics have been proposed. For example, [5] points out that accuracy alone cannot distinguish between strategies. Two systems – brains or algorithms – may achieve similar accuracy with very different strategies. In their study, they conclude that the consistency between human errors and errors made by deep learning models is not far away from what can be expected by chance alone, indicating that machines still employ very different perceptual mechanisms. There are also some benchmarks that approach decentralisation with respect to accuracy. In [7] the authors propose to study the impact of errors and compare them according to their type, but both the proposed error classification they offer and the associated impact focus only on pose estimation algorithms.

In [8], the author also proposes to pay particular attention to the analysis of error from a quantitative perspective. He proposes to focus this analysis on those errors whose correction has the greatest impact in terms of improving the accuracy of the algorithm. Even though this analysis is fundamental, we propose to focus on improving our models in qualitative terms. Given a context, if a type of error is sufficiently serious, illogical or risky, it does not matter if it is made infrequently: we should work on minimising this type of error in order to minimise the possible harmful consequences. Another important discussion that the author mentions is how to define human-level performance in order to compare it with the performance of a machine. This highlights the importance of considering the context in which the machine learning model is applied. The human-level performance we choose to consider will depend on the task itself, or the harm risks it may pose.

To solve the problem of the misalignment between state-of-the-art benchmarks and human-centered tasks, some of the works mentioned above propose as the main solution either to correct the labels in data sets or redefine the way we store and represent these labels. Even when these corrections are essential, using correctly labeled or redefined labels to evaluate models could still not be sufficient to cover the diversity of human perception. While working on improving the representations associated with the inputs in ML systems is very necessary, we need to do a similar effort on improving the way we interpret and analyze the outputs. These outputs are mainly characterized by two elements: successes and failures. Until now, ML evaluation metrics have been mainly based on successes, giving visibility to the accuracy of the algorithm over other possible ways of measuring the overall performance of the algorithm. We propose to change the focus and start prioritizing the analysis of errors, as well as their classification according to the potential damage they may cause in the context in which they occur.

[1] https://github.com/ealmenzar/non-human-errors



(a) (b)

Figure 1: Visual representation of human and ML performances. On the left, the red triangle represents the ML model, the blue ellipse represents the human, and the green sphere represents the ground truth. They are positioned in the solution space of a binary prediction problem. For every figure, *positive* answers are inside, and *negative* answers are outside, being the correct answers determined by the green sphere. On the right we can see the yellow region representing the correct answers obtained by both the human and the ML model.

3. Non-Human Errors

Finding new methodologies and metrics that allow us to exploit the valuable information in errors done by machines is not trivial. To illustrate in a simplified way the error exploration that we are proposing, let us consider a problem with a binary solution space. In this space, we are given the ground truth, so we are capable of determining whether a point from the space is a correct or an incorrect answer, as well as its impact. For example, a typical assumption when solving a binary classification problem, is to consider that false negatives have the same weight of false positives. This is not always correct, because their harm might be quite different. Indeed, when predicting an illness, a physician will prefer to see many more healthy patients just to avoid missing any ill one. One solution is to use a weighted accuracy but still the operational point might be different because here the recall of ill patients is much more relevant than the overall accuracy (weighted or not).

In Figure 1 we can see this ground truth represented as a green sphere, positioned on the solution space, such that those points that fall into the green area are the true positive answers, and the rest of the points are the true negative answers. We can see two more shapes in this space that symbolize the perceptual agents; a red triangle representing the machine predictions, and a blue ellipse representing the human predictions. Following the previous logic, in Figure 1b we can see the true answers correctly predicted by both the model and a human. In Figure 2, we focus on the errors. Here we are able to distinguish between two kinds of errors: false positives and false negatives. And we make another distinction based on the entity that is making the error (human and/or ML model).

In this general and abstract scenario, where no concrete use case is specified, we wonder whether we can determine which errors are the most relevant in terms of human harm risk. Although harm risk may be perceived very differently depending on the performer (human or machine), it is clear that we are interested in avoiding harmful consequences for the humans involved, directly or indirectly, in the task at hand. It is reasonable to think that the errors related to these consequences are those that are unexpected and atypical for humans, and therefore those that are difficult for us to explain and control. Since, as humans, we are accustomed to human errors, we might expect that those errors that are furthest away from the errors that a human might make could be considered risky: we refer to these types of disparate errors as *non-human* errors (see Figure 3).

We can also formalise this idea in terms of mathematical sets. This will help us to formally define the different types of errors mentioned and graphically expressed above. We denote S as the green sphere, T as the red triangle, and E as the blue ellipse (see in Figure 1a). Following the logic explained above, we could consider these sets of points in the solution space (and their complementary sets, noted as $\overline{S}$, $\overline{T}$, and $\overline{E}$ respectively) as follows:

$S \equiv$ true positives

$T \equiv$ positives predicted by the model

$E \equiv$ positives predicted by the human

$\overline{S} \equiv$ true negatives

$\overline{T} \equiv$ negatives predicted by the machine

$\overline{E} \equiv$ negatives predicted by the human

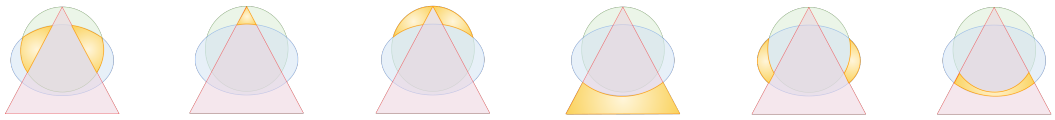
Focusing on the errors shown in Figure 2, now we denote the false positives errors made only by the machine as P_m (Figure 2d), and we define this set of errors as

$$P_m = T \cap \overline{E} \cap \overline{S}$$

Similarly, we denote and define the false positives errors made only by the human (Figure 2e), the false positive errors made by both the machine and the human (Figure 2f), the false negative errors made only by the machine (Figure 2d), the false negative errors made only by the human (Figure 2b), and the false negative errors made by both the machine and the human (Figure 2c) as follows, respectively:

$$P_h = \overline{T} \cap E \cap \overline{S}$$
$$P_b = T \cap E \cap \overline{S}$$
$$N_m = \overline{T} \cap E \cap S$$
$$N_h = T \cap \overline{E} \cap S$$
$$N_b = \overline{T} \cap \overline{E} \cap S$$

Note that all these sets are disjoint because of the exclusivity imposed when considering which agent commits



(a) (b) (c) (d) (e) (f)

Figure 2: Visual representation of false negatives and false positives attributed to the model, the human, or to both. The first three diagrams ((a), (b) and (c)) represent false negatives errors done by only the model, only the human, or by both, respectively. The next three diagrams ((d), (e) and (f)) represent the false positives errors done by only the model, only the human, or by both, respectively. Error areas are overstated to emphasize the idea.

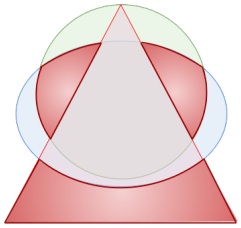


Figure 3: Non-human errors stressed in red: both false negatives and false positives done by the ML model but not by humans (cases (a) and (d) in Figure 2).

the error. Now the sets of interest arise from the union of some of the previous sets. We note M as the non-human errors (those committed by the machine but not by the human) explained above, H as those errors committed by the human but not by the machine, and B as those errors committed by both the human and the machine together:

$$M = N_m \cup P_m$$
$$H = N_h \cup P_h$$
$$B = N_b \cup P_b$$

In this paper we focus on M, non-human errors, which we believe are the errors that we should address first because of the harm risks that could be involved in making errors that escape human logic. But how can we precisely determine these errors? How can we measure how far an answer should be from human logic in order to call it a non-human error? We address these challenges in the next section.

4. Proof of Concept

Approaching a problem by adopting the previous abstract perspective allows us to visualize it with some independence from the use case or real-world application, which is good for understanding the wide range of operating points. The distinction of non-human errors is based on the distinction between the successes and mistakes made by the different perceptual agents (human and machine). Also, this distinction only makes sense in a context in which we can expect reasonable human performance. Thus, the category of non-human errors can be found in those human centered tasks that can be at least partially solved in a reasonable way by the humans and where a ML algorithm is applied instead. However, in practice, consideration of the specific use case will be decisive.

We next apply this idea to a simple but illustrative problem: classifying images of dogs and cats according to their breed [9]. This translates into a fine-grained image classification problem that is mainly solved by using deep neural networks. Based on expert sources in the classification of these animals (FIFe and FCI Federations), we have been able to construct a taxonomy that represents the possible errors that can be made in this task. Following our definition of non-human errors, in this problem we can identify them as those errors that are fundamentally different from the errors that a human solving this task would commit. Therefore, we define as non-human errors those cases in which the machine classifies a dog as a cat, or vice versa (see Figure 4). Notice that there might be other non-human errors when comparing among only cats or dogs, but those are much less important and less common than the definition that we use for this proof of concept and provides a lower bound for non-human errors.

So far, we have selected one of the top-ranked algorithms for solving this specific task, the Big Transfer (BiT) model from [10], which achieved 93% of accuracy. In Figure 6 we give the full confusion matrix of 3,312 prediction pairs among 25 dog breeds (top-left) and 12 cat breeds (bottom-right), where we can see that there are 4 pairs that are hard to classify (two breeds of Terriers and 3 pairs of cat breeds). Notice that this confusion matrix is in general non-symmetric, as the output of the model may differ because the input and the prediction for each pair is different.

Here we found that more than 3% of the errors were non-human errors (8 of 241 errors), which appear as light

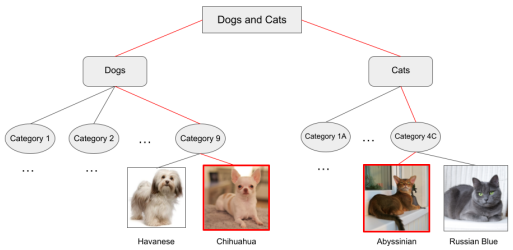


Figure 4: Part of the error taxonomy obtained from the analysis made with Oxford-IIIT Pets data set [9]. In red, one of the most common non-human errors committed by the BiT model [10].

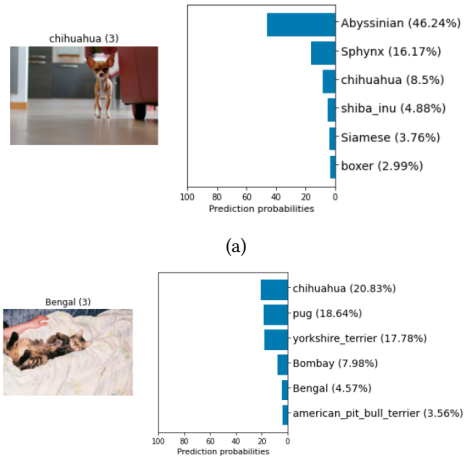


Figure 5: Two of the non-human errors obtained when running the BiT model [10] over the Oxford-IIIT Pets data set [9]. In (a) a Chihuahua is mistaken for an Abyssinian cat with a confidence of 46.24%. In (b) a Bengal cat is mistaken for a Chihuahua with a confidence of 20.83%, a percentage very close to the one of the second option in the list of breeds sorted by their probability of being selected as the tag for that image.

squares in the top-right and bottom-left of Figure 6. Two of these errors are shown in Figure 5, where a Chihuahua is classified as an Abyssinian cat (Figure 5a) and a Bengal cat is classified as a Chihuahua (Figure 5b). However, there is a notable difference between these two errors: the certainty of the answer provided by the algorithm. This supports the need to start providing new metrics. In this case, for instance, it would be interesting to focus on the extent to which an algorithm is, under unreliable certainty, either predicting correctly or erring, regardless of whether the answer is right or wrong.

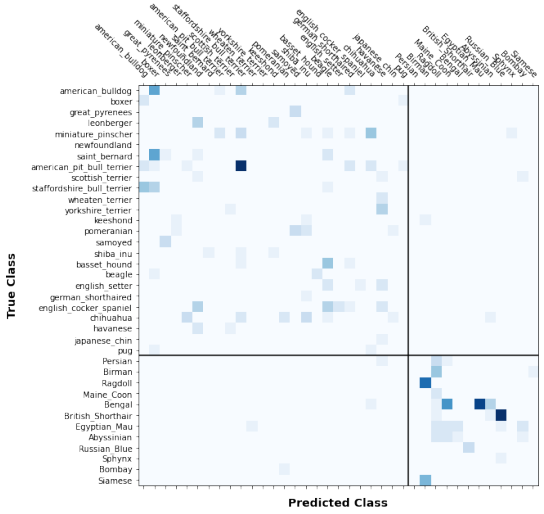


Figure 6: Confusion matrix for the Oxford-IIIT Pets data set [9]. Darker the squares, more errors were made for that pair of breeds.

5. Discussion

Why should we care if algorithms mistake dogs for cats? This is clear when similar tasks are proposed in fields where the lives of human beings and their fundamental rights are at risk of being left unprotected. In these fields, even in the case of a low percentage of non-human errors, the consequences could have a catastrophic and irreversible impact. One concrete such example happened in 2018, when a Uber self-driving car was not able to recognize a woman in a bicycle crossing a road at night in Tempe, Arizona.[2]. A human most probably would have recognized the woman and hence this is a non-human error. We do not know if the backup driver could have reacted on time, but she was seeing a video as the car was working well until then. Finally, she was charged of negligence, as Uber quickly settled with the family of the victim to avoid being sued [11].Hence, this event at the end impacted the lives of two women.

One related issue that we do not discuss is another bad habit: predicting an answer even when we have low confidence. For example, in Figure 5 (b), any smart/honest person would say “I don’t know” with such low confidence. Even in case (a), if there is a harm risk, not giving an answer might be a safer output. In the Uber example is the same. Predicting ”I don’t know” and stopping, might be safer than predicting ”there is no human in front of me and is safer to run over the object” (notice that the later assumption might be still dangerous for the passengers).

[2]https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe

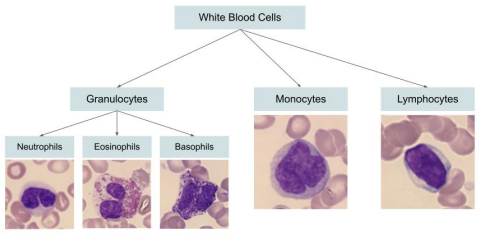


Figure 7: Classification of white blood cells. As it happened with cats and dogs, after further investigation on the risk associated to mistake one cell for another [12, 13], this tree could be used as a taxonomy to define disparate errors. Images collected by [14].

In fact the self-driving car did predict a bicycle one of the times [11].

We are currently working on a problem that is technically very similar to the classification of dogs and cats according to breed, but is a real-life application that is much more relevant to humans: the classification of white blood cells. This problem is also formulated as a fine-grained image classification problem and, even when the number of different classes of elements is much smaller than in the previous example (see Figure 7), their differentiation is very important. Indeed, [12] points out that neutrophil levels were associated with breast cancer risk, including advanced stages of breast cancer. In the meta-analysis proposed by [13], it was shown that breast cancer patients with a higher ratio of neutrophils to lymphocytes had a higher relapse and lower overall survival.

The importance of including an in-depth study of the errors that an algorithm could make in this classification is evident, just as it is fundamental that in these complex use cases, both the evaluations of the algorithms and their publication are accompanied by the corresponding parameters or new metrics that make visible the different errors made, their frequency and their associated risk based on professional knowledge. Moreover, these parameters could provide not only transparency and explainability to the model, but also valuable clues to researchers that would allow the algorithms to be improved in terms of human-centered responsibility and accountability.

References

[1] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, A. Madry, Robustness may be at odds with accuracy, arXiv preprint arXiv:1805.12152 (2018).

[2] D. Tsipras, S. Santurkar, L. Engstrom, A. Ilyas, A. Madry, From imagenet to image classification: Contextualizing progress on benchmarks, in: International Conference on Machine Learning, PMLR, 2020, pp. 9625–9635.

[3] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021, ACM, 2021, pp. 610–623.

[4] C. G. Northcutt, A. Athalye, J. Mueller, Pervasive label errors in test sets destabilize machine learning benchmarks, arXiv preprint 2103.14749 (2021).

[5] R. Geirhos, K. Meding, F. A. Wichmann, Beyond accuracy: Quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency, arXiv preprint arXiv:2006.16736 (2020).

[6] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, A. Madry, Adversarial examples are not bugs, they are features, arXiv preprint arXiv:1905.02175 (2019).

[7] M. Ruggero Ronchi, P. Perona, Benchmarking and error diagnosis in multi-instance pose estimation, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 369–378.

[8] A. Ng, Machine learning yearning, 2018. URL: https://info.deeplearning.ai/machine-learning-yearning-book.

[9] O. M. Parkhi, A. Vedaldi, A. Zisserman, C. Jawahar, Cats and dogs, in: 2012 IEEE conference on computer vision and pattern recognition, IEEE, 2012, pp. 3498–3505.

[10] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, N. Houlsby, Big transfer (bit): General visual representation learning, in: 16th European Conference on Computer Vision, Part V 16, Springer, 2020, pp. 491–507.

[11] L. Smiley, 'I'm the operator': The aftermath of a self-driving tragedy, Wired (2022).

[12] Y. Okuturlar, M. Gunaldi, E. E. Tiken, B. Oztosun, Y. O. Inan, T. Ercan, S. Tuna, A. O. Kaya, O. Harmankaya, A. Kumbasar, Utility of peripheral blood parameters in predicting breast cancer risk, Asian Pacific Journal of Cancer Prevention 16 (2015) 2409–2412.

[13] B. Wei, M. Yao, C. Xing, W. Wang, J. Yao, Y. Hong, Y. Liu, P. Fu, The neutrophil lymphocyte ratio is associated with breast cancer prognosis: an updated systematic review and meta-analysis, OncoTargets and therapy 9 (2016).

[14] X. Zheng, Y. Wang, G. Wang, J. Liu, Fast and robust segmentation of white blood cell images by self-supervised learning, Micron 107 (2018) 55–71. URL: https://www.sciencedirect.com/science/article/pii/S0968432817303037.