To err is (only) human. Reflections on how to move from accuracy to trust for medical AI

Federico Cabitza[1], Andrea Campagner[1], and Edoardo Datteri[2]

[1] Dipartimento di Informatica, Sistemistica e Comunicazione
Università degli Studi di Milano-Bicocca,
Viale Sarca 336, 20126, Milano, Italy federico.cabitza@unimib.it
[2] Dipartimento di Scienze Umane per la Formazione
Università degli Studi di Milano-Bicocca,
Piazza dell'Ateneo Nuovo, 1, 20126, Milano, Italy

Abstract. In this paper, we contribute to the deconstruction of the concept of accuracy with respect to machine learning systems that are used in human decision making, and specifically in medicine. We argue that, by taking a socio-technical stance, it is necessary to move from the idea that these systems are "agents that can err", to the idea that these are just tools by which humans can interpret new cases in light of the technologically-mediated interpretation of past cases, like if they were wearing a pair of tinted glasses. In this new narrative, accuracy is a meaningless construct, while it is important that beholders can "believe in their eyes" (or spectacles), and therefore trust the tool enough to make sensible decisions.

Keywords: Accuracy · Decision Support Systems · Medical Artificial Intelligence · Machine Learning

1 Introduction

Machine Learning (ML) techniques (or, broadly speaking, Artificial Intelligence - AI) are becoming more and more common in Decision Support Systems (DSSs) employed in an increasing number of business processes, especially in regard to discriminative tasks, like disease detection and classification in medicine.

To address the so called "trust chasm" [9] and hence gain impact on situated practices of decision making, DSSs are usually associated with "vanity" measures that are supposed to relate to their intrinsic and context-independent quality, and hence the extent they are trustworthy in real-world practices. This sort of quality is universally denoted as *accuracy*. Both in its narrow sense, and its broadest one, *classification accuracy* is always related to the concept of *error*: in the former case, the relation is straightforward, as accuracy is but the complement of error rate; but also all the other more popular measures, like specificity, sensitivity, precision, F-score, G-mean, C-statistic, are grounded on the so called "confusion matrix", and therefore to the tally of the various types of errors that have occurred in (several) executions of the classification algorithm.

This contribution stems from the recognition that accuracy, broadly meant, is a partial and imperfect quality measure. It is partial because it only regards discriminative performance, while the quality of a DSS used in real settings should also (or rather) encompass estimations of calibration quality, and fit-for-use metrics [13], like efficiency and utility (or net benefit), not to mention more socio-technical and human-related measures, like user satisfaction, fairness (in terms of error disparity across different population groups [37]) as well as social and human sustainability: in particular the former kind of sustainability regards "how organisational activities affect people's physical and mental health and well-being" [36], while the latter relates directly to human development and continuous learning as necessary components of work practices to achieve quality work results and excellence [4]. For this reason, an extremely accurate system that brought an "excess of efficiency" with itself in terms of throughput, or that led to unrealistically higher expectations by the consumer [39], or that induced forms of over-reliance on its services by its users, and subsequent automation bias and deskilling [10, 23], would be a technology with very low social and human sustainability and potentially harming the company where it is operated, regardless of its accuracy.

Moreover, also these considerations apart, accuracy is an imperfect measure because it entails two ill-grounded ideas:

- accuracy on past data is equal to accuracy on unseen data. This is true only on a probabilistic perspective and if the unseen (new) data are similar to past data, that is, e.g., data are taken from similar populations and measured in similar ways (which is often not the case [30]) and, most notably, the so-called *concept drift* [40] has not occurred. Only if the data used to train the predictive model are not significantly different from the new instance that the model has to classify, the frequentist statement "the system was accurate for 90 cases out of 100" can become a probabilistic one like "the system has a probability of 90% to be accurate on this new instance". However, data similarity between new and old data, despite being important for generalization of a DSS [2], is almost never verified and the uncertainty about past performance is seldom represented (e.g., in terms of confidence intervals or average confidence scores).
- the accuracy of a DSS is the accuracy of decision making (and, related but different, this latter could not be lower than the DSS accuracy). Both these assertions could be plausible only if we considered human decision making as occurring in the vacuum, as a structured evaluation of alternative options, and if human decision makers did not rather rely on gut feeling, intuition [29], and contextual information that are hard (if not impossible) to codify or represent as data; or, the other way round, as if they did not develop forms of *automation complacency* and *automation bias* quite easily [34, 23], especially if supported by allegedly "very accurate" decision aids.

To circumscribe our argumentation, we will focus on the second point (although we will also briefly touch on the first) and will take the medical decision making as reference application domain, not only for our extensive experience

in this field (e.g., [8, 9]), but also for the challenges that medical data and medical processes pose to those who want to build useful and trustworthy decision support systems, especially in terms of variability [11] and concept drift [26].

To put this point within the frame of socio-technical research, we can rephrase it in terms of the necessary shift between a *technical* way to assess classification accuracy to a more socio-technical one, that is one taking into account how decisions are made in naturalistic settings [28], and by whom.

2 A framework for a different narrative

The traditional way to look at decision support systems embedding models that have been built with machine learning techniques, and their output, which is usually denoted as a *prediction* (even if they do not regard future events), grounds on two basic assumptions.

1. predictions regard objects that are out-there, in the real world.
2. predictions are statements that can be assimilated to judgements, that is assertions with a true / false value.

The first attitude characterizes what we can call an "externalist" view of machine learning, because it is related to external things with respect to the data (symbolic) representation of these objects. The second perspective, on the other hand, is what allows many commentators to assimilate the main functions of these system to *cognition* (i.e., recognition, understanding, interpretation, judgment and similar terms that are often associated with machine learning systems), and hence what inspires those researchers that like calling this kind of technological support *cognitive computing* [15].

To this perspective we want to counterpose a dual one: a perspective that we could call both *internalist* (as opposite to externalist) and *perceptual* (as opposite to cognitive, and on the same metaphorical level); from this twofold standpoint predictions are, on one hand, symbolic representations that do not refer to external objects but, rather, they complete an internal, and purposely left incomplete, representation. On the other, predictions are not the expression of agents endowed with any form of cognition (or with a behavior that is assimilable to cognition in its capability to state the truth), but rather the "tint" (metaphorically speaking) of a translucent medium through which humans can perceive an object. To this respect, DSSs do not assert any statement about the external world but rather facilitate human observers in seeing objects associated with a specific symbolic representation, in light of the assumptions taken to build a mathematical model that describes the representations of other (past) objects.

In this new narrative, we need to move from appraising the accuracy of decision aids to assessing their *trustworthiness* (and hence the reliability of their utterances or interpretations). The former concept regards *truth*, which in its turn is beyond the scope of any computational system and entirely within the network of meanings that constitute (and is constructed by) a human collective

and community; the latter concept, on the other hand, regards more context-dependent and situated aspects, like the users' perceptions, attitudes and preferences, and also an idea of the *integrity* and *benevolence* of the machine vendor, which are (along with competence) the main components of trust [16].

This entails some small but significant shifts: it means to move from seeing "intelligent" decision support systems as autonomous agents and *truth enunciators*, to tools that represent (instantiate) a symbolic model of a third actor, the designer (so that these tools act as *designer's deputies* [19]) and are endowed with a memex-like function [7], that is the capability to help humans recall cases, experiences, past interventions, and thus help them establish (and make) sense of new cases through an ever-new network of signs [21]. Another shift requires us to move our focus from accuracy-related metrics to other relevant dimensions that characterize the possible roles of decision support tools in human agencies, like utility, causability [25] and acceptability, which all regard the capability of these systems to contribute to the discourses that motivate action (beyond classification decisions) and provide post-hoc justification for those actions.

This also means to move from the world of objectivity (i.e., of the truth that is indisputable and manifest to anybody) into the world of inter-subjectivity [31], where prospective users of these systems are first involved to create a representative *ground truth* (often by taking true labels on a majority vote); and then their actual users are made aware of the intrinsic reliability of this reference truth, and finally involved to assess the extent they would recommend such a system to their peers (recommendability), or would keep relying on its advice to make their decisions, and be responsibly accountable about them towards any stakeholder.

Concretely, this means to attach to any DSS response a whole network of human experiences and perceptions, related to, e.g., (to keep the analysis quantitative):

- how many times decision makers (DMs) and the DSS agreed upon a case (concordance);
- how many times DMs believe the DSS is right (confidence);
- how many times the DSS is proved right after the fact (accuracy);
- how many times DMs changed their mind for the DSS' advice (performance impact);
- how many times DMs have perceived the DSS useful (usefulness);
- how many times DMs believed to have received interesting elements to factor in their decisions (utility);
- how many times they believe to have been faster in their decision making or, rather, hampered by the DSS (satisfaction);
- how many times the DSS output has facilitated or censured discussion with their colleagues or the patients (collaborative impact);
- how many times it has facilitated learning or relieved from the "burden" of recalling, analogical reasoning and deductive inference (cognitive impact).
- also, how many times users believe such a tool can have nurtured confirmation bias, defensive medicine, disciplinary bandwagon effects and other biases (like automation bias and complacency).

Technically speaking, this also means to discover how probability (or confidence) scores are calculated; how to make explicit and comprehensible the model assumptions; how to represent the uncertainty that affects both the input and the output of these systems; how to take into consideration the similarity between the case to be classified and the cases of the training set, and the similarity between the former case and all of the other cases that are considered belonging to the same class.

3 A case from current events

To illustrate what we practically mean with the above concepts, we will outline a real-life case, taken from our current research activities. In [6], we presented a machine learning model that, on the basis of few hematochemical parameters extracted from routine blood examinations, is capable to determine whether the patients from whom the blood samples at hand were drawn are positive to the Sars-COV-2 virus, i.e., suffer from COVID-19. This model has also been embedded in an online service[3] that can provide the above "prediction" in few seconds once a short questionnaire has been filled in with the blood parameters (see Figure 1).

The service is provided as-is and we made it available to the broader community of Internet users as a proof-of-concept of the feasibility of using routine blood exams for COVID-19 screening, as well as to assess its usefulness either as integration or substitution of the more complex RT-PCR[4] test: in short, we did not intend this machine-learning service to provide any medical advice, but rather we want to assess its utility, not only in those settings where there is a shortage of nasopharingeal swabs or molecular test reagents, but also in any setting where blood exams can be done fast and cheap and in a matter of minutes, instead than the many hours necessary for the molecular assays, to diagnose COVID-19.

In what follows, we will imagine that the accuracy and reliability of such a system has been validated and that hence the intended use of this system is twofold: to support the fast screening and management of COVID-19 suspects, while doctors wait for the result of any gold standard reference test (either the molecular or serologic assay, CT scans,...); and to complement the result of the reference test in case this latter one were found negative even in presence of serious COVID-like symptoms. This assumption makes our system akin to any software-as-a-medical-device, whose intended use is to support diagnosis with explicit advice given to physicians, when provided with a number of data attributes regarding the signs and symptoms of a given patient.

Thus, in this light such a system asserts:

1. what disease the person who has those symptoms suffers from.

[3] Available at https://covid19-blood-ml.herokuapp.com/.

[4] This is the acronym for Reverse transcriptase-polymerase chain reaction, a laboratory technique for the quantification of viral RNA in research and clinical settings.

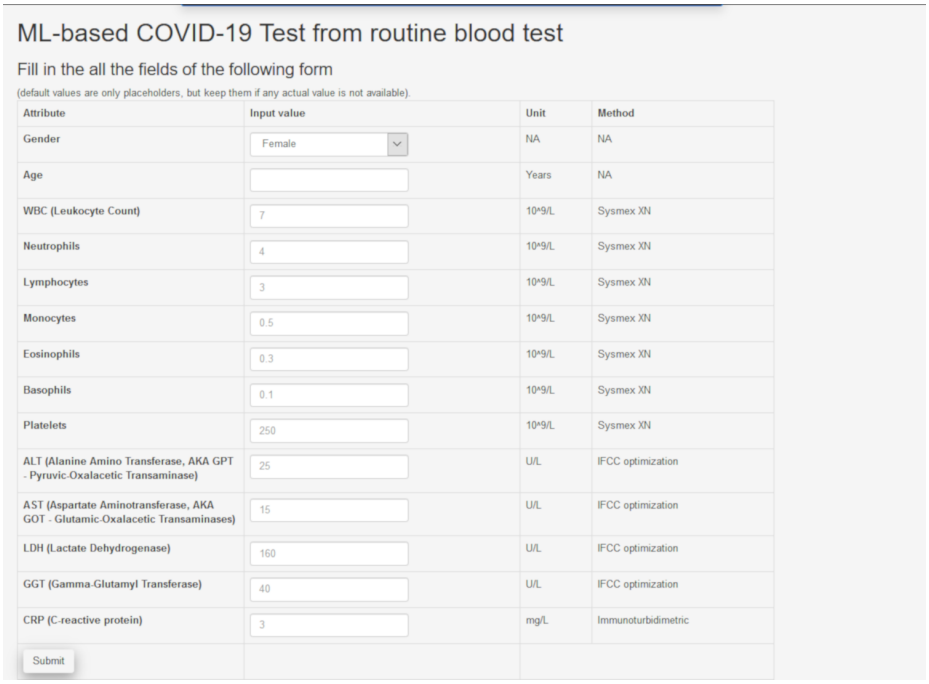

Fig. 1. A screenshot from the diagnostic online service for fast COVID-19 screening.

2. what disease an ideal person suffers from, who manifests only those symptoms and, (we assume) who has all the other physiological parameters within the normal range of values.
3. which, among the records that the system received as a training set, the record at hand resembles the most.

The output of the diagnostic software does not change across these three alternative ways to interpret it, only the expectations of the users does; and the underlining idea of the role of the software within a decision making setting: an externalist, cognitive, and potentially substitute (for the sake of efficacy and efficiency) role in the first case; a more prudent but still cognitive and externalist view in the second case; an internalist and perceptual (in the metaphorical sense above) sense the third and last one. Moreover, and more importantly for our argumentation, the first two views can be interpreted in terms of error, and hence accuracy. Conversely, the third one is more open to analogical reasoning, and to the further interpretation by the physicians involved.

In Figure 2, we see the three ways in which the above system can (and actually does in the current version online) display the result of its computation. The first one (a in Figure 2) only gives the predicted target label: this is a common approach in the machine learning community but also in medicine, as it is how exam results are given, to either patients or the prescribing physicians, in the

case of so-called qualitative tests. This response is the most straightforward one, as it is easy to convene, and it addresses the original inquiry that motivated the prescription and collection of the test: is this patient positive to COVID-19?

The second method (b in Figure 2) presents the result in a tabular form: in particular, the table reports the so-called *confidence scores* (also known as probability scores); intuitively (but the next section will get into the details of this aspect), the system exposes how much confident it is (or the human decision maker can be, by taking the machine's response at face value) that the patient does not suffer from COVID-19: in particular, the model estimates that there is one possibility out of 5 that the patient should be isolated from other patients, admitted to a specific hospital ward, and put, e.g., under anti-inflammatory steroids as soon as possible.

The third method (c in Figure 2) renders the same information above in a visual manner, and more in particular in terms of a *vague visualization* [1], that is a visualization where quantification aspects are purposely concealed from the decision makers, so that these latter ones are nudged towards a more comprehensive assessment of the case at hand. In this specific case, the horizontal position of the circle expresses the confidence (or probability, as mentioned above) so that the closer the circle to one extreme of the colored bar, the stronger the confidence; while the dimension of the circle graphically represents the confidence interval of the probability estimate so that the larger the circle, the higher the uncertainty of the estimate. Moreover, whenever the confidence is lower than a specific threshold, and the circle moves into the middle section of the bar, its color tends to blur with the color of the bar itself so that, in extreme cases, the visualization does actually hide the machine's response from view, and it acts as a *programmed inefficiency* [10] within the decision aid. The ways in which a vague visualization can be rendered are many: our system could have used transparency, instead of position and hue, so as to mimic real serological tests, where even a faint line in the test region is noted to be positive; or other metaphors could be used, like those mentioned in [1]: in any case, the reader would have had to interpret the result, instead of receiving it plain and simple in numeric or quantitative terms.

Both cases b and c in Figure 2 deal with the concept of confidence, in either numeral or visual form. In the next section, we will see this concept in more details, and outline further alternative ways in which the same information resulting from the DSS computation can be expressed ,according a more internalist and perceptual standpoint.

4 Inside the confidence machine

As we previously argued, the shift of emphasis from an *externalist/cognitive* perspective to an *internalist/perceptual* one may require the construction of a network of additional information, as a sort of meta information, which is aimed at clarifying and describing different (internal) aspects of the DSS: namely, how its output is to be interpreted and understood; on which grounds this output

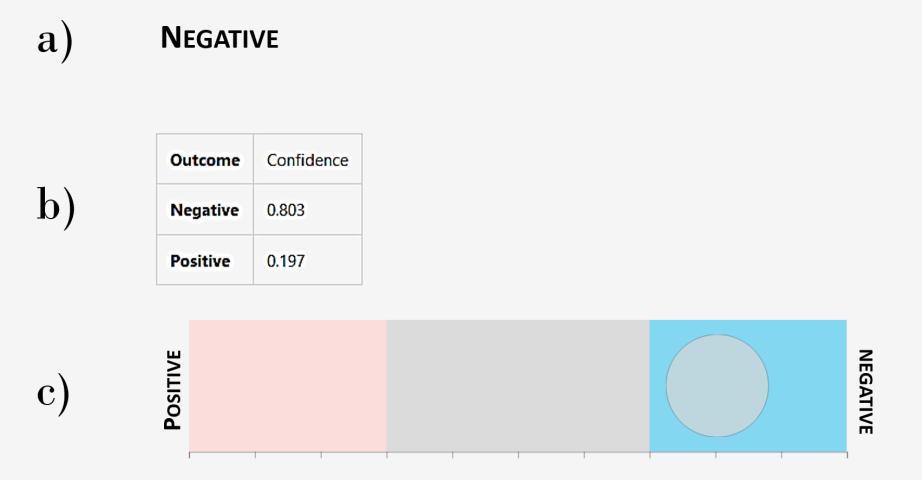

Fig. 2. Three alternative (or complementary) ways to present the output of a prediction model for COVID-19 screening: a) the target label for the case at hand; b) a pair of confidence scores; c) a visual, and purposely under-specified, rendering, or vague visualization.

is computed; how the inner workings of the DSS can be employed to provide additional informative pieces of data.

The first aspect that requires to be understood regards the concept of *confidence score*, which could intuitively be seen as a form of uncertainty representation. As we mentioned above, in the externalist perspective, the DSS is often supposed to return a single label as output (hence the emphasis on computing error): nonetheless, the actual output of DSS is usually provided as a weight vector $w \in R^{|Y|}$ that attaches to any possible label $y \in Y$ a so-called confidence score $w(y)$, whose underlying semantics is that the greater the value of $w(y)$, the more confident the model is in assigning label y to the case at hand. This weight vector is usually required to be *normalized*, that is $\sum_y w(y) = 1$: in this case, the scores $w(y)$ are called *probability scores* and are assumed to define a probability distribution over the class labels.

While, intuitively, having a quantification of the uncertainty attached to the DSS response may be seen as useful, it should be made clear that the meaning and usefulness (from a decision support perspective) that the decision maker can draw from the confidence scores may be affected by how these scores are actually computed. From a technical perspective, this heavily depends on the specific algorithmic family to which the DSS belongs to; as examples: in ensembles of Decision Trees (a class that encompasses popular algorithms such as Gradient Boosting [22] and Random Forests) the probability of a given a label, for a specific case x, is computed as the (possibly weighted) fraction $p_x(y)$ of trees associating y as response to case x; on the other hand, in regard to the Logistic Regression (and, by extension, Artificial Neural Networks) this same probability

$p_x(y)$ is computed through the application of a non-linear function (i.e. the logistic function $\sigma(t) = \frac{1}{1+e^{-t}}$) to the response of a linear model.

In order for the confidence scores to be useful as uncertainty quantification mechanisms, these should be required to satisfy some intuitively useful properties, such as *calibration* [5] (a requirement for the probability scores provided by the DSS to be well-aligned with observed likelihoods), or the ability to distinguish and properly represent *aleatoric* (uncertainty due to variability in data) vs *epistemic* (uncertainty which is only due to the nature of the adopted DSS and its training process) uncertainty [27]. Notice that not all classes of DSS satisfy these properties (or others) by default:

- As regards calibration, it is widely known that logistic regression (and, by extension, classical non-regularized neural networks) or Bayesian models (such as Gaussian Processes) are well-calibrated; the same does not holds true for most other algorithms [14] (such as tree ensembles, modern deep learning algorithm [24] or support vector machine) that typically require the application of post-processing techniques such as isotonic regression or Platt scaling [32];
- In the same way, most model classes do not provide a clear distinction between aleatoric and epistemic uncertainty sources and, more in general, this second property is seen as an open problem in the current debate within the ML community [27].

Apart from technical considerations on the confidence scores provided by a DSS, the shift towards an internalist/perceptual perspective also requires to reflect on to how the DSS could be used to provide additional information such as, as we mentioned in the previous Sections, the collection of cases which are most similar to the case at hand, in order to provide the decision maker with some form of analogical ground for the DSS predictions. To this aim, the most intuitive approach requires the definition of a *similarity function* associating each pair of cases to a number that represents their similarity with a positive real number: the above mentioned information is provided by a direct application of this function to the relevant cases. Notice, however, that this method completely ignores the structure of the DSS itself, as most common approaches to implement DSS do not directly rely on this kind of similarity functions: indeed, among the popular approaches used to implement DSSs, only k-nearest neighbors and support vector machines could be properly interpreted as similarity-based [17]. In all other cases, simply applying an external similarity function might provide results that are in contrast, and completely unrelated, both to the response provided by the DSS and to how the DSS actually uses the past experience to provide that response: in those cases, more meaningful measures of similarity (hence, means to provide the above mentioned network of information) can be obtained on the basis of a technical understanding of the model assumptions that underlie the specific DSS under consideration.

As a simple illustration, consider the cases of tree ensembles and neural network models. In the former case, the notion of similarity between two cases x, x' can be defined as the number (or proportion) $s(x, x')$ of trees assigning the same

label y to both x, x' [20]; on the other hand, in the latter case this same quantity could be meaningfully defined as the similarity between the representations of x and x' computed by the last hidden layer of the neural network [33]: in both cases the provided definition of similarity aligns well with the DSS assumptions (e.g. in a tree ensemble, if two cases are frequently classified in the same way, than we also expect the ensemble as a whole to assign them to the same label) and it provides useful summary information about the conceptual structure through which the DSS interprets its past experience.

Finally, we notice that similarity can be employed not only to describe the most similar (or dissimilar) cases for a given case at hand, but it could also be applied to evaluate the similarity of a given case with respect to the training set *as a whole*, for instance by looking at whether the average similarity of the case at hand with all the cases in the training set is compatible with the distribution of similarities *inside* the training set itself, using an approach that reminds of *nonconformity scores* [41] for hypothesis testing or multi-variate permutation tests. This information could ultimately be useful to assess whether should decision makers trust the predictions and information provided by the DSS for a specific case or, more in general, to evaluate the robustness of the DSS itself.

5 Conclusions

In the next future, DSS will be increasingly more part of the networks of agents that are mobilized to make faster, more accurate, more sensible decisions in sensitive fields like the medical one, e.g., to decide whether a patient is ill or not, will benefit from a treatment more than she will be harmed from it, or even whether she should receive a treatment or not. For this reason, we feel the urgent need to advocate a radical shift in considering the role of DSS in human decision making, especially in sensitive fields where decisions can produce "legal or similarly significant effect on individuals" (cf. art. 24 of the Regulation EU 2016/679, also known as General Data Protection Regulation, or GDPR), that is have an impact on individuals' life, health and well-being.

The shift we advocate is the one from the naive perspective that sees AI-based DSS as actors that can discern the right from the wrong, and hence be right or make mistake; to the perspective seeing these computational systems as tools by which users can "mine" (i.e., retrieve and analyze) past experience and get clues for significant correlations and associations. However, attaching significance and making sense out of these hints will be the call of humans, who are the only ones who can make mistake, according to their local, yet public, sense of right, wrong, and truth.

Thus, due to the socio-technical nature of errors (and also to mitigate the risk of technology-related risks, like over-reliance, automation bias and deskilling [12]), we argue that DSS should be considered more as *perceptual lenses* (not devoid of aberrations), that is as tools by which decision makers can inspect new objects (cases) in the light of other past ones to which a community of experts (through some of its representative members, the raters) attached some value of

contingent truth in the past (labels), rather than oracular aids that have "the capability to state the truth" [38] on those ojects. A similar point has been proposed by Pasquinelli [35] through the provocative idea to see AI as a *nooscope*, that is an "instrument of knowledge" or logical magnification that "maps and perceives complex data patterns that are beyond the reach of the human mind".

According to this perspective, we propose to abandon the discourses that mention accuracy in regard to the performance of machines that we call "decision support systems" and adopt alternative narratives, like those that relate these systems to their capability to enable a more comprehensive interpretation of the cases at hand, abstaining from the production of "machinic" interpretations. In this sense, we also support a semiotic engineering stance to DSS design [18], through which the developers of these systems tell their users about interpretations of the past, which only the users of the present have the right to let inform their current case interpretation and choice of action course.

Summing things up. In this article we have argued that the concept of accuracy is closely related to that of error, intended as an objective (that is objectively established within a normative system) difference (or deviation) from the right answer, choice or belief. We have briefly reflected on the oft-neglected inductivist and probabilistic nature of the concept of accuracy whenever this is related to error rate, and also to some unintended consequences that this mindset brings in, like calling the data that a DSS produces and associates with a new case a "prediction" (a statement about something that still does not exist), rather than what it essentially is, i.e., a *post-diction*, that is a statement *after* (and, to some extent, *about*) the past cases used to train the DSS.

As a consequence of this mindset, ML predictions are also considered new elements to be put into the discursive and generative practices of the decision makers (e.g., medical diagnosticians) while, we have argued, they are but sort of metadata, which computational procedures attach to data in light of both the previous records upon which the ML model has been trained, and the implicit assumptions (what, in the technical jargon, is referred to as inductive bias [3]) underlying the ML model itself.

In this new light, DSS should rather be called "medical experience miners", more than predictive models, and be appreciated not for their "divinatory" skills but rather for their capability to present the case (or cases) from the past that resemble the new case at hand more closely, as well as to allow for counterfactual reasoning on this past-present relationship, like asking "to what extent these two cases would be more or less similar if these data were different?"; or to prune information in order to understand what features of the case at hand contribute in suggesting a particular categorization more than others (cf. feature selection).

However, as we also argued in [9], accuracy is still considered an ontological attribute of DSSs, i.e., something that belongs to these machines regardless of the conditions in which they operate, or of contextual conditions that usually are not to be found in the data (e.g., the difficulty, complexity or rarity of a medical case) but nevertheless characterize the socio-technical setting. Failing to see accuracy as a relational attribute that emerges from the situated interaction

between the system and the user(s), still prevents these systems from bridging what we called the *chasm of trust* in the last mile of AI implementation [9]: that is in building a *trust relationship* with these users, as a necessary condition for responsible and appropriate use.

References

1. Michela Assale, Silvia Bordogna, and Federico Cabitza. Vague visualizations to reduce quantification bias in shared medical decision making. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 3: IVAPP, 209-216, 2020 , Valletta, Malta*, pages 209–2016.
2. Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. A theory of learning with similarity functions. *Machine Learning*, 72(1-2):89–112, 2008.
3. Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.
4. Peter M Bednar and Moufida Sadok. Socio-technical toolbox for business systems analysis and design. In *STPIS@ CAiSE*, pages 20–31, 2015.
5. Antonio Bella, Cèsar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. Calibration of machine learning models. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pages 128–146. IGI Global, 2010.
6. Davide Brinati, Andrea Campagner, Davide Ferrari, Giuseppe Banfi, Massimo Locatelli, and Federico Cabitza. Detection of covid-19 infection from routine blood exams with machine learning: A feasibility study. *Journal of Medical Systems*, 44(8):135, 2020.
7. Vannevar Bush. As we may think. *interactions*, 3(2):35–46, 1996.
8. Federico Cabitza. Biases affecting human decision making in ai-supported second opinion settings. In *11676 LNAI, International Conference on Modeling Decisions for Artificial Intelligence*, pages 283–294. Springer, 2019.
9. Federico Cabitza, Andrea Campagner, and Clara Balsano. Bridging the "last mile" gap between ai implementation and operation:"data awareness" that matters. *Annals of Translational Medicine*, 8(7), 2020.
10. Federico Cabitza, Andrea Campagner, Davide Ciucci, and Andrea Seveso. Programmed inefficiencies in dss-supported human decision making. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 201–212. Springer, 2019.
11. Federico Cabitza, Angela Locoro, Camilla Alderighi, Raffaele Rasoini, Domenico Compagnone, and Pedro Berjano. The elephant in the record: on the multiplicity of data recording work. *Health informatics journal*, 25(3):475–490, 2019.
12. Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. Unintended consequences of machine learning in medicine. *Jama*, 318(6):517–518, 2017.
13. John M Carroll and Mary Beth Rosson. Getting around the task-artifact cycle: how to make claims and design by scenario. *ACM Transactions on Information Systems (TOIS)*, 10(2):181–212, 1992.
14. Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, 2006.

15. Min Chen, Francisco Herrera, and Kai Hwang. Cognitive computing: architecture, technologies and intelligent applications. *IEEE Access*, 6:19774–19783, 2018.
16. Sandy C Chen and Gurpreet S Dhillon. Interpreting dimensions of consumer trust in e-commerce. *Information technology and management*, 4(2-3):303–318, 2003.
17. Yihua Chen, Eric K Garcia, Maya R Gupta, Ali Rahimi, and Luca Cazzanti. Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 10(Mar):747–776, 2009.
18. Clarisse Sieckenius De Souza. *The semiotic engineering of human-computer interaction*. MIT press, 2005.
19. Clarisse Sieckenius De Souza, Simone Diniz Junqueira Barbosa, and Raquel Oliveira Prates. A semiotic engineering approach to user interface design. *Knowledge-based systems*, 14(8):461–465, 2001.
20. Dmitry Devetyarov and Ilia Nouretdinov. Prediction with confidence based on a random forest classifier. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 37–44. Springer, 2010.
21. Umberto Eco. Metaphor, dictionary, and encyclopedia. *New Literary History*, 15(2):255–271, 1984.
22. Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
23. Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012.
24. Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017.
25. Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.
26. Hamish Huggard, Yun Sing Koh, Gillian Dobbie, and Edmond Zhang. Detecting concept drift in medical triage. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1733–1736, 2020.
27. Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction. *arXiv preprint arXiv:1910.09457*, 2019.
28. Gary Klein. Naturalistic decision making. *Human factors*, 50(3):456–460, 2008.
29. Raanan Lipshitz. Decision making as argument-driven action. *Decision making in action: Models and methods*, pages 172–181, 1993.
30. Kim Luijken, Rolf HH Groenwold, Ben Van Calster, Ewout W Steyerberg, and Maarten van Smeden. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective. *Statistics in medicine*, 38(18):3444–3459, 2019.
31. Andrew Maul, Luca Mari, and Mark Wilson. Intersubjectivity of measurement across the sciences. *Measurement*, 131:764–770, 2019.
32. Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.
33. Harris Papadopoulos, Volodya Vovk, and Alex Gammermam. Conformal prediction with neural networks. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, volume 2, pages 388–395. IEEE, 2007.
34. Raja Parasuraman and Dietrich H Manzey. Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3):381–410, 2010.

35. Matteo Pasquinelli. How a machine learns and fails. *spheres: Journal for Digital Cultures*, (5):1–17, 2019.
36. Jeffrey Pfeffer. Building sustainable organizations: The human factor. *Academy of management perspectives*, 24(1):34–45, 2010.
37. Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.
38. Eŕic Sadin. *L'intelligence artificielle ou l'enjeu du siecle: anatomie d'un antihumanisme radical*. L'ećhappeé, 2018.
39. Edward Tenner. *The Efficiency Paradox: What Big Data Can't Do*. Vintage, 2018.
40. Alexey Tsymbal. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, 106(2):58, 2004.
41. Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.