Reframing the Accuracy/Interpretability Trade-Off in Machine Learning

ZHANGHAN YIN (UC BERKELEY) and BORIS BABIC (UNIVERSITY OF TORONTO)

A central open question in the machine learning explainability/interpretability literature is the extent to which restricting consideration to interpretable models decreases classification accuracy. This has so far proven difficult to evaluate rigorously. The question depends on too many assumptions about the empirical context and the user's subjective state of mind. In this paper, we reframe the problem in a more tractable way. We focus on cases where a human decision-maker must collaborate with a machine and we evaluate the effect of a model's opacity on the team's collaborative performance. We first explain methodologically under what conditions the team performance benefits from a transparent collaboration, and we then demonstrate how interpretability improves team performance in practice, using an experiment on a synthetic classification problem.

ACM Reference Format:
Zhanghan Yin (UC Berkeley) and Boris Babic (University of Toronto). 2024. Reframing the Accuracy/Interpretability Trade-Off in Machine Learning. 1, 1 (August 2024), 15 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

A common suggestion (and sometimes assumption) in the literature on interpretable machine learning is that there may be a trade-off between the interpretability of a model and its classification accuracy [1]. The extent to which an accuracy/interpretability trade-off ("AIT") exists is in many respects the million-dollar question in interpretable ML. But this relationship is hard to prove (or disprove) because interpretability eludes a rigorous mathematical definition [2]. Indeed, systematic discussions of the AIT are thus far largely based on empirical observations in specific contexts where accuracy seems to suffer with more interpretable models [3–5].

Meanwhile, some recent work demonstrates that high stand-alone ML classification accuracy – in the sense of high specificity and sensitivity, for instance – does not always translate into better real-life performance when humans are part of the ultimate decision procedure [1, 6–8]. This is partly because humans can be fallible and biased, thereby failing to appropriately incorporate a probabilistic prediction into their subjective beliefs [9, 10]. Yet given the importance of human-ML collaboration, especially in medical and criminal justice applications, it is particularly valuable to examine ML performance in collaborative settings.

As a result of these two observations – (1) the difficulty of evaluating AIT in the abstract and (2) the importance of human-ML collaboration – in this paper we reframe the problem into one that is both particularly salient to real-life ML applications and which can be addressed rigorously. Instead of asking: "Does accuracy suffer when we restrict our consideration to interpretable models only?" we are going to ask: "When is it more effective to present the human with a more interpretable, but potentially less accurate model?"

By framing the problem in terms of human-ML collaboration, we can explore and quantify the general cost of misinterpretation while avoiding the need to provide a universal definition of what makes a model interpretable. In

Authors' address: Zhanghan Yin (UC Berkeley); Boris Babic (University of Toronto).

Manuscript submitted to ACM

particular, we demonstrate that the AIT is not as simple as the widely accepted negative relation between interpretability and accuracy. A more delicate trade-off emerges instead: between the stand-alone performance advantage of a black-box ML and the collaborative performance loss as a result of misinterpretation by the human decision-maker.

This paper proceeds as follows. In Section 2, we discuss some background literature relevant to the AIT. In Section 3, we describe our general framework and formalize it in a classification setting. In Section 4, we mathematically develop a notion of misinterpretation and its associated accuracy cost. This section contains the construction of our complete formal framework and in it, we demonstrate the more subtle relationship between interpretability and accuracy that emerges in a collaborative setting. In section 5, we illustrate this relationship using an experiment on a synthetic dataset.

2 BACKGROUND

There is substantial interest in interpretable machine learning. In scenarios with overarching concerns of justice, for example, such as recidivism risk prediction [11–13], welfare need assessments [14], or medical resource allocation [15, 16], understanding the algorithmic output becomes essential for trust and accountability [17, 18]. Model transparency is also instrumental to the effective implementation of ML systems in medicine, where their performance can depend on how health care professionals actually use them [1, 19]. This prompts a natural question, which is whether and to what extent there exists a trade-off between models which are easy to understand and models which perform best.

However, a major challenge in evaluating the AIT is the difficulty in precisely articulating what makes a model interpretable. Like art or beauty, interpretability is at least in part in the eye of the beholder. On one extreme, we find canonical black-box algorithms such as a convolutional neural network. On the other extreme, we find the simplest white-box algorithms such as short decision trees and perhaps linear models. For the many models in between these extremes, whether and to what extent they are interpretable would be an open question. As a result, there is little consensus on whether such a trade-off exists, in part because it is hard to agree on what we are trading off in the first place.

Nonetheless, there are principled reasons to doubt the existence of the AIT in many contexts, however, one construes interpretability. In [20], Cynthia Rudin explores many areas where there is no apparent advantage to using black-box models. For example, a three-rule model obtained by the Certifiably Optimal Rule Lists (CORELS) algorithm (which is surely interpretable) attains approximately the same accuracy as the well-known proprietary COMPAS recidivism model on the Broward County, Florida data [21]. Rudin argues more generally, using a Rashomon-set strategy, that when a problem allows the same level of accuracy to be attained by a large collection of models, there is likely one which is interpretable. Though finding such a simultaneously accurate and interpretable model is challenging, the observed negative relation between accuracy and interpretability (e.g., as in [5]) could be due to us not having found such a model, rather than it not existing.

But there are also principled reasons to believe that the AIT would hold in general. For example, increasing complexity gives a model more power and more flexibility to approximate highly non-linear feature-label relationships. Indeed, it is well-known that any Borel-measurable function on a finite-dimensional feature space can be approximated arbitrarily accurately by a neural network with only a single hidden layer, given a sufficiently large number of neurons [22]. In [23], the authors propose a formal framework in which instead of trying to define interpretability, we evaluate how the *act of enforcing interpretability* affects performance. Enforcing interpretability (however one defines it) effectively forces us to work with a smaller set of classifiers, and as a result, we may end up with a worse classifier than the one we would use without such a restriction. In their setting, we consider a collection of admissible classifiers, where admissible just means “in consideration", since it would be impossible and impractical to work with all classifiers. Interpretable classifiers then constitute a subset. By using the empirical risk minimizer over the subset of interpretable classifiers, one would end up with a worse classifier than if one were to use the empirical risk minimizer over the set of all admissible classifiers – when the *Vapnik–Chervonenkis (VC) dimension* of the collection of admissible classifiers is finite and when the number of training examples goes to infinity.

While this approach is insightful, the result is not unique to interpretability constraints. It could be applied to any type of restriction on the classifier that is expressible in terms of complexity since any such requirement would result in optimization over a smaller subset of admissible classifiers. In this paper, we suggest a different formal reframing – focusing on how misinterpretation affects *collaborative* performance. While it is true that we too will not "solve" the AIT problem, full stop, we believe our reframing sheds further light on the accuracy-interpretability relationship from a different direction.

3 TEAM PERFORMANCE AND CLASSIFICATION ACCURACY

A model's performance can be evaluated both independently and as part of a human-ML collaboration. For example, an ML-based medical device for detecting diabetic retinopathy (IDx-DR) was recently cleared by the US Food and Drug Administration (FDA) through its De Novo pathway for diagnostic use by clinicians primarily on the basis of its specificity and sensitivity as a stand-alone diagnostic tool [24, 25]. In the context of IDx-DR, this setting is reasonable because the tool is designed to help non-specialists identify high-risk patients and refer them to ophthalmologists if necessary. Hence, humans are not intended to play a substantial role in the classification decision.

Meanwhile, OsteoDetect is another medical device recently cleared by the FDA for identifying, locating, and annotating wrist fractures on X-ray images [26]. Unlike IDx-DR, OsteoDetect is designed to be a collaborative tool – where the clinician (a specialist) would base her diagnosis concurrently on the annotated and unannotated X-ray images. In this case, FDA clearance involved both stand-alone performance testing and an assessment of how clinicians perform with and without the aid of OsteoDetect. And what we see in the OsteoDetect case is likely true in many if not most real-life settings, from ophthalmology and radiology to recidivism and lending: the model's stand-alone performance is at best a partial guide of how it will fare in real-world conditions, because what really matters is how the human makes use of its outputs. In the case of OsteoDetect both stand-alone and collaborative performances were reasonably strong, but this is not always the case [6]. Accordingly, we develop a framework to evaluate a model not as a stand-alone product, but as part of the whole system within which it operates, taking our motivation from the policy perspective expressed in [19].

Let $\mathcal{Y}$ be the space of labels and $\mathcal{X}$ be a measurable space of features. We have a probability distribution on the labels, $\mathbb{P}_Y$, and conditioned on each label $y \in \mathcal{Y}$, we have a probability distribution $\mathbb{P}_{X|Y=y}$ on $\mathcal{X}$. Together $\mathbb{P}_Y$ and $\mathbb{P}_{X|Y=y}$ define the true feature-label joint distribution $\mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$. The feature distribution is given by $\mathbb{P}_X = \int_{\mathcal{Y}} \mathbb{P}_{X|Y=y} \, d\mathbb{P}_Y(y)$.

We define an *agent(s)* (this could be a machine or a human) to be a function $C : \mathcal{X} \to \mathcal{I}$ where $\mathcal{I}$ is the space of *suggestions*. For example, in the case when a doctor is using OsteoDetect to diagnose fractures, $\mathcal{I}$ would be the space of all annotated X-ray images. In the special case when $\mathcal{I} = \mathcal{Y}^k$ for some $k \in \mathbb{N}$ and $\mathcal{Y}$ is a finite label space, we also think $\mathcal{I}$ intuitively like a "space of votes". A function $p : \mathcal{I} \to \mathcal{Y}$ is called a *collaboration process*.

To further motivate and illustrate this conceptual framing, consider a real-life example: the radiological double reading task on diagnosing knee lesions (as in [27]). Let $\mathcal{X}$ be the feature space of MRI images. Let $C_1, ..., C_k$ be a collection of radiologists and ML readers, which are maps from $\mathcal{X} \to \mathcal{I}_i := \{0, 1\}$ (0 indicates there is no knee lesion, and 1 indicates there is a knee lesion). The process, p, could then be to decide whether or not a knee lesion exists on the basis of a majority vote. In this case, $C = C_1 \times C_2 \times ... \times C_k$ and $\mathcal{I} = \mathcal{X}^k$ is the space of votes. See Figure 1, right-panel.

$$\begin{tikzcd}
 & \mathcal{I}_1 \arrow[dr, dashed, no head] & & \\
\mathcal{X} \arrow[ur, "C_1"] \arrow[r, "C_2"] \arrow[ddr, "C_n"'] & \mathcal{I}_2 \arrow[r, dashed, no head] & {} \arrow[r, "p"] & \mathcal{Y} \\
 & \vdots & & \\
 & \mathcal{I}_n \arrow[uur, dashed, no head] & &
\end{tikzcd}
\qquad
\begin{tikzcd}
 & \{0,1\} \arrow[dr, dashed, no head] & & \\
\textbf{MRI Output} \arrow[ur, "\textbf{reader } 1"] \arrow[r, "\textbf{reader } 2"] \arrow[ddr, "\textbf{reader } n"'] & \{0,1\} \arrow[r, dashed, no head] & \textbf{vote} \arrow[r, "\textbf{majority}"] & \{0,1\} \\
 & \vdots & & \\
 & \{0,1\} \arrow[uur, dashed, no head] & &
\end{tikzcd}$$

Fig. 1. A group of n classifiers using collaboration process p (left), and a group of n readers of MRI output for knee lesion diagnoses, collaborating via majority rule (right).

Given a classifier $f : \mathcal{X} \to \mathcal{Y}$, its accuracy is given by

$$\mathscr{A}(f) := \int_{y \in \mathcal{Y}} \mathbb{P}_{X|Y=y}(f(x) = y)\, d\mathbb{P}_Y(y).$$

The *optimal collaboration process* for agent C is the accuracy maximizer

$$p^* := \underset{p}{\textbf{argmax}}\, \mathscr{A}(p \circ C).$$

Now we will illustrate how the choice of process p can affect the team's accuracy, $\mathscr{A}(p \circ C)$, in a way that can come apart from the classifiers' stand-alone performance. A simple but instructive example is the well-known XOR separation problem. XOR is a logical operator which represents exclusive disjunction (exclusive "or"), meaning that a statement of the form "A xor B" is true if and only if either only A is true or only B is true, and false otherwise. The XOR logical operator can be depicted graphically as a two-dimensional classification problem where the feature distribution takes on values $(\pm 1, \pm 1)$, and each observation is labelled 0 if the two coordinates are equal – i.e., the feature values are $(-1, -1)$ or $(1, 1)$ – and 1 if the coordinates are not equal – i.e., the feature values are $(-1, 1)$ or $(1, -1)$. This classification problem is depicted in Figure 2.

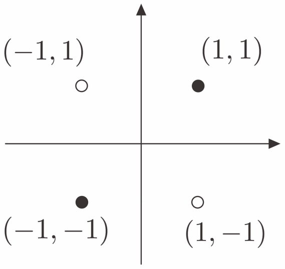

Fig. 2. The XOR separation problem. Formally, let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0, 1\}$ (0 is black, 1 for white), and $\mathbb{P}_Y = \textbf{Unif}\{0, 1\}$, $\mathbb{P}_{X|Y=0} = \textbf{Unif}\{(1, 1), (-1, -1)\}$, $\mathbb{P}_{X|Y=1} = \textbf{Unif}\{(1, -1), (-1, 1)\}$.

Consider Team 1, consisting of linear classifiers $C_1(x_1, y_1) = I_{\{y_2<0\}}$ (where I_A denotes the indicator function of set A) and $C_2(x_1, x_2) = I_{\{x_1 \geqslant 0\}}$. C_1 (resp. C_2) classifies the upper (resp. left) half-plane as 0 and the lower (resp. right) half-plane as 1. Meanwhile, Team 2 consists of $C_1'(x_1, x_2) = I_{\{x_2 < x_1 - 1\}}$ and $C_2'(x_1, x_2) = I_{\{x_2 \geqslant -x_1 - 1\}}$, which draw diagonal classification boundaries with one half-space containing 3 points and the other half-space containing 1 point. See Figure 3.

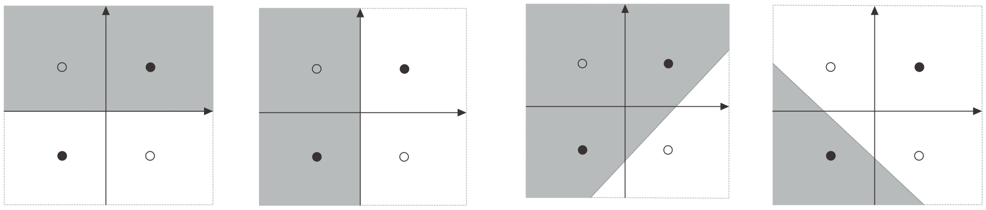

Fig. 3. From left to right: classification partitions of C_1, C_2, C_1', C_2'. As can be seen from the plots, they achieve classification accuracies of 0.5, 0.5, 0.75, and 0.75, respectively.

The optimal process for Team 1 is the function $p^*(0,0) = p^*(1,1) = 1$, $p^*(0,1) = p^*(1,0) = 0$, i.e. $p^*(x_1, x_2) = I_{\{x_1=x_2\}}$. The optimal process for Team 2 is the function $p'^*(0,0) = 0$ and $p'^*(0,1) = p'^*(1,0) = p'^*(1,1) = 1$, i.e. $p'^*(x_1, x_2) = 1 - I_{(0,0)}$. See Figure 4.

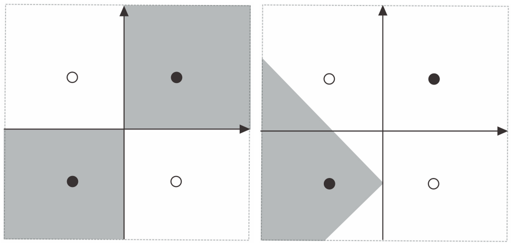

Fig. 4. The optimal collaborative classifier of Team 1 and Team 2: $p^*(C_1, C_2)$ and $p'^*(C_1', C_2')$, respectively. They achieve classification accuracy of 1 and 0.75, respectively.

Here is the lesson from this illustration: Team 1 is individually dominated by Team 2 because each player from Team 1 is weaker than every player from Team 2. For Team 1, both C_1 and C_2 have a classification accuracy of 0.5 each, whereas, for Team 2, both C_1' and C_2' have a classification accuracy of 0.75 each.

However, under optimal collaboration process, $p^*(C_1, C_2)$, the two players from Team 1 working together achieve perfect accuracy. Meanwhile, and perhaps even more counter-intuitively, the two players from Team 2 working together, *even under* optimal collaboration process, still cannot achieve perfect classification accuracy! In this example, then, two individually stronger players cannot outperform two individually weaker players even under ideal conditions. Hence, we see how the collaboration process p can have counter-intuitively substantial effects on performance.

This captures, in some respects, Russian chess legend Garry Kasparov's well-known remark that a weak human working with a computer under a good process could outperform both a strong computer alone and, surprisingly, a strong human with a computer under a bad process [28].

4 INACCURACY AND IMPRECISE INTERPRETATION

The preceding example is an idealization in the sense that we can determine objectively what the optimal collaboration process is. Let us call the person (or machine) who comes up with the collaboration process the *curator*. In real-life cases, unlike in the XOR separation problem, we do not have an omniscient curator. For example, the curator might not have full knowledge of the agent(s), i.e., of C. Rather, we have a fallible curator who must use an agent's output in order to make a final decision. In such cases, the curator might come up with a sub-optimal collaboration process, and the quality of that process depends on the extent to which they understand, or can interpret, the agent's output.

From a statistical perspective, it should not be too surprising that the human's understanding of the machine's partition of the feature space should improve collaborative classification accuracy because research in expert forecast aggregation demonstrates that combining total *information/evidence*, or full distributions, is generally much more effective than averaging point estimates alone [29]. And a human who understands the machine will be better able to infer its underlying information/evidence for a particular classification outcome. In this sense, we can think about our contribution to this paper as a development of the literature on forecast aggregation in the area of human-ML collaboration.

4.1 Two Conditions of Interpretability

First, without defining interpretability, we think it is fair to say that whatever interpretability requires, it is surely a concept that depends on both the agent being interpreted and the curator making the interpretation. For example, a sentence like "this ML model is interpretable", would leave the reader wondering, "interpretable by whom?". By comparison, a sentence like "this ML model is interpretable by Dr. Stephen Strange" sounds more natural. Consider the following hypothetical. Suppose that two doctors, Peter and Jocelyn, are equally good at detecting tumors from reading MRI images. They work in the same hospital, and tend to get very competitive against one another. So when the hospital introduces an ML reader tool to assist in the interpretation of MRI images, Peter applies some under-the-table sleight of hand to obtain the secret proprietary information about how the tool works which, it turns out, is a set of cleverly designed, yet flawed, simple decision rules. When the two doctors begin to actually use the ML tool, it is thereby much more interpretable to Peter than it is to Jocelyn. Even though it is the same tool, Peter has knowledge that Jocelyn lacks. Likewise, we can imagine a simple tool, such as a linear classification model, being used by a statistician, and a person with nearly zero mathematical knowledge. Again it is the same tool, but the statistician will find it more readily interpretable.

Secondly, and again without fully defining interpretability, we make another hopefully uncontroversial observation about this concept: Perfect interpretability is effectively equivalent to perfect replication. We will refer to this as the Principle of Replication. We understand that this principle might be a little bit more objectionable than the previous one, but we think it is a natural, one might say behaviorist, understanding of interpretability. It essentially states that if an ML model C is perfectly interpretable to Dr. Stephen Strange, then Dr. Stephen Strange, given any feature x, can perfectly replicate the output $C(x)$ of the ML model. Conversely, if Dr. Stephen Strange can perform perfect replication of C, then we must grant him that his interpretation of C is perfect. In a similar vein, Cynthia Rudin points out that if one ML model is a globally faithful approximation of another, then it is effectively the same model [20]. The point of this condition is to avoid getting too bogged down in mentalistic properties of interpretation – that is, to avoid objections such as: how can we ever know that someone truly understands a model, regardless of how they behave? Our view is that if Dr. Stephen Strange can perfectly replicate the model, then for our purposes in this project at least, the model is interpretable to Dr. Stephen Strange.

Imperfect replication is hence a result of imperfect interpretability. Therefore, we will evaluate the effect of imperfect replication on the quality of the collaboration process which is set by the curator. We do this in order to avoid having to stipulate a full definition of interpretability, which is surely bound to be controversial.

4.2 The Formal Framework

Now we will formalize the above concepts. Let $\omega \in \Omega$ be a measurable space, equipped with probability distribution $\mathbb{P}_\Omega$ that represents the source of all relevant noise – which could consist of mental or systemic factors such as the curator's own biases or the irreducible randomness involved in the underlying algorithm. We assume that $\mathbb{P}_\Omega$ is independent of the feature-label joint distribution. The curator's interpretation is a replica of agent C, denoted by $\hat{C} : \Omega \times \mathcal{X} \rightarrow \mathcal{I}$. In words, it is what the curator thinks C is. The noisiness/uncertainty of the curator's interpretation is modelled by $\hat{C}$'s dependence on the noise space Ω. For ease of notation, we denote $\hat{C}_\omega(x) = \hat{C}(\omega, x)$. See Figure 5.

$$\begin{tikzcd}
\mathcal{X} \arrow[r, dash, dashed] \arrow[rr, bend left, "C"] & \hat{C} \arrow[r] & \mathcal{I} \arrow[r, dashed] & p \arrow[r] & \mathcal{Y} \\
\Omega \arrow[ru, dash, dashed] \arrow[rrru, dash, dashed] & & & &
\end{tikzcd}$$

Fig. 5. C is the actual agent, and $\hat{C}$ is what the human curator assumes it is. $\hat{C}$ might depend on the curator's mental state represented by space Ω. p is the collaboration process.

We can measure how far away C and $\hat{C}$ are from each other through a *fidelity* function $\gamma : \mathcal{I} \times \mathcal{I} \rightarrow \mathbb{R}$. For example, γ could be the $(L_2 \text{ distance})^{-1}$, or the indicator $\gamma(a, b) = I_{(a=b)}$. While there is no "right" choice of a fidenlity function, in this project we will use what we call γ-fidelity between C and $\hat{C}$, which is given by $fid_\gamma := fid_\gamma(C, \hat{C}) := \mathbb{E}_\Omega \mathbb{E}_X \gamma(C(x), \hat{C}_\omega(x))$. This is therefore an assessment of how good the curator's interpretation is. Of course, calculating fid_γ is impossible in most real-life situations, and we propose the empirical estimate by the sample mean $\widehat{fid}_\gamma = \frac{1}{n} \sum_{i=1}^{n} \gamma(C(x_i), \hat{C}_{\omega_i}(x_i))$ where x_i, ω_i are iid samples from $\mathbb{P}_X$ and $\mathbb{P}_\Omega$ – which we refer to as the *empirical* γ*-fidelity* of the curator's interpretation.

There are two ways for the curator to misinterpret the machine's output (see Figure 6, below). First, the curator's interpretation can be *inaccurate*. In this context, inaccuracy means that the curator's understanding of where the machine's classification boundary lies is different from where it actually lies. Second, the curator's interpretation can be *imprecise*. By imprecision, we mean that the curator's estimate of the classification boundary is noisy. We can measure the accuracy of interpretation (with respect to fidelity function γ) by $\mathbb{E}_X \gamma(C(x), \mathbb{E}_\Omega \hat{C}_\omega(x))$; and we can measure imprecision by the expected noise-induced variance $\mathbb{E}_X Var_\Omega \gamma(C(x), \hat{C}_\omega(x))$.

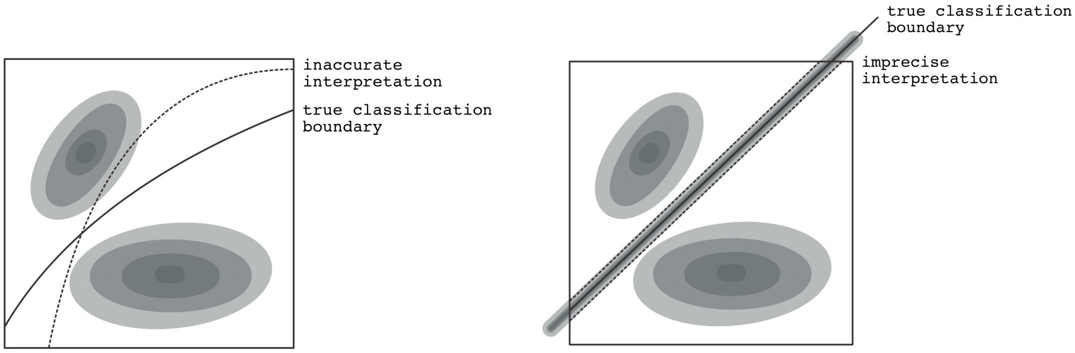
inaccurate interpretation
true classification boundary
true classification boundary
imprecise interpretation

Fig. 6. Inaccurate interpretation (left) and imprecise interpretation (right).

Based on the potentially flawed interpretation $\hat{C}$, the best possible process that the agent could employ is the *subjectively optimal process*:

$$p_\omega^{*,subj} := \underset{p}{\mathbf{argmax}}\, \mathcal{A}(p \circ \hat{C}_\omega).$$

Trivially, the subjectively optimal process is sub-optimal. For each $\omega \in \Omega$,

$$\mathcal{A}(p_\omega^{*,subj} \circ C) \leqslant \max_p \mathcal{A}(p \circ C) = \mathcal{A}(p^* \circ C). \tag{1}$$

Now, we would like to consider the performance deficit caused by imperfect replication (resulting from imperfect interpretation), namely the *interpretation error* given by $IE := \mathbb{E}_\Omega[\mathcal{A}(p^* \circ C) - \mathcal{A}(p_\omega^{*,subj} \circ C)]$. One should expect this quantity to be negatively related to γ-fidelity. Unfortunately for us, such $p^*, p^{*,subj}$ may not always exist when $\mathcal{I} \times \mathcal{Y}$ is not finite. Even in the finite case, the computational complexity of finding $p^*, p^{*,subj}$ by checking all possible mappings is of order $|\mathcal{Y}|^{|\mathcal{I}|}$, which is computationally infeasible even in simple cases. This is, however, not an issue. In some sense, most machine learning techniques would be useless if we could simply obtain the best classifier by checking all possibilities. Rather, we introduce the notion of a *collaboration protocol.* Whereas the collaboration *process* is a function from $\mathcal{I}$ to $\mathcal{Y}$, a collaboration *protocol* is the *procedure* by which the curator obtains that process function.

More precisely, a collaboration protocol is a function $\pi : \Omega \times \mathcal{C} \longrightarrow \mathcal{P}$, where $\mathcal{C}$ is the set of all possible agents and $\mathcal{P}$ the set of all possible collaboration processes, given $\mathcal{X}, \mathcal{I}, \mathcal{Y}$. We will denote $\pi_\omega(C)$ as $\pi(\omega, C)$. The collaboration protocol has an element of noise inherited from Ω. For instance, π could be constant in a majority rule process, or π could be fitting a decision tree – which would introduce an independent layer of randomness. $p^*, p^{*,subj}$ corresponds to the case where π is the protocol which selects the optimal and subjectively optimal classifier.

Given a collaboration protocol π, we denote the π*-chosen* process as $p_\omega^\pi = \pi(\omega, C)$ and the *subjective* π*-chosen* process as $p_\omega^{\pi,subj} = \pi(\omega, \hat{C}_\omega)$. As a replacement for $IE(\omega) := \mathcal{A}(p^* \circ C) - \mathcal{A}(p_\omega^{*,subj} \circ C)$ we consider the π*-interpretation error* ,

$$IE_\pi := \mathbb{E}_\Omega[\mathcal{A}(p^\pi \circ C) - \mathcal{A}(p_\omega^{\pi,subj} \circ C)].$$

Since population accuracy is mostly unobtainable in practical settings, we consider the *empirical* π*-interpretation error* given by

$$\widehat{IE}_\pi := \frac{1}{n}\sum_{i=1}^{n} I(p^\pi \circ C(x_i) = y_i) - \frac{1}{n}\sum_{i=1}^{n} I(p_{\omega_i}^{\pi,subj} \circ C(x_i) = y_i),$$

where x_i, y_i, ω_i are sampled from the feature-label-noise joint distribution.

4.3 The Ensuing System

In this subsection, we will bring together the formal concepts introduced above in order to describe our final model for representing human-ML collaboration. To summarize, we have so far introduced the following: In a learning problem with feature and label spaces $\mathcal{X}, \mathcal{Y}$ and feature-label joint distribution $\mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$, a *system* is a quintuple $(\mathcal{I}, \Omega, \mathbb{P}_\Omega, \delta, \pi)$ where:

- $\mathcal{I}$ is a space of suggestions.
- $(\Omega, \mathbb{P}_\Omega)$ is a noise space that is independent of the feature-label joint distribution.
- $\delta : \Omega \times \mathcal{C} \to \mathcal{C}$, is the interpretation map, where we denote $\hat{C}_\omega = \delta(\omega, C)$ for $C \in \mathcal{C}$ and $\mathcal{C}$ is the space of all agents from $\mathcal{X}$ to $\mathcal{I}$.
- $\pi : \Omega \times \mathcal{C} \to \mathcal{P}$ is the collaboration protocol, where $\mathcal{P}$ is the space of all collaboration processes from $\mathcal{I}$ to $\mathcal{Y}$. We denote the *π-chosen* process as $p^\pi_\omega = \pi(\omega, C)$ and the *subjective π-chosen* process as $p^{\pi,subj}_\omega = \pi(\omega, \hat{C}_\omega)$.

Given a fidelity function $\gamma : \mathcal{I} \times \mathcal{I} \to \mathbb{R}$, with n iid samples x_i, ω_i from $\mathbb{P}_X, \mathbb{P}_\Omega$ respectively, the empirical γ-fidelity is defined by $\widehat{fid}_\gamma(\omega) = \frac{1}{n}\sum_{i=1}^n \gamma(C(x_i), \hat{C}_{\omega_i}(x_i))$. Given n iid samples x_i, y_i, ω_i from the feature-label-noise distribution, the empirical π-interpretation error is given by

$$\widehat{IE}_\pi := \frac{1}{n}\sum_{i=1}^n I(p^\pi \circ C(x_i) = y_i) - \frac{1}{n}\sum_{i=1}^n I(p^{\pi,subj}_{\omega_i} \circ C(x_i) = y_i).$$

Our perspective, therefore, sees interpretability through the lens of a *system*, thereby formalizing the system-view idea first introduced in [19]. Within our framework, one's assessment of the relationship between interpretability and accuracy cannot be separated from the overall system within which the ML model is employed.

Furthermore, we can formalize the AIT problem as the study of the relationship between the fidelity of the curator's interpretation of the agent (measured in practice by $\widehat{fid}_\gamma$) and the collaborative performance of the system (i.e., $\frac{1}{n}\sum_{i=1}^n I(p^{\pi,subj}_{\omega_i} \circ C(x_i) = y_i)$).

Our formal framework provides a unifying perspective of the exceedingly large interpretability and explainability literature: whereas scholars interested in interpretability are concerned with the choice of agent C (without changing the interpretability map δ), scholars who are interested in so-called post-hoc explainability are concerned with finding a better interpretation map δ for a given choice of agent C. Regardless of the approach, the goal is to improve fidelity in practice. See Figure 7 for a visual summary of our full formal model.

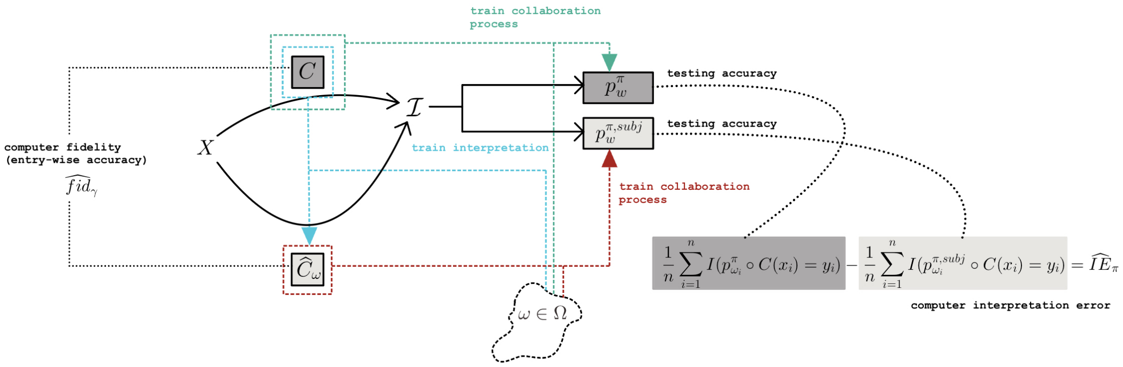
train collaboration process
C
X
$\mathcal{I}$
p_w^π
$p_w^{\pi,subj}$
testing accuracy
testing accuracy
computer fidelity (entry-wise accuracy)
$\widehat{fid}_\gamma$
train interpretation
train collaboration process
$\widehat{C}_\omega$
$\omega \in \Omega$
$\frac{1}{n}\sum_{i=1}^{n} I(p_{\omega_i}^{\pi} \circ C(x_i) = y_i) - \frac{1}{n}\sum_{i=1}^{n} I(p_{\omega_i}^{\pi,subj} \circ C(x_i) = y_i) = \widehat{IE}_\pi$
computer interpretation error

Fig. 7. This diagram depicts our formal construction of a system, the computation of fidelity, and the computation of the interpretation error.

5 EXPERIMENTAL RESULTS

In this section, we put the theoretical model to a test, and devise a synthetic experiment to examine the AIT under our system-view framework. We first sketch the design of our computational experiment and then present our main results. For the sake of readability, the granular details of the experiment are provided in the Appendix, and the full associated code is available on GitHub [30].

While an ideal experiment would involve a large human factors study of many decision-makers and state of the art ML models making judgments on the basis of large real world datasets, we start with something a little bit simpler but nonetheless effective. We use a synthetic system that mimics this ideal as closely as possible, by creating a scenario where the curator is a decision tree trying to interpret and collaborate with a collection of multi-layered perceptrons (MLPs).

To begin, the following characterizes our learning problem under consideration: We start with a 10-dimensional lattice $\{-1, 1\}^{10}$ which consists of 2^{10} lattice points. Half of the lattice points are assigned label 0 uniformly and at random, while the rest are assigned label 1. We generate data from a uniform distribution in a small 10-dimensional cube centered at each lattice point, inheriting the label of the center. To understand this setup, see Figure 8. This gives us a binary classification problem with feature space $\mathcal{X} = \mathbb{R}^{10}$ and label space $\mathcal{Y} = \{0, 1\}$.

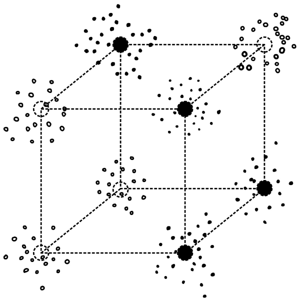

Fig. 8. In the above diagram, uniform clusters are generated around each lattice point of the cube and they inherit the labels of the lattice point. This is a visualization for a 3-dimensional cube, whereas our synthetic dataset is its 10-dimensional analogue.

5.1 Synthetic Experiment Design

To formally construct an **agent**, we train k multi-layered perceptrons (MLPs), $M_1, ..., M_k$, which are meant to play the role of ML models, in this synthetic study, and we fit a decision tree M_{cur} on the same task to play the role of a human curator. The agent function is then the tuple $C = (M_1, ..., M_k, M_{cur})$ and the suggestion space is given by $\mathcal{I} = \{0, 1\}^{k+1}$.

To construct the **interpretation map**, we train a decision tree with exactly the same hyper-parameter as M_{cur} to learn from the ML model's k-tuple $(M_1, ..., M_k)$ (together) to obtain an interpretation $\vec{M}_{int,\omega}$ with outputs in $\{0, 1\}^k$. This framing is similar to Geoffrey Hinton's well-known knowledge distillation process, where the MLP k-tuple is the teacher and the decision tree is the student [31]. The interpretation map is then given by $\delta : (M_1, ..., M_k) \mapsto (\vec{M}_{int,\omega}, M_{cur})$. It is constant on the last index because the classifier M_{cur} is the curator. The ω captures the random noise involved in the decision tree training algorithm.

The **collaboration protocol** is a decision tree with exactly the same hyper-parameter as the one used for M_{cur} and $\vec{M}_{int,\omega}$. This design is motivated by the idea that the same curator is making the interpretation as well as the collaboration process. Together with $\vec{M}_{int,\omega}$, the randomness involved in the tree-fitting constitutes the noise space $(\Omega, \mathbb{P}_\Omega)$. Finally, as our **fidelity function**, we use $\gamma(\vec{a}, \vec{b}) = \frac{1}{k+1} \sum_{i=1}^{k+1} I(a_i = b_i)$. This corresponds to entry-wise accuracy, measuring the percentage of agreement between two vectors, for $\vec{a}, \vec{b} \in \mathcal{I} = \{0, 1\}^{k+1}$.

5.2 Quantities Measured

Each iteration of the experiment consists of choosing a new labelling of the $\{-1, 1\}^{10}$ lattice and regenerating the uniform point clusters, then retraining the MLPs and the trees. From each iteration, we collect the following quantities:

- *Fidelity:* as previously defined, where we choose the fidelity function to be entry-wise accuracy.
- *Collaboration Accuracy:* which is given by $\frac{1}{n} \sum_{i=1}^{n} I(p_{\omega_i}^{\pi,subj} \circ C(x_i) = y_i)$, measuring how the system performs.
- *Ideal Collaboration Accuracy:* which is given by $\frac{1}{n} \sum_{i=1}^{n} I(p_{\omega_i}^{\pi} \circ C(x_i) = y_i)$, measuring how the system would perform had the curator's interpretations been perfect.
- *Interpretation Error:* which is given by ideal collaboration accuracy minus the collaboration accuracy $\frac{1}{n} \sum_{i=1}^{n} I(p_{\omega_i}^{\pi} \circ C(x_i) = y_i) - \frac{1}{n} \sum_{i=1}^{n} I(p_{\omega_i}^{\pi,subj} \circ C(x_i) = y_i)$.
- *(True) Curator Improvement:* which is given by collaboration accuracy minus the accuracy of M_{cur}, measuring how much the curator is able to improve with the help of the MLPs.

- *Ideal Curator Improvement:* which is given by ideal collaboration accuracy minus the accuracy of M_{cur}, measuring how much the curator would have been able to improve from the help of the MLPs if his interpretation is perfect.

The entirety of the system is summarized in the diagram in Figure 9.

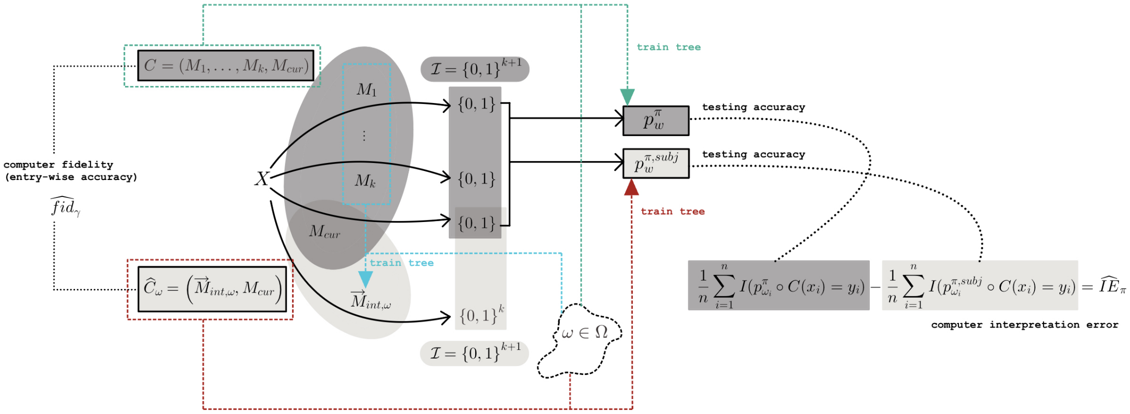
$C = (M_1, \ldots, M_k, M_{cur})$
computer fidelity
(entry-wise accuracy)
$\widehat{fid}_\gamma$
$\widehat{C}_\omega = \left(\overrightarrow{M}_{int,\omega}, M_{cur}\right)$
X
M_1
$\vdots$
M_k
M_{cur}
$\overrightarrow{M}_{int,\omega}$
train tree
$\mathcal{I} = \{0,1\}^{k+1}$
$\{0,1\}$
$\{0,1\}$
$\{0,1\}$
$\{0,1\}^k$
$\mathcal{I} = \{0,1\}^{k+1}$
$\omega \in \Omega$
train tree
p_w^π
testing accuracy
$p_w^{\pi,subj}$
testing accuracy
train tree
$\frac{1}{n}\sum_{i=1}^{n} I(p^\pi_{\omega_i} \circ C(x_i) = y_i) - \frac{1}{n}\sum_{i=1}^{n} I(p^{\pi,subj}_{\omega_i} \circ C(x_i) = y_i) = \widehat{IE}_\pi$
computer interpretation error

Fig. 9. This diagram outlines the system used for our synthetic experiment. The agent consists of k MLPs and a prediction decision tree. An interpretation decision tree is fitted to predict the tuple output of the k MLPs which, together with the prediction tree, forms the interpretation on which a collaboration process decision tree is fitted. From this system, we compute the empirical fidelity (as entry-wise empirical accuracy), the collaborative accuracy, the ideal collaborative accuracy, and the interpretation error.

5.3 Results

We collected 50 data points, consisting of fidelity, collaboration accuracy, ideal collaboration accuracy, and interpretation error, for each of $k = 4, 5, 6, 7, 8, 9, 10, 11, 12$ – a total of 450 samples. These data points are pooled together for analysis. We will present the data as dot plots, depicting fidelity against collaboration accuracy, ideal collaboration accuracy, and interpretation error, respectively, and we compute the associated Pearson correlation coefficients. The results are summarized in Figure 10, below. Following the figure, we describe the main takeaways from this exercise.

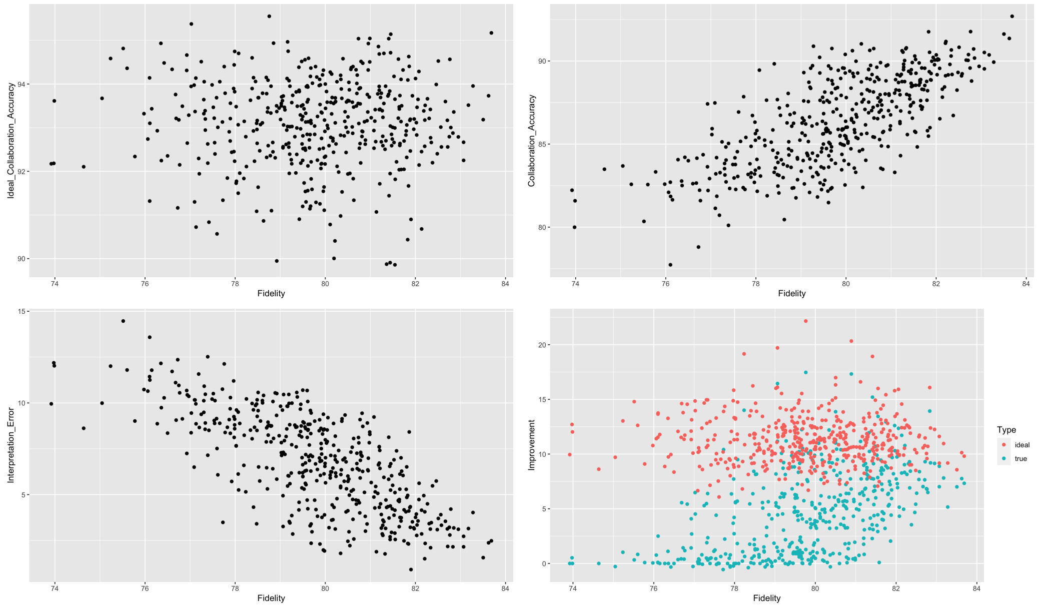

Fig. 10. First three plots: dot plots displaying fidelity against ideal collaboration accuracy, collaboration accuracy, and interpretation error, respectively. The last plot: dot plot displaying fidelity against (true) curator improvement and ideal curator improvement.

As expected, the ideal collaboration accuracy seems to have no relationship with fidelity, admitting a correlation coefficient of about 0.0153, which is close to being completely uncorrelated. The collaboration accuracy, however, bears a strong positive relationship with fidelity, admitting a correlation coefficient of about 0.720. Interpretation error, on the other hand, bears a strong negative relationship with fidelity, admitting a correlation coefficient of about −0.732. Correlation coefficients of this magnitude are typically considered to be quite strong. Our experimental result demonstrates that when the ideal collaboration accuracy (which is to be thought of as the "raw power" of the agent) does not vary with fidelity (which is to be thought of as the quality of the curator's interpretation), interpretation fidelity can have a strong positive effect on collaboration accuracy by lowering the interpretation error.

Moreover, by looking at the last dot plot in Figure 10, we can see that the curator's performance gain through using the MLP models' input increases and approaches optimal accuracy with increasing fidelity. On the other hand, when fidelity is relatively low, the curator benefits very little from collaborating with the MLP models. In fact, we have kept track of the MLPs' individual classification accuracies (namely, the individual accuracies of $M_1, ..., M_k$), and they almost always unanimously outperform the curator's tree, M_{cur}. The lesson here is clear: without the interpretability of ML models, we run the risk that the human decision-maker might benefit very little from their adaption in the intended systems of use, even if these models all significantly outperform the human curator. This is precisely what we first illustrated with our very simple XOR separation problem in Section 3. This result helps us to understand the recent literature in explainable ML demonstrating that high stand-alone ML classification accuracy – in the sense of high specificity and sensitivity, for instance – does not always translate into better real-life performance when humans are part of the ultimate decision procedure [1, 6–8].

6 CONCLUSION

The lesson of this project is short and simple, despite the apparent complexity of the preceding discussion: we have explained – first theoretically, and second through a synthetic experiment – how a human-machine collaboration can benefit from an ML model's transparency. When an ML model is interpretable to a particular human decision maker, then that human decision maker can do better, in terms of overall accuracy, than they could as compared to taking advice from a model that is opaque to them. Somewhat surprisingly, this is true *even when* the opaque model is objectively better (i.e., more accurate). In the current literature on machine learning, one often finds an uncritical enthusiasm for the most elaborate state-of-the-art models, regardless of the underlying need for such complex algorithms. The motivating idea, we suspect, is that developers assume more accuracy is always better. But our results should temper such unbridled enthusiasm for overly complex models. Sometimes, as we have shown, less is more.

A APPENDIX

Synthetic Dataset Hyper-Parameters: Our synthetic dataset is constructed as uniform cube clusters around the 2^{10} lattice points in the 10-dimensional lattice $\{-1, 1\}^{10}$. Each uniform cube has a side length of 0.7. Half of the clusters are uniformly randomly assigned label 1 and the rest label 0. For each iteration of the experiment, the lattice labels are reassigned. In order to train our agent, interpretation, collaboration process, and lastly, testing, we need to synthetically generate four datasets:

- *Agent training dataset:* This dataset consists of 30 points per cluster, which amounts to 30720 labeled training data points for the MLPs, $M_1, ..., M_k$ and the tree M_{cur}.
- *Interpretation training dataset:* This dataset consists of 10 points per cluster, which amounts to 10240 data points for the interpretation decision tree to to fit the k-tuple of MLPs.
- *Collaboration training dataset:* This dataset consists of 10 points per cluster, which amounts to 10240 data points for fitting the collaboration protocol, which is a decision tree.
- *Testing dataset:* The dataset we use for testing – computing fidelity and the relevant accuracies – consists of 20 points per cluster, which amounts to 20480 test data points.

MLP Hyper-Parameters: Each one of $M_1, ..., M_k$ are MLPs with four fully connected layers, with respective in-features-out-features pair of $(10, 2048)$, $(2048, 1024)$, $(1024, 512)$, $(512, 2)$, and ReLU activation. The output is generated via the Softmax function. The MLPs are trained with a batch size of 10 for only 1 epoch and a flat learning rate of 0.1. Our experiment required minimal hyperparameter tuning. The choice of hyper-parameters is quite arbitrary. The only issue that concerned us was that we want to ensure that each MLP is trained to just above 80% accuracy. This would ensure that it is non-trivially good at solving the problem, but also imperfect enough to allow room for improvement to make the effects of interpretation and collaboration apparent.

Curator Decision Tree Hyper-Parameters: The curator decision trees are all trained using the entropy method with *max_tree_depth* of 10 and *min_samples_leaf* parameter of 6. The choice of these hyper-parameters, just like for the MLPs, is somewhat arbitrary. We wanted to ensure that (1) the classifier, interpretation and collaboration trees are trained with the same hyper-parameters, since they are intended to model the learning of the same curator, and, again (2) the classifier should achieve accuracy between 70% and 80%, so that it is sufficiently well-performing while allowing room for improvement and being outperformed by the individual MLPs.

The full experiment that we have performed here can be reproduced using the code available at the GitHub repository which we have made public (see [30]).

REFERENCES

[1] B. Babic, S. Gerke, T. Evgeniou, and I.G. Cohen. Beware Explanations from AI in Health Care. *Science*, 373(6552):284–286, 2021.

[2] Z.C. Lipton. The Mythos of Model Interpretability. *Queue*, 16(3):31–57, 2018.

[3] U. Johansson, C. Sönströd, U. Norinder, and H. Boström. Trade-off Between Accuracy and Interpretability for Predictive in Silico Modeling. *Future Medicinal Chemistry*, 3(6):647–663, 2011.

[4] A. Nesvijevskaia, S. Ouillade, P. Guilmin, and J. Zucker. The Accuracy Versus Interpretability Trade-off in Fraud Detection Model. *Data & Policy*, 3(12):1–20, 2021.

[5] S. Goethals, D. Martens, and T. Evgeniou. The Non-Linear Nature of the Cost of Comprehensibility. *Journal of Big Data*, 9(30):1–23, 2022.

[6] M. Jacobs, M. Pradier, T. McCoy, P. Roy, F. Doshi-Velez, and G. Krzysztof. How Machine Learning Recommendations Influence Clinician Treatment Selections. *Translational Psychiatry*, 11(108):1–9, 2021.

[7] B.J. Dietvorst, J.P. Simmons, and C. Massey. Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science*, 64(3):1155–1170, 2018.

[8] A.F. Markus, J.A. Kors, and P.R. Rijnbeek. The Role of Explainability in Creating Trustworthy Artificial Intelligence for Health Care. *Journal of Biomedical Informatics*, 113(103655):1–11, 2021.

[9] R. M. Hamm and S. L Smith. The Accuracy of Patients' Judgments of Disease Probability and Test Sensitivity and Specificity. *Journal of Family Practice*, 47:44–52, 1998.

[10] B. Babic, S. Gerke, T. Evgeniou, and I.G. Cohen. Direct-to-Consumer Medical Machine Learning and Artificial Intelligence Applications. *Nature Machine Intelligence*, 3:283–287, 2021.

[11] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, 67(23):1–23, 2017.

[12] Z. Lipton, J. McAuley, and A. Chouldechova. Does Mitigating ML's Impact Disparity Require Treatment Disparity? *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

[13] A. Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, 2017.

[14] A. Brown, A. Chouldechova, E. Putnam-Hornstein, A. Tobin, and R. Vaithianathan. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-Making in Child Welfare Services. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*, 2019.

[15] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science*, 366(6464):447–453, 2019.

[16] I. Glenn Cohen, Boris Babic, Sara Gerke, Xia Qiong, Theodoros Evgeniou, and Klaus Wertenbroch. How AI Can Learn from the Law: Putting Humans in the Loop only on Appeal. *Nature Digital Medicine*, 6(160), 2023.

[17] M.T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*, 2016.

[18] C. Rudin, C. Wang, and B. Coker. The Age of Secrecy and Unfairness in Recidivism Prediction. *Harvard Data Science Review*, 2(1), 2020.

[19] S. Gerke, B. Babic, T. Evgeniou, and G. Cohen. The Need for a System View to Regulate Artificial Intelligence/Machine Learning-Based Software as Medical Device. *Nature Digital Medicine*, 3(53), 2020.

[20] C. Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1:206–215, 2019.

[21] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin. Learning Certifiably Optimal Rule Lists for Categorical Data. *Journal of Machine Learning Research*, 18(234):1–78, 2018.

[22] K. Hornik, M. Stinchcombe, and H. White. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2(5):359–366, 1989.

[23] G. K. Dziugaite, S. Ben-David, and D. M. Roy. Enforcing Interpretability and its Statistical Impacts: Trade-Offs between Accuracy and Interpretability. *arXiv*, https://arxiv.org/abs/2010.13764, 2020.

[24] M.D. Abràmoff, P.T. Lavin, M. Birch, N. Shah, and J.C. Folk. Pivotal Trial of an Autonomous AI-Based Diagnostic System for Detection of Diabetic Retinopathy in Primary Care Offices. *Nature Digital Medicine*, 1(39), 2018.

[25] FDA. De Novo Classification Request for IDx-DR – Decision Summary, 2018.

[26] FDA. Evaluation of Automatic Class III Designation for OsteoDetect – Decision Summary, 2018.

[27] F. Cabitza, A. Campagner, and L. M. Sconfienza. Studying Human-AI Collaboration Protocols: the Case of Kasparov's Law in Radiological Double Reading. *Health Information Science and Systems*, 9(8), 2021.

[28] G. Kasparov and M. Greengard. *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins.* Perseus Books, USA, 2018.

[29] R.T. Clemen. Combining Overlapping Information. *Management Science*, 33(3):373–380, 1987.

[30] Authors and GitHub Link Omitted for Blind Review. Github Repository To be Adeed.

[31] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. 2015.