On Transparency of Machine Learning Models: A Position Paper

Yan Zhou[1*], **Murat Kantarcioglu**[1]
[1]Computer Science Department
University of Texas at Dallas
{yan.zhou2, muratk}@utdallas.edu

Abstract

An ongoing challenge in machine learning is to improve the transparency of learning models, helping end users to build trust and defend fairness and equality while protecting individual privacy and information assets. Transparency is a timely topic given the increasing application of machine learning techniques in the real world, and yet much more progress is needed in addressing the transparency issues. We propose critical research questions on transparency-aware machine learning on two fronts: *know how* and *know that*. Know-how is concerned with searching for a set of decision objects (e.g. functions, rules, lists, and graphs) that are cognitively fluent for humans to apply and consistent with the original complex model, while know-that is concerned with gaining more in-depth understanding of the internal justification of the decisions through external constraints on accuracy, consistency, privacy, reliability, and fairness.

1 Introduction

As human beings we make decisions everyday, sometimes explicitly and sometimes implicitly. In any case, we live our lives under relative constraints that give rise to *rational decision making*—neutralizing inevitable losses by gaining something we treasure more. Machine learning takes rational decision making to a whole new level. Rigorous mathematical models empower machine learning algorithms to search for optimal solutions for problems under constraints, often way more complicated than any human mind can comprehend. Despite the enormous potential of success in real life, making critical decisions with machine learning algorithms comes with many great challenges. The most apparent one is transparency. No one should feel comfortable when decisions, especially critical ones, are made by and within a black box. Without transparency, we cannot speak for machine learning algorithms in justifying the decisions they deduce and their ultimate consequences on trust, fairness, privacy, and security. But what is transparency?

When concerns about transparency of machine learning algorithms are raised, we often coarsely relate transparency to simplicity or understandability while neglecting the panoramic vision of the issue. Is transparency solely tied to human understandability? There are quite a few machine learning algorithms that directly or indirectly produce human comprehensible output, such as a linear model, a decision rule, or a decision list [Alvarez Melis and Jaakkola, 2018; Ribeiro *et al.*, 2016a; Datta *et al.*, 2016; Letham *et al.*, 2015]. Suppose we can trace the chain of reasoning of each decision such an algorithm makes, can we claim the algorithm is transparent? The answer is unfortunately no. The chain of reasoning only tells us "how" a decision is made for a given input but not "that"—the justification that is only accessible internally. Knowing "how" is not sufficient for justifying that the decision is made consistently, accurately, reliably, and validly. Traditional epistemology makes a clear distinction between "know how" (understand an action) and "know that" (understand a concept). A learning model is truly transparent only when we know both "how" and "that". Unfortunately, "know that" is often an esoteric exercise and requires years of training. However, a learning model can provide the insights into "know that" by the justification of decisions that can be gauged externally. For example, a transparent model can supply what evidence supports the decisions, whether it makes the consistent decisions across the entire distribution, whether it is more susceptible to adversarial attacks, and whether it gives away private information.

2 Model Transparency

For any given input, a transparent learning model not only supplies chains of reasoning such as linear functions, graphs, lists of rules, but provides the justification of decisions in terms of accuracy, consistency, reliability, and security. Together the "know how" and "know that" elements of a transparent model serve as a technical-cognitive prosthesis between human and machine. In this section, we discuss each element and potential research challenges.

2.1 Transparency—"Know How"

Extensive research has been done to increase the interpretability of different types of classification models. Letham et. al. (2013) use decision lists to simplify a high-dimensional, multivariate feature space. Martens and

*Contact Author

Provost (2014) define explanation as a minimal set of words that explains the class membership of a document. Martens et. al. (2011) propose an assessment for the overall performance of classification models from a user perspective in terms of accuracy, comprehensibility, and justifiability. Lim et. al. (2009) examine different types of explanations for improving transparency of rule based systems. More recently, Ribeiro et. al. (2016b) present a sparse linear model (LIME) for local exploration—providing interpretable representation locally faithful to the classifier. Datta et al. proposed a family of Quantitative Input Influence (QII) that can be used to measure the influence of the inputs of a decision making model on its output [Datta *et al.*, 2016]. These QII measures can be used to produce transparency reports on the inputs that were most influential on the model.

In general, the problem of "know how" is to search for a set of decision objects that is cognitively fluent for human to follow. This set of decision objects serves as a transparent substitute for the original complex and possibly black-box decision model. Given a data collection $\mathcal{D}$ and a decision model $\mathcal{H}$ built on $\mathcal{D}$, a transparent model $\mathcal{T}$ can be built from $\mathcal{H}$ consisting of a set of interpretable decision-making objects $\mathcal{R}$ such as decision rules, local linear models, or feature weights. As a transparent model $\mathcal{T}$ has to satisfy two requirements: *consistency* and *coverage*. First, the set of decision objects in the transparent model $\mathcal{T}$ must faithfully represent the decision model $\mathcal{H}$. Since statistical reasoning techniques used to build $\mathcal{H}$ and $\mathcal{T}$ are likely different, the decisions induced by the two models may be inconsistent for some input. Therefore, one research challenge is to search for a subset of decision objects in $\mathcal{R}$ that is consistent with $\mathcal{H}$. In addition to consistency, we also need to consider the coverage of the decision objects in the transparent model $\mathcal{T}$. An input x is said to be covered by $\mathcal{T}$ if there is at least one decision object in $\mathcal{T}$ that can be applied to x. A transparent model is said to have a λ-*good* coverage for any given x if x is covered by $\mathcal{T}$ with a probability of λ. The greater the value of λ, the greater the coverage. Therefore, the research challenge is to devise a transparent model with a bounded λ-good coverage for any x with minimum inconsistency.

Let $\mathcal{R} = \{R_1, \ldots, R_m\}$ given a transparent model $\mathcal{T}$. Given an input $x \in \mathcal{D}$, the decision d output by $\mathcal{H}$ can be mapped to a subset of $\mathcal{R}$ with a probability. Let $\mathbb{1}_{R_s}(x)$ be a function that returns a boolean vector that indicates which decision objects in $\mathcal{R}$ can be applied to a given input x. If a subset $R_s \subseteq \mathcal{R}$ has a boolean value of "1", x is covered by R_s; otherwise, x is not covered and the prediction for x is reduced to the class prior. For all examples $\{(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)\}$ in $\mathcal{D}$, where $y_i = h(x_i)$ is the decision made by $\mathcal{H}$ for the input x_i, the general optimization problem is as follows:

$$\begin{aligned} \underset{R_s \subseteq \mathcal{R}}{\arg\min} \quad & \textstyle\sum_{i=1}^{n} \mathcal{L}(\mathcal{T}(R_s, x_i), y_i) \\ subject\ \ to \quad & \textstyle\sum_{i=1}^{n} \mathbb{1}_{R_s}(x_i) \geq \lambda n \end{aligned}$$

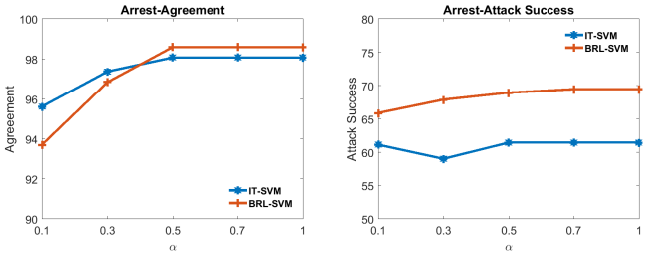
where $\mathcal{L}$ is the loss function that measures the inconsistency between $\mathcal{T}$ and $\mathcal{H}$ for a given x. The solution to the above optimization problem is composed of a subset of $\mathcal{R}$ that satisfies a predefined coverage with minimum inconsistency.

2.2 Transparency—"Know That"

The internal justification for decisions made by machine learning models is difficult for end users to grasp. An indirect approach to justifying machine-made decisions can be conceived by enforcing constraints on privacy, reliability, and fairness. These are commonly raised concerns on legal rights of individuals (e.g., HIPAA 1996), safety of autonomous vehicles, and fair procedure in credit scoring systems.

Privacy Model transparency can be both beneficial and pernicious. While greater transparency in the decision processes can help users better understand how decisions are reached inside a machine learning model, it may also introduce and exacerbate biases and privacy/security risks.

Figure 1 demonstrates one of the potential challenges we have to face: the trade-off between model consistency and privacy in the presence of inversion attacks [Fredrikson *et al.*, 2014]. Recall that *consistency* measures the extent to which the transparent model agrees with the original machine learning model on the prediction for a given x. Inversion attacks refer to malicious attempts for reverse engineering sensitive information embedded in the data used to train the machine learning models.



(a) (b)

Figure 1: Trade-off between model consistency and attack success rate on the *Arrest* dataset. α is a hyperparameter that controls the tolerance for privacy leak.

The two transparency models in the figure are IT-SVM [Alufaisan *et al.*, 2017] and the BRL transparency model [Letham *et al.*, 2015], tested on the *Arrest* dataset [nls, 2017]. For both transparency models the attack success rate increases as the model consistency grows. The empirical demonstration reveals one scenario of the potentially conflicting objectives: consistency, coverage and privacy protection. A transparent model with better consistency and coverage may have a higher risk of privacy violation. Therefore, the research challenge is to find the transparent model that strikes the right balance between increasing coverage and consistency and reducing the risk of privacy violation. One potential solution to this problem is to model the problem as a multi-objective optimization task and explore multiple heuristics such as using the weighted sum approach with a convex combination of objectives to solve the multi-objective optimization problem.

Reliability *Reliability* is concerned with the robustness of a machine model when it faces adversarial attacks. It has been shown that standard machine learning techniques are

susceptible to adversarial attacks [Goodfellow *et al.*, 2014; Szegedy *et al.*, 2013; Kantarcioglu *et al.*, 2011; Zhou *et al.*, 2012; Papernot *et al.*, 2016]. One of the important lines of attack against standard machine learning techniques is called evasion attack (e.g., [Dalvi *et al.*, 2004a; Lowd and Meek, 2005a]). In an evasion attack, given a classifier C, an instance (x, y) where x is the feature vector for the instance (e.g., PDF malware related features) and y is the class value (e.g., y='malware') controlled by the attacker, x can be modified to x' by the attacker such that

$$\begin{aligned} &\arg\min_{x'} && d(x, x') \\ &subject\ \ to && x' \in F_y, C(x') = t \end{aligned}$$

for a set of feasible instances F_y[1], some domain specific distance function d, and the target class t (e.g., for a PDF malware, the target class would be "benign PDF file"). There are two unique research challenges regarding the robustness of a transparent model: 1.) whether the existing threat from evasion attacks can be exacerbated by the additional information provided by the transparent model; and 2.) for successful evasion attacks, how to enhance the transparent models to reduce the effectiveness of the attacks.

To address the question whether it is possible that a transparent model would make evasion attacks more successful and reduce the cost of finding x', we can start by understanding the impact of releasing the set of decision objects R. For a given instance x, the attacker can find the decision rules in R that match x and its target class value t. The main research challenge would be to define appropriate matching function M such that given x and R, $M(x, R)$ returns a suitable set of rules that can be used for evasion attacks. A good candidate of M should also be able to model attackers' capabilities. If an attacker can modify some features more easily than others (e.g., the number of bytes in a certain section of the PDF malware can be modified more easily than control flow relate features), the function M can give more weights to the rules matching such features. Once we identify the set of matching rules, we can find the minimum perturbation that turns x to x'. We can use the feature modification costs as a proxy and define a distance function $d(x, x')$. The existence of rules $M(x, R)$ will allow us to find the desired x' by limiting the search space and the number of queries that needs to be issued to the classifier C.

To defend against such attacks, the first step is to explore whether the existing attack-resilient classifiers (e.g., [Dalvi *et al.*, 2004b; Lowd and Meek, 2005b; Zhou *et al.*, 2012; Bruckner and Scheffer, 2009; Bruckner and Scheffer, 2011; Kantarcioglu *et al.*, 2011]) remain robust against the attacks given the transparent models. The next step is to explore more advanced techniques such as modifying rules (e.g., making them less general), deleting rules (e.g., deleting some rules that are useful in the attacks) or adding some fake rules that may reduce the effectiveness of the transparency model based evasion attacks.

[1] In the case of image processing, if initially x was classified as y where y= 'truck', we may want to find x' that can be still recognized as a truck *by a human* so that the attack may not be captured by the human eye. In other domains, there may be other constraints. For example, for a PDF file that contains malware, the attacker may want to modify the malware in such a way that the modified PDF should be a valid PDF file.

Fairness Fairness is concerned with whether a transparent model is "fair" for a protected or sensitive group. There are two scenarios to consider: 1.) is the bias in the original decision model transferable to its transparent counterpart? 2.) is there a trade-off between transparency and fairness?

Given data $X \in \mathbb{R}^n$, labels $Y \in \{0, 1\}$, and sensitive attributes $A \in \{0, 1\}$, the goal of fair learning is to predict outcomes that are accurate with respect to Y but fair with respect to A. The formulation of loss functions often depends on the definitions of fairness. Below are the common ones:

- Demographic Parity: $P(\hat{Y} = 1|A) = P(\hat{Y} = 1)$
- Equalized Odds: $P(\hat{Y} \neq Y|A = 0, Y = y) = P(\hat{Y} \neq Y|A = 1, Y = y), \forall y \in \{0, 1\}$
- Equalized Opportunity: $P(\hat{Y} \neq Y|A = 0, Y = 1) = P(\hat{Y} \neq Y|A = 1, Y = 1)$

where $\hat{Y}$ is the prediction made by a learning model. Let f be the original model function and g be the transparent counterpart of f. To learn a fair transparent model, we can consider an adversarial predictor a that aims to predict the sensitive attributes. Let $L_Y(f(X), g(X))$ be the transparency loss for mimicking the prediction of Y and $L_A(a(g(X)), A)$ be the adversary's loss for predicting A given g. To learn a fair transparent model, we can define an objective function that minimizes the inconsistency between f and g while maximizing the loss for the adversary:

$$\min \left[\sum_{x \in X} L_Y(f(x), g(x)) - \sum_{x \in X, A} L_A(a(g(x)), A) \right]$$

Existing research [Edwards and Storkey, 2016; Kim *et al.*, 2019; Madras *et al.*, 2018; Zhang *et al.*, 2018; Beutel *et al.*, 2017] demonstrates some promising results on de-biasing learning models. However, it is often the case that a better fairness measure is obtained at the cost of model accuracy. It is not uncommon that fairness on sensitive group is achieved through hurting the accuracy on the non-sensitive group. Therefore, to learn a fair transparent model, the major research challenge would be to improve fairness without hurting the normal groups. More importantly, transparent models should be self-censored so that transparency is not supplied in a harmful way that can be easily explored by the adversary.

3 Conclusions

In this paper, we address the importance of improving transparency in machine learning models. We discuss what it means by transparency and what are the major challenges in the process of making machine learning models transparent. We propose potential solutions to some of the challenges.

4 Acknowledgement

This work was supported by ARO award W911NF-17-1-0356 and NSF IIS-1939728.

References

[Alufaisan *et al.*, 2017] Yasmeen Alufaisan, Yan Zhou, Murat Kantarcioglu, and Bhavani Thuraisingham. From myths to norms: Demystifying data mining models with instance-based transparency. In *In Proceedings of the 2017 IEEE 3rd International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 2017.

[Alvarez Melis and Jaakkola, 2018] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7775–7784. Curran Associates, Inc., 2018.

[Beutel *et al.*, 2017] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *CoRR*, abs/1707.00075, 2017.

[Bruckner and Scheffer, 2009] M. Bruckner and T. Scheffer. Nash equilibria of static prediction games. In *Advances in Neural Information Processing Systems*. MIT Press, 2009.

[Bruckner and Scheffer, 2011] M. Bruckner and T. Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2011.

[Centers for Medicare & Medicaid Services, 1996] Centers for Medicare & Medicaid Services. The Health Insurance Portability and Accountability Act of 1996 (HIPAA). Online at http://www.cms.hhs.gov/hipaa/, 1996.

[Dalvi *et al.*, 2004a] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108, New York, NY, USA, 2004. ACM Press.

[Dalvi *et al.*, 2004b] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 99–108, New York, NY, USA, 2004. ACM.

[Datta *et al.*, 2016] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *In Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617, 2016.

[Edwards and Storkey, 2016] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference in Learning Representations (ICLR2016)*, 2 2016.

[Fredrikson *et al.*, 2014] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 17–32, San Diego, CA, 2014. USENIX Association.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014.

[Kantarcioglu *et al.*, 2011] Murat Kantarcioglu, Bowei Xi, and Chris Clifton. Classifier evaluation and attribute selection against active adversaries. *Data Min. Knowl. Discov.*, 22:291–335, January 2011.

[Kim *et al.*, 2019] Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 247–254, New York, NY, USA, 2019. Association for Computing Machinery.

[Letham *et al.*, 2013] Benjamin Letham, Cynthia Rudin, Tyler H. Mccormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model, 2013.

[Letham *et al.*, 2015] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Ann. Appl. Stat.*, 9(3):1350–1371, 09 2015.

[Lim *et al.*, 2009] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 2119–2128, New York, NY, USA, 2009. ACM.

[Lowd and Meek, 2005a] Daniel Lowd and Christopher Meek. Adversarial learning. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 641–647, New York, NY, USA, 2005. ACM Press.

[Lowd and Meek, 2005b] Daniel Lowd and Christopher Meek. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 641–647, 2005.

[Madras *et al.*, 2018] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3384–3393. PMLR, 10–15 Jul 2018.

[Martens and Provost, 2014] David Martens and Foster Provost. Explaining data-driven document classifications. *MIS Q.*, 38(1):73–100, March 2014.

[Martens *et al.*, 2011] David Martens, Jan Vanthienen, Wouter Verbeke, and Bart Baesens. Performance of classification models from a user perspective. *Decis. Support Syst.*, 51(4):782–793, November 2011.

[nls, 2017] National longitudinal surveys home page, 2017.

[Papernot *et al.*, 2016] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik,

and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016*, pages 372–387, 2016.

[Ribeiro *et al.*, 2016a] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM.

[Ribeiro *et al.*, 2016b] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *SIGKDD*, pages 1135–1144, 2016.

[Szegedy *et al.*, 2013] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

[Zhang *et al.*, 2018] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA, 2018. Association for Computing Machinery.

[Zhou *et al.*, 2012] Yan Zhou, Murat Kantarcioglu, and Bhavani M. Thuraisingham. Sparse bayesian adversarial learning using relevance vector machine ensembles. In Mohammed Javeed Zaki, Arno Siebes, Jeffrey Xu Yu, Bart Goethals, Geoffrey I. Webb, and Xindong Wu, editors, *ICDM*, pages 1206–1211. IEEE Computer Society, 2012.