On Mismatched Detection and Safe, Trustworthy Machine Learning

Kush R. Varshney
IBM Research
Yorktown Heights, NY, USA
krvarshn@us.ibm.com

***Abstract*—Instilling trust in high-stakes applications of machine learning is becoming essential. Trust may be decomposed into four dimensions: basic accuracy, reliability, human interaction, and aligned purpose. The first two of these also constitute the properties of safe machine learning systems. The second dimension, reliability, is mainly concerned with being robust to epistemic uncertainty and model mismatch. It arises in the machine learning paradigms of distribution shift, data poisoning attacks, and algorithmic fairness. All of these problems can be abstractly modeled using the theory of mismatched hypothesis testing from statistical signal processing. By doing so, we can take advantage of performance characterizations in that literature to better understand the various machine learning issues.**

***Index Terms*—signal detection theory, data poisoning, adversarial robustness, distribution shift, fairness**

I. Introduction

Despite artificial intelligence's promise to reshape different sectors, there has not yet been wide adoption of the technology except in certain pockets such as electronic commerce and media. Like other general purpose technologies, there are many short-term costs to the changes required in infrastructure, organizations, and human capital [1]. In particular, it is difficult for many businesses to collect and curate data from disparate sources. Importantly, there is a *lack of trust* in artificial intelligence and machine learning in critical enterprise workflows. For example, a study of business decision makers released in 2018 found that only 21% of them have a high level of trust in different types of analytics; the number is likely smaller for machine learning, which is a part of analytics in business parlance [2]. Trust is particularly important in high-stakes decision making applications such as health care, criminal justice, lending, and employment.

But what is trust and how should we model trustworthy machine learning? In the remainder of this paper, we will define trust and trustworthiness both in general and in the context of machine learning, relate it to safety in machine learning, and discuss how to model it using the theory of mismatched detection.

II. Trust

The concept of trust is defined and studied in many different fields including philosophy, psychology, sociology, economics, and organizational management. Trust is the relationship between a *trustor* and a *trustee*: the trustor trusts the trustee. A definition of trust from organizational management is particularly appealing and relevant for defining trust in machine learning because machine learning systems in high-stakes applications are typically used within organizational settings. Trust is defined in [3] to be:

> The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party.

This definition can be put into practice as a foundation for desiderata of machine learning systems.

Embedded within this definition is the idea that the trustee has certain properties that make it *trustworthy*, i.e. the qualities by which the trustor can expect the trustee to perform the important action. Being trustworthy does not automatically imply that the trustee is trusted. The trustor must consciously make a decision to be vulnerable to the trustee based on its trustworthiness and other factors. It is possible for a system to not be trusted no matter how worthy of trust it is.

In much of the literature on the topic, both the trustor and the trustee are people. For our purposes, however, an end-user or other person is the trustor and the machine learning system is the trustee. Although the specifics may differ, there are not many differences between a trustworthy person and a trustworthy machine learning system.

Building upon the above definition of trust and trustworthiness, one can list many different attributes of a trustworthy person: availability, competence, consistency, discreetness, fairness, integrity, loyalty, openness, promise fulfilment, and receptivity to name a few [4]. Similarly, one can list several attributes of a trustworthy information system, such as: correctness, privacy, reliability, safety, security, and survivability [5]. The 2019 International Conference on Machine Learning listed the following topics under trustworthy machine learning: adversarial examples, causality, fairness, interpretability, privacy-preserving statistics and machine learning, and robust statistics and machine learning. The European Commission's High Level Expert Group on Artificial Intelligence listed the following attributes in 2019: lawful, ethical, and robust (both technically and socially).

Such long and disparate lists give us some sense of what people deem to be trustworthy characteristics, but are difficult to use as anything but a rough guide. However, we can

TABLE I
ATTRIBUTES OF TRUSTWORTHY PEOPLE AND ARTIFICIAL INTELLIGENCE

Source	Attribute 1	Attribute 2	Attribute 3	Attribute 4
[6]	competent	reliable	open	concerned
[7]	credibility	reliability	intimacy	low self-orientation
[8]	competent	use fair means to achieve its goals	take responsibility for all its impact	motivated to serve others' interests as well as its own
[9]	ability	integrity	predictability	benevolence
[10]	technical competence	reliability	understandability	personal attachment

distill these attributes into a set of *separable* sub-domains that provide an organizing framework for trustworthiness. Several pieces of work converge onto a nearly identical set of four such separable attributes; a selected listing is provided in Table I [6]–[10]. The first three rows of the table are attributes of trustworthy people. The last two rows are attributes of trustworthy artificial intelligence. Importantly, through separability, it is implied that each of the qualities is conceptually different and we can examine each of them in isolation of each other.

In considering the four attributes of trustworthiness from Table I in the machine learning context, we take Attribute 1 to be basic performance such as accuracy, Attribute 2 to include reliability, safety, security and fairness, Attribute 3 to consist of various aspects of human interaction with the machine learning system and its openness (including interpretability), and Attribute 4 to be the alignment of the machine learning system's purpose with a society's wants.

We use the following working definition of trustworthy machine learning in the remainder of the paper. A trustworthy machine learning system is one that has sufficient:

1) basic performance,
2) reliability,
3) human interaction, and
4) aligned purpose.

III. SAFETY AND MODEL MISMATCH

As we look at the first two elements of the definition of trustworthy machine learning, we see that they recapitulate the definition of safety in machine learning that we proposed in earlier work [11], [12]: minimizing both risk and epistemic uncertainty, i.e. the probability of expected harms and the possibility of unexpected harms. Risk minimization is the central tenet of statistical machine learning and yields basic performance.

Uncertainty is the state of current knowledge in which something is not known. There are at least two types of uncertainty: *aleatoric uncertainty* and *epistemic uncertainty* [13], [14]. Aleatoric uncertainty, also known as statistical uncertainty, is inherent randomness or stochasticity in an outcome that cannot be further reduced. Etymologically derived from dice games, aleatoric uncertainty is used to represent phenomena such as vigorously flipped coins and vigorously rolled dice, thermal noise, and quantum mechanical effects.

Epistemic uncertainty, also known as systematic uncertainty, refers to knowledge that is not known in practice, but could be known in principle. The acquisition of this knowledge would reduce the epistemic uncertainty. As an example of epistemic uncertainty, [13] presents a person who does not know whether *kichwa* is *head* or *tail* in Kiswahili. This uncertainty can be reduced by observing coins being tossed in Nairobi.

Epistemic uncertainty is not part of the typical formulation of training machine learning models. The most common instance of epistemic uncertainty to arise in machine learning is a mismatch between the data distribution of the given training set and an unknown ideal distribution or true distribution that will be encountered at the time of inference by the model. The training and test data are not identically distributed with the desired distribution.

Mismatch can take several forms, such as distribution shift, data poisoning, and algorithmic fairness. In a distribution shift setting, we train on a dataset drawn from the past but would have liked to have trained on a dataset from the present. In a data poisoning adversarial attack (including label modification, data injection, and data modification), we train on a corrupted dataset but would have liked to have trained on a dataset that has not been attacked. Similarly, in case of algorithmic fairness, we are given a training dataset that includes unwanted biases that yield systematic disadvantages to certain groups and individuals (defined by protected attributes such as race and gender), but we would like to train on a dataset that represents a fair and just world in which those biases are not present.

Methods for achieving out-of-distribution generalization to deal with the brittleness of machine learning models in new contexts arising from spurious correlations [15], [16], for defending against data poisoning attacks [17], [18], and for mitigating unwanted bias [19], [20] are active areas of research, but typically dealt with separately. We contend that all of these problems share enough similarity that it is useful to view them abstractly through a single lens in order to gain intuition about them. There are several different possible approaches for such abstract modeling, including causal modeling. In the remainder of the paper, we focus on mismatched hypothesis testing from the statistical signal processing and information theory literature as the modeling approach.

IV. MISMATCHED HYPOTHESIS TESTING

Hypothesis testing is the common framework for posing the signal detection task in signal processing and information theory. The standard supervised classification problem in machine learning is a version of hypothesis testing when we have access to a finite number of samples as a dataset rather than full access to the probability distributions governing the hypotheses. If we take the limit as the number of samples goes to infinity, also known as the population setting of machine learning, we can analyze supervised classification as hypothesis testing. For simplicity, let us consider *binary* classification and hypothesis testing.

Let the features or observations be X and the labels or hypotheses be $Y \in \{0, 1\}$. The overall joint distribution $p_{X,Y}$ is broken down into the prior probabilities of the hypotheses $\pi_0 = Pr(Y = 0)$ and $\pi_1 = Pr(Y = 1)$, and likelihood functions $p_0(x) = p_{X|Y}(x \mid Y = 0)$ and $p_1(x) = p_{X|Y}(x \mid Y = 1)$. The Bayes-optimal detection rule to predict $\hat{y}$ from x is the likelihood ratio test:

$$\frac{p_1(x)}{p_0(x)} \quad \substack{>^{\hat{y}(x)=1} \\ <_{\hat{y}(x)=0}} \quad \frac{\pi_0}{\pi_1}. \tag{1}$$

In mismatched hypothesis testing [21]–[23], we do not have access to the ideal distribution $p_{X,Y}$, but only get to see some other distribution $\tilde{p}_{X,Y}$ and must perform a likelihood ratio test on $\tilde{p}_1(x)/\tilde{p}_0(x)$. As discussed before, $\tilde{p}_{X,Y}$ can refer to a past distribution that has shifted, a poisoned distribution, or a distribution containing unwanted biases against protected groups. In all of these cases, the ideal distribution $p_{X,Y}$ is not known in practice and this mismatch thus constitutes epistemic uncertainty.

The performance of the optimal likelihood ratio test is characterized using the Chernoff information $C(p_0, p_1)$: the maximum error exponent of the Bayesian error probability [24]. The work of [25] proposes and defines a *generalized* Chernoff information $C(p_0 \to \tilde{p}_0, p_1 \to \tilde{p}_1)$ as the maximum error exponent of the Bayesian error probability of the mismatched likelihood ratio test.

Chernoff information-based characterizations can be used to provide insights on various aspects of trustworthy machine learning. For example, the work of [26] uses this mismatched hypothesis testing framework in the space of algorithmic fairness to precisely show when there is and when there is not a tradeoff between fairness and accuracy. The prevailing wisdom is that there is always such a tradeoff, but that wisdom is predicated on measuring the probability of error with respect to $\tilde{p}_{X,Y}$. When the probability of error is measured with respect to $p_{X,Y}$, the trade-off disappears. Similarly, although it has not been done, one can use generalized Chernoff information to characterize fundamental limits in out-of-distribution generalization and adversarial robustness.

The third element of trust: human interaction, can be related to explainable and interpretable machine learning. If the interaction between a machine learning model and a human is modeled as a two-node distributed detection system where the human is a fusion center, then Chernoff information-based analysis shows that increased explainability (more information passing from machine learning to human) yields increased overall system accuracy [27]. This result contradicts a different prevailing wisdom: that there is a tradeoff between explainability and accuracy. One can imagine combining the analyses for standard detection, mismatched detection, and distributed detection to come up with an overall unified characterization for the first three elements of trustworthy machine learning.

V. Conclusion

Trustworthy machine learning is starting to become an important topic of study. In this paper, we have defined the elements of trust and highlighted the fact that several important issues in making machine learning trustworthy and safe can be traced to reliability and robustness when the training data is not identically distributed to an unknown desired distribution. This mismatch yields epistemic uncertainty that must be minimized.

We have discussed how the machine learning paradigms of being robust to unknown distribution shift, defending against data poisoning attacks, and mitigating unwanted biases for fairness, are all examples of mismatch and can be abstractly characterized using the theory of mismatched detection and generalized Chernoff information. Pulling together such characterizations along with similar characterizations of other elements of trust, such as explainability, may yield an abstract unified theory of trust that will allow us to better understand which elements are in conflict and which ones can be simultaneously satisfied. Understanding these relationships will also help system designers take input from policymakers and other stakeholders in setting parameters for machine learning systems that respect the values they desire.

This paper is a call to action that does not introduce any new technical results. It sets forth a future technical research agenda that is grounded in defining trust for machine learning from the synthesis of existing literature in different fields such as organizational management and capturing different considerations of machine learning performance that go beyond predictive accuracy in a single abstract framework.

Acknowledgment

The author thanks Pin-Yu Chen, Sanghamitra Dutta, Sijia Liu, Dennis Wei, and Hazar Yüksel for discussions.

References

[1] B. Bergstein, “This is why AI has yet to reshape most businesses,” *MIT Technology Review*, Feb. 2019. [Online]. Available: https://www.technologyreview.com/s/612897/this-is-why-ai-has-yet-to-reshape-most-businesses

[2] M. Korolov, “Explainable AI: Bringing trust to business AI adoption,” *CIO*, Sep. 2019. [Online]. Available: https://www.cio.com/article/3440071/explainable-ai-bringing-trust-to-business-ai-adoption.html

[3] R. C. Mayer, J. H. Davis, and F. D. Schoorman, “An integrative model of organizational trust,” *Academy of Management Review*, vol. 20, no. 3, pp. 709–734, Jul. 1995.

[4] G. Dietz and D. N. Den Hartog, “Measuring trust inside organisations,” *Personnel Review*, vol. 35, no. 5, pp. 557–588, Sep. 2006.

[5] F. B. Schneider, Ed., *Trust in Cyberspace*. Washington, DC, USA: National Academy Press, 1999.

[6] A. K. Mishra, *Organizational Responses to Crisis: The Centrality of Trust*. Newbury Park, California, USA: Sage, 1996, pp. 261–287.

[7] D. H. Maister, C. H. Green, and R. M. Galford, *The Trusted Advisor*. New York, New York, USA: Touchstone, 2000.

[8] S. J. Sucher and S. Gupta, “The trust crisis,” *Harvard Business Review*, Jul. 2019. [Online]. Available: https://hbr.org/cover-story/2019/07/the-trust-crisis

[9] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. Gonzalez Zelaya, and A. van Moorsel, “The relationship between trust in AI and trustworthy machine learning technologies,” arXiv:1912.00782, 2019.

[10] M. Ashoori and J. D. Weisz, “In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes,” arXiv:1912.02675, 2019.

[11] K. R. Varshney, “Engineering safety in machine learning,” in *Proceedings of the Information Theory and Applications Workshop*, La Jolla, California, USA, Feb. 2016.

[12] K. R. Varshney and H. Alemzadeh, “On the safety of machine learning: Cyber-physical systems, decision sciences, and data products,” *Big Data*, vol. 5, no. 3, pp. 246–255, Sep. 2017.

[13] E. Hüllermeier and W. Waegeman, “Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction,” arXiv:1910.09457, 2019.

[14] J. Z. Liu, J. Paisley, M.-A. Kioumourtzoglou, and B. A. Coull, “Accurate uncertainty estimation and decomposition in ensemble learning,” in *Advances in Neural Information Processing Systems*, Vancouver, British Columbia, Canada, Dec. 2019.

[15] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, “Invariant risk minimization,” arXiv:1907.02893, 2019.

[16] D. Arpit, C. Xiong, and R. Socher, “Predicting with high correlation features,” arXiv:1910.00164, 2019.

[17] P. W. Koh, J. Steinhardt, and P. Liang, “Stronger data poisoning attacks break data sanitization defenses,” arXiv:1811.00741, 2018.

[18] S. Liu, S. Lu, X. Chen, Y. Feng, K. Xu, A. Al-Dujaili, M. Hong, and U.-M. O’Reilly, “Min-max optimization without gradients: Convergence and applications to adversarial ML,” arXiv:1909.13806, 2019.

[19] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney, “Fairness GAN: Generating datasets with fairness properties using a generative adversarial network,” *IBM Journal of Research and Development*, vol. 63, no. 4/5, p. 3, Jul./Sep. 2019.

[20] D. Wei, K. N. Ramamurthy, and F. P. Calmon, “Optimized score transformation for fair classification,” in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Palermo, Italy, Jun. 2020.

[21] P. J. Huber, “A robust version of the probability ratio test,” *Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, Dec. 1965.

[22] D. Kazakos, “Signal detection under mismatch,” *IEEE Transactions on Information Theory*, vol. IT-28, no. 4, pp. 681–684, Jul. 1982.

[23] S. A. Kassam and H. V. Poor, “Robust techniques for signal processing: A survey,” *Proceedings of the IEEE*, vol. 73, no. 3, pp. 433–481, Mar. 1985.

[24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, New Jersey, USA: John Wiley & Sons, 2006.

[25] Y. Lee and Y. Sung, “Generalized Chernoff information for mismatched Bayesian detection and its application to energy detection,” *IEEE Signal Processing Letters*, vol. 19, no. 11, pp. 753–756, Nov. 2012.

[26] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. R. Varshney, “An information-theoretic perspective on the relationship between fairness and accuracy,” arXiv:1910.07870, 2019.

[27] K. R. Varshney, P. Khanduri, P. Sharma, S. Zhang, and P. K. Varshney, “Why interpretability in machine learning? An answer using distributed detection and data fusion theory,” in *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*, 2018, pp. 15–20.