PERSPECTIVE
https://doi.org/10.1038/s42256-022-00445-z

nature machine intelligence

Stable learning establishes some common ground between causal inference and machine learning

Peng Cui[1,2] ✉ and Susan Athey[3]

Causal inference has recently attracted substantial attention in the machine learning and artificial intelligence community. It is usually positioned as a distinct strand of research that can broaden the scope of machine learning from predictive modelling to intervention and decision-making. In this Perspective, however, we argue that ideas from causality can also be used to improve the stronghold of machine learning, predictive modelling, if predictive stability, explainability and fairness are important. With the aim of bridging the gap between the tradition of precise modelling in causal inference and black-box approaches from machine learning, stable learning is proposed and developed as a source of common ground. This Perspective clarifies a source of risk for machine learning models and discusses the benefits of bringing causality into learning. We identify the fundamental problems addressed by stable learning, as well as the latest progress from both causal inference and learning perspectives, and we discuss relationships with explainability and fairness problems.

Machine learning has been incorporated to make predictions within a wide variety of digital services, ranging from search engines to e-commerce to social media platforms, thereby nurturing the booming digital economy. In these scenarios, the prediction accuracy and efficiency of machine learning techniques are the objectives of optimization, but the potential risks from erroneous predictions are less important. For applications such as predicting clicks or classifying images, models can be updated frequently, and errors are not too costly. Thus, these application areas are well-suited to black-box techniques combined with ongoing performance monitoring.

Over recent years, however, machine learning has been applied in a wider variety of domains, even entering high-stakes areas such as healthcare, industrial manufacturing, financing and the administration of justice. In these areas, mistakes made by machine learning algorithms may bring tremendous risks, and mistakes have substantial consequences for social issues such as safety, ethics and justice—especially when algorithmic predictions play substantial roles in a decision-making process. In such settings, the environment may change more quickly than the model is updated, and properties beyond short-term predictive performance become increasingly important.

In particular, we regard the lack of stability, explainability and fairness guarantees as the most critical and urgent factors that must be addressed in today's machine learning.

Key factors to address

Stability. Predicting future outcome values on the basis of their observed features using a model estimated on a training dataset is a standard machine learning problem. Many learning algorithms have been proposed and shown to be successful when the test data and training data come from the same distribution. However, the best-performing models for a given distribution of training data typically exploit subtle statistical associations among features, making them potentially more prone to prediction error when applied to test data whose distribution differs from that of the training data. In real applications, a distribution shift from training to testing is often inevitable; the consequence is that machine learning algorithms will have unstable performance when applied in different testing environments with unknown distribution shifts, thus making them unreliable.

Explainability. Owing to the high complexity encountered in many applications, it is not realistic to expect machine learning algorithms to produce predictive or prescriptive results with such high accuracy that humans can rely on them. Particularly in high-stakes areas, or in settings where it is difficult to quantify all of the important consequences of a decision, it may be preferable to let humans remain in the decision loop and serve as the final gatekeeper[1]. This necessitates a common language for the algorithms and humans to understand and collaborate. Most off-the-shelf machine learning models at present are black-box models: both the algorithmic process and the prediction results cannot easily be explained to humans. Although a strand of research on explainable AI exists, most studies try to partially explain black-box models, rather than design inherently interpretable models [2].

Fairness. With the recent trend of applying machine learning to societal problems, fairness issues have raised significant concerns among researchers and the public. Mainstream machine learning algorithms could amplify the bias existing in data, which could result in 'unfair' outcomes. For example, COMPAS is a widely used tool in US courts to judge whether a defendant will commit a crime in the future. However, it was reported to produce higher false positive rates for Black defendants than for white defendants—a finding widely interpreted as unfair to Black defendants[3]. This is only one of the many cases where machine learning may negatively impact social outcomes if fairness concerns are not sufficiently addressed.

In the following, we discuss some key drivers of risks like these, as well as the opportunities and challenges for the ideas from causal inference to address them. We then introduce the development of stable learning with the goal of finding the common ground between causal inference and machine learning and its implications for addressing explainability and fairness problems.

[1]Department of Computer Science, Tsinghua University, Beijing, China. [2]Beijing Academy of Artificial Intelligence, Beijing, China. [3]Graduate School of Business, Stanford University, Stanford, CA, USA. ✉e-mail: cuip@tsinghua.edu.cn

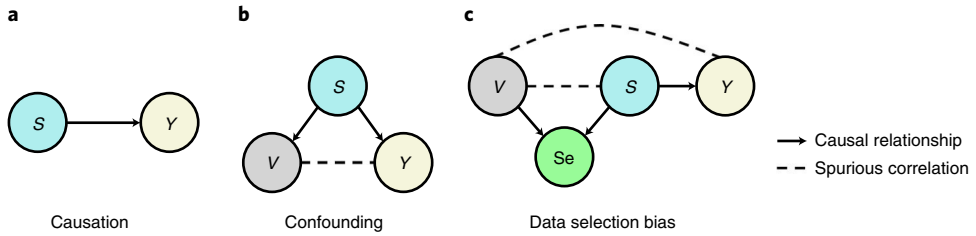

Fig. 1 | Three ways of generating correlations. *Y* is the outcome variable, *S* is the input variable corresponding to the direct cause of *Y*, *V* is another input variable that may have spurious correlation with *Y* and Se is a selection variable that affects whether a sample with certain *V* and *S* is included in a dataset. **a**, A causal relationship between *S* and *Y*. **b**, A spurious correlation between *V* and *Y* arising from a confounder *S*. **c**, Selection (Se) based on both *V* and *S* can lead to spurious correlation between *V* and (*S*,*Y*) in the selected dataset.

Spurious correlation is a key source of risk

Machine learning models, taking supervised learning as an example, eventually learn linear or nonlinear correlation relationships between input variables and output variables. That is to say, correlation is the statistical basis of these learning algorithms. Correlations in data can arise for a variety of reasons. Several important scenarios are illustrated in Fig. 1.

Causality. When one of the two variables is the direct or indirect cause of the other, there is an association between them, as shown in Fig. 1a. For example, weather (that is, *S*) affects crop yields (*Y*), therefore the weather in a season is correlated with crop yields in that season. This type of relationship reflects intrinsic and universal dependency among variables and remains invariant across different settings, even if the magnitude depends on context.

Confounding. When two variables share common causes (that is, confounders), they will be associated with one another, as shown in Fig. 1b. For example, the condition of a patient (that is, *S*), especially the seriousness of his/her disease, is the common cause of ICU treatment (that is, *V*) and recovery rate (*Y*). If we directly measure the correlation between ICU treatment and recovery rate without properly balancing the conditions of patients, we will get an erroneous conclusion that ICU treatment leads to a lower recovery rate—a spurious correlation. This kind of (unconditional) correlation is usually difficult to interpret. Meanwhile, since the strength and even sign of a correlation depend on the correlation between *S* and *V*, its stability is weak when the joint distribution of features varies across environments.

Data selection bias. Data selection bias is common and even inevitable in real cases where data is selected in a way that differs from the target population. A typical case is shown in Fig. 1c, where analysing data only for observations with a high value of the selection variable Se will result in spurious correlations among *V*, *S* and *Y*. Consider an example of image classification in the 'dog' category. We might collect a training dataset in which most positive samples depict dogs on grass; the features of grass (that is, *V*) will then be spuriously correlated with the features of dog (*S*), and thus lead to a spurious correlation between features of grass and the 'dog' label (*Y*). Given that data selection bias is usually induced unintentionally, such spurious correlations can be difficult to identify in advance. If the training data distribution differs from the test data, predictions will be inaccurate.

Among these three ways of generating a correlation, only the correlations generated by a causal relationship reflect the intrinsic dependency among variables; the other two types are spurious correlations sensitive to the joint distribution of features and the data collection processes. Nevertheless, in today's off-the-shelf machine learning, black-box models do not even try to differentiate the three different ways in which these correlations are generated. Therefore, their predictive performance depends heavily on how much the test distribution shifts from the training distribution, leading to unstable performance under varying test distributions. Meanwhile, a predictive model based on spurious correlation may also be unfair. To fundamentally address the risks of stability, explainability and fairness, we need to embrace and emphasize causality in the machine learning framework.

Challenges and opportunities for causality in predictive modelling

A causal model matches the underlying process the generating data. In Fig. 2 we showcase the physical process for generating datasets, occurring over time. By the very nature of prediction problems, an analyst is attempting to use the pre-outcome variables to predict future and unseen outcomes. In product recommendation systems, a user characterized by his/her attributes shows varying levels of interest in products with different attributes, and finally generates purchasing behaviour simultaneously caused by his/her and product attributes. In image classification problems, a photo is first selected into the dataset, then an image annotator observes the photo content and extracts features, and finally he/she annotates the photo with a category label according to his/her understanding of the visual content. Therefore, features of an image are the causes and its label is the effect. Although this scenario is described as an example of an anticausal case in ref. [4], the causal structure they propose is a description of the data generation mechanism of $P(\mathbf{X}, Y)$, rather than $P(Y|\mathbf{X})$ which is the object of interest for predictive modelling. If the generating process is described alongside relevant features of the environment, the process is fundamentally stable. This could serve as an important motivation for machine learning researchers to incorporate causality into machine learning prediction problems[5].

Estimating causal effects using observational data requires strong assumptions. One of the most popular approaches can be described as follows. First, the researcher observes potential confounders, and assumes that after adjusting for these observed confounders, treatment assignment is independent of a unit's potential outcomes. This assumption is referred to as unconfoundedness[6]. A second assumption is the stable unit treatment value assumption (that is, the response of a particular unit depends only on its treatment, not the treatments of other units). Third, the overlap assumption requires that conditional on each possible realization of observed confounders, all units have a non-zero probability of assignment to each treatment condition[6,7]. Unfortunately, these assumptions are mostly untestable (although in practice, researchers conduct a variety of supplementary analyses to assess the credibility of their assumptions[8]). Outside large-scale, multi-treatment randomized controlled trials, it can be extremely challenging to find settings where the relevant assumptions can be justified when there are many possible treatments. Meanwhile, owing to what has been called the 'fundamental problem of causal inference'[9], where we do

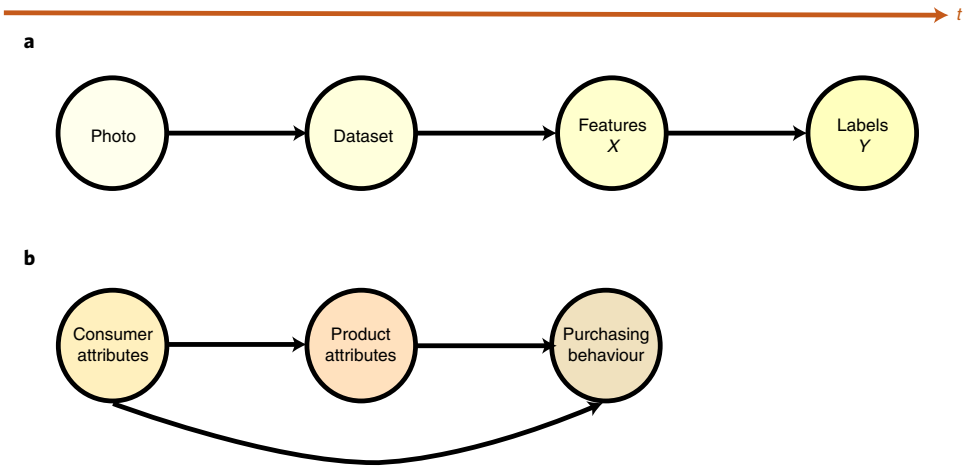

Fig. 2 | The physical processes for generating datasets used in predictive modelling, occurring over time. a, Image classification. **b**, Recommendation systems. *t*, time.

not observe a unit simultaneously treated and untreated, there is a missing data problem that makes it difficult to determine the validity of a causal model.

In general, if the true causal structure can be identified and estimated (that is, if the data generation process can be uncovered), the prediction problem can be naturally solved as a side product. But to follow this technical path, we have to solve all of the challenges for causal inference. Doing so may be impossible in a realistic dataset. Despite this, we argue that predictive modelling does not need to reconstruct the true data-generating process, and the best predictive model will balance considerations of bias and variance in model selection. Thus, the strict goals common to causal inference, such as consistent estimation of causal effects, are not required, and approximations or improvements may be possible even when we do not have the data necessary to fully solve causal inference problems. Another reason that predictive modelling is easier is that the ground truth of the predicted outcome is available, so the correctness of a model can be quantitatively evaluated in held-out test sets. Therefore, the challenges for validation arising from the fundamental problem of causal inference can be avoided.

Thus, we argue here that common ground between machine learning and causal inference should be built. The framework of stable learning is thus proposed and developed as one approach to meeting this goal.

The positioning and development of stable learning

Different from traditional machine learning settings, we do not maintain the assumption that the test dataset comes from the same distribution as the training data. Given training data $D^e = (X^e, Y^e)$ from environment $e \in \mathcal{E}$, where X^e are features and Y^e is the outcome variable, stable learning aims to learn a predictive model that can achieve uniformly good performance on any possible environment in $\mathcal{E}$. Of course, there must be some common link across environments to make positive progress. In this Perspective, we focus on the setting of covariate shift generalization[10], where $P^e(Y|X)$ does not vary with e, but $P^e(X)$ varies.

Ref. [11] formalizes the objective for stability based on Average_Error and Stability_Error, which refer to the mean and standard deviation, respectively, of the predictive error over all possible environments $e \in \mathcal{E}$. Note that the stability here is defined over prediction performance, rather than estimation stability, as in ref. [12]. When evaluating the stability error of a stable learning model, the analyst cannot anticipate all possible test environments. Although it is often possible to simulate a range of test environments by repeatedly creating nonrandom subsamples of test data for the evaluation of stability, this introduces a subjective element into the algorithm.

We illustrate the relationships of different learning paradigms in Fig. 3. The most common situation is learning with the assumption[13] that the training and test data are independent and identically distributed (IID). However, the test distribution may shift arbitrarily from the training distribution. Transfer learning (or domain adaptation) methods[14] assume that we have previous knowledge of the target distribution that we may encounter in the test phase. More recently, the problem of domain generalization has attracted increasing attention. These methods mostly require the training data to be composed of different environments, and their performances highly depend on the diversity of predefined or pre-identified training environments[15].

Compared with the learning paradigms mentioned above, stable learning aims for a more realistic problem setting. On the one hand, we do not assume the availability of any strong prior on the test distribution, as in the problems of 2D learning or transfer learning. On the other hand, we do not assume the availability of multiple environments in training data, as in domain generalization. Meanwhile, stable learning poses a higher standard for a model's generalization ability. The learned model should achieve a good performance on average in unseen environments. Such a high standard poses more challenges for machine learning models and forces us to rethink the generalization problem more fundamentally.

Stable learning from the causality perspective. In contrast to causal inference, which seeks consistent estimates of the effects of treatments and their interactions, stable learning aims to learn the mapping between a potentially larger number of treatment variables and the outcome. To interpret stable learning from a causality perspective, we begin by considering the case where (1) there are no effect variables of the outcome variable Y in the system; that is, the outcome variable cannot be the cause of any other variables; and (2) the above-mentioned three assumptions (unconfoundedness, overlap and the stable unit treatment value assumption) are satisfied for every pair of (X_i, Y).

The original idea of stable learning is motivated by the literature on covariate balancing strategies in causal inference[16–18], which are used to estimate the average effect of manipulating a single treatment in the presence of many potential confounders. Such methods attempt to construct sample weights that balance covariates' distributions between the treated and the control group, after which the correlation between treatment and outcome variable is a consistent

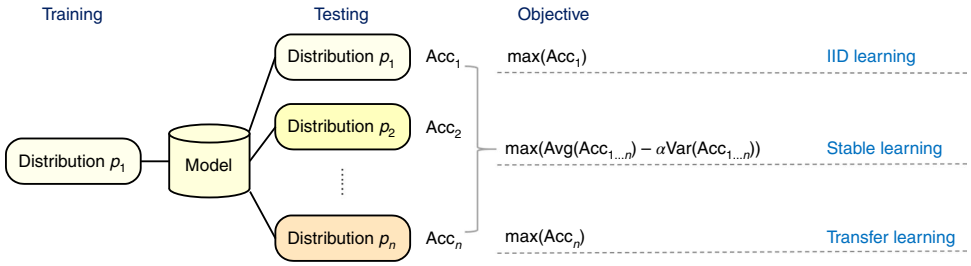

Fig. 3 | Comparison of different learning paradigms. We differentiate the learning problems into IID learning, transfer learning and stable learning on the basis of the learning objectives with respect to testing distributions. Acc means accuracy, and α is a hyperparameter to tradeoff the average accuracy and variance across different distributions.

estimate of the causal effect. Although there are other approaches to estimating causal effects under unconfoundedness[19], this approach based on reweighting is particularly useful as a starting point for stable prediction. However, existing methods in covariate balancing are tailored to environments with few treatments. Stable learning, when viewed through the causal inference lens, poses a more ambitious question: if we regard each input variable as the treatment iteratively and all remaining input variables as its covariates, does a set of sample weights that can realize covariate balancing globally (that is, global balancing) for whichever input variable acts as the treatment exist? If so, the set of sample weights can allow us to consistently estimate the causal effect of each input feature on the basis of the correlation between the feature and the outcome in the reweighted data.

Note that the causal effect estimated by global balancing can be interpreted as a direct effect. Consider the causal structure $Y \leftarrow X_2 \longrightarrow X_1 \longrightarrow Y$ as an example, where for simplicity the features are binary and the direct effect of X_2 is additive to that of X_1. When considering X_1 as the treatment, X_2 plays the role of a confounder, so balancing X_2 between the X_1-treated and X_1-control groups leads the correlation between X_1 and Y to serve as an estimate of the direct effect (equals the average treatment effect, ATE) of X_1 on Y. When considering X_2 as the treatment, X_1 plays the role of a mediator between X_2 and Y[20]. Balancing X_1 between the X_2-treated and X_2-control groups eliminates the effect of X_2 on Y through the mediator X_1, and the resulting correlation between X_2 and Y is the controlled direct effect of X_2 on Y; that is, the part that is not mediated by X_1[21]. In linear systems, the direct effect is independent of the value at which we hold X_1, whereas in nonlinear systems, the direct effect depends of the value of X_1 after balancing. For the goal of prediction when both X_2 and X_1 are observed, this is all we need: we do not require an estimate of the direct effect of X_2 on X_1, which might vary across environments. If we know the direct effects of both X_1 and X_2, with this causal structure, we can predict outcomes even if the joint distribution of (X_1, X_2) changes. If the direct effect of X_1 depends on the value of X_2 (that is, if there is an interaction effect in the outcome model), then the average (over X_2) of the effect of X_1 depends on the joint distribution of (X_1, X_2), so it is important to incorporate such interactions in the predictive model to achieve stability with respect to the joint distribution of features.

In pursuit of such a set of sample weights for global balancing, we proposed an approach in ref. [11] for removing the correlations among the features, so that for each feature, when considering that feature as a treatment variable, the covariate distribution is balanced between the treated group and the control group. We showed that under the law of large numbers and the overlap assumption mentioned above, there exists an optimal weight W^* that reduces the global balancing loss to zero. However, with many features and realistic sample sizes, the overlap assumption may fail, and we may not have observations associated with every combination of features.

Therefore, a series of algorithms are proposed to optimize the sample weights towards global balancing. The process starts with a global balancing loss designed for binary input variables[22] that can be easily plugged into standard learning tasks as a regularizer. It is demonstrated that, after integrating the global balancing loss into a standard logistic regression model, the learned regression coefficients possess both predictive power and causal implication. To relieve the overlap assumption especially with small sample size or high-dimensional feature space, an unsupervised representation learning module is integrated into the global balancing stage, forming a 'deep' version of the original regularizer[11]. By introducing the criterion of continuous variable independence in ref. [23], the regularizer of global balancing is extended from binary variables to continuous variables[24], which is the common setting in learning scenarios.

By extending the confounder balancing techniques from causal inference into machine learning problems, we have seen promising results in improving the stability of machine learning models. But, as mentioned above, we need strict assumptions to make causal interpretations of stable learning. This motivated us to explore other theoretical support for stable learning.

Stable learning from the statistical learning perspective. Formally speaking, the advantages of stable learning are attained by sample reweighting. Hence, there arises a natural question: why does sample reweighting improve the stability of a correlation-based model (such as a linear regression)? Can stable learning algorithms still improve stability without fully achieving the more ambitious goals of causal inference?

To answer these questions, in ref. [25] we investigated the stable learning problem in a linear regression framework with model misspecification, where the true data-generating process was characterized by nonlinearities or interactions not included by the analyst. Suppose that:

$$Y = X^\top \bar{\beta}_{1:p} + \bar{\beta}_0 + b(X) + \epsilon \tag{1}$$

The nonlinear term $b(X)$ is constrained to be smaller than a small volume δ, and ϵ is a noise term. If we can correctly estimate the coefficients $\bar{\beta}$ and use them for prediction, the model may produce uniformly good prediction results for any sample, leading to stable performances under arbitrary distributions. Therefore, the stability of a model can be quantified by model estimation error $||\hat{\beta} - \bar{\beta}||_2$ where $\hat{\beta}$ represents the estimated coefficients. We theoretically prove that this estimation error is upper bounded by $O(\delta/\lambda)$ where λ is the smallest eigenvalue of the design matrix, indicating the degree of collinearity among the input variables[25]. If a misspecified model is used at the training stage, the existence of collinearity among input variables can inflate a small misspecification error to an arbitrarily large size. Both the theorem and empirical results tell us that reducing the collinearity among input variables, which is the direct effect of global balancing via sample reweighting, is an effective way to improve stability.

It should be noted that the three assumptions required by causal inference are not explicitly discussed here. Unconfoundedness is not strictly required by stable learning; that is, latent variables are allowed. As long as the joint distribution of latent variables with observed variables is stable, the stability of a stable learning model can be guaranteed; or else, stable learning does not inflate the harmfulness of latent variables compared with traditional predictive models. Meanwhile, the covariate decorrelation process does not depend on the overlap assumption, but a dataset that better satisfies the overlap assumption leads to lower estimation variance. The stable unit treatment value assumption is also not necessary if we focus on prediction performance, rather than on causal interpretation. In stable learning, the estimation of the causal effects is a means to an end, rather than the primary goal. Stable learning thus has the potential to optimize the trade-offs involved in balancing bias and variance, using the data to make those trade-offs.

Bridging the causality and learning perspectives. Although from both viewpoints, stable learning algorithms adopt sample reweighting as the technical way to improve model stability, the notions underpinning the idea are different. Here we try to bridge these two notions within the regression framework.

As we iteratively regard each input variable as the treatment in stable learning, we suppose that all of its confounding covariates are contained in the remaining input variables. In the cases of binary treatment variables, the learned sample weights eventually make the treatment variable independent of the remaining variables. Extending this interpretation into global balancing, we conclude that the learned global sample weights can make all input variables independent of each other. Thereafter, when we conduct regression over the weighted samples, each input variable's regression coefficient represents its partial effect on the outcome, which is regarded as the causal effect[26]. Similarly, from the view of statistical learning, the effect of removing the collinearity among input variables tends to make input variables independent. Therefore, making input variables independent is the common objective of these two perspectives of stable learning, which also provides a common ground for causal inference and machine learning.

Furthermore, we prove that making input variables independent can help to identify true variables for predictions. Consider a data generation process $Y=f(S)+\epsilon$, where $\mathbf{X}=(S,\mathbf{V})$ and $Y\perp\mathbf{V}|S$, which covers all the cases shown in Fig. 1, where both confounding and selection bias cases are included. If $f(.)$ is a nonlinear function, model misspecification occurs when we use a linear regression model such as ordinary least squares (OLS), resulting in non-zero coefficients on $\mathbf{V}$. With the sample weights learned from stable learning methods such as sample reweighted decorrelation operator (SRDO)[25], weighted OLS can guarantee zero coefficients on $\mathbf{V}$ in the learned function, which means only the variables (S) are used for prediction, so that prediction is stable even when the joint distribution of $(S,\mathbf{V})$ changes[10].

The causal inference framework provides a fundamental view of understanding the stability of a learning model like regression. Still, the causal interpretations of regression coefficients can only be justified by relying on much stricter assumptions than are needed for predictive inference[27]. Comparatively, the learning perspective can help to weaken these assumptions for better performance in more complicated tasks. Therefore, as one aspect of the common ground between causal inference and machine learning, stable learning can extend in the theoretical foundation and practical predictive power.

Implications of explainability and fairness

In real applications, the stability, explainability and fairness properties are often jointly required. As these properties are inherently related to causality, causality-inspired stable learning may potentially provide implications of explainability and fairness.

A recent article[2] appeals to the community to stop explaining black-box models and use inherently interpretable models instead. As human models are often based on causality with the ultimate aim of understanding the underlying mechanisms[28], it is natural to incorporate causality to form a common ground for human and predictive models. Therefore, in the Explainable AI project sponsored by Defense Advanced Research Projects Agency (DARPA), causal models are regarded as a prominent technical path[29]. On the other side, in the study of explainable AI, partial dependence plots are commonly employed as a diagnostic technique to generate insights into the importance of specific features in the model's predictions[30]. Stable learning models are consistent with the notion of causality and evaluation metrics of explainability such as partial dependence plots, naturally guaranteeing their explainability[22,31].

With respect to fairness, mainstream studies propose various metrics for measuring group fairness[32,33] and individual fairness[32,34]. In contrast to the existing metrics that are typically applied directly to observational data, causal inference may provide a generative angle to frame fairness problem[35]. In stable learning, under reasonable assumptions, we eventually exploit direct causal variables to predict the outcome, which can avoid the fairness issues caused by spurious correlations[36]. Recently, researchers established a close connection between fairness problems and stability (or robustness) problems[37,38], making stable learning the preferred candidate for addressing fairness problems.

Conclusion

The stability, explainability and fairness problems of machine learning algorithms are urgently needed to be addressed if we expect these algorithms to be widely deployed. However, most studies on these topics try to rectify today's models (for example, deep learning models), which are inherently divergent from these goals. We argue that these problems are fundamental limitations of today's learning paradigm that need to be addressed radically. The whole question of 'what should the basis for prediction, correlation versus causality be?' needs to be rescrutinized, despite the long historical debate. The recent progress of causal inference—especially in observation studies—can provide more insights into, and theoretical support for, machine learning. Stable learning is presented as an attempt to find common ground between these two directions. How to reasonably loosen strict assumptions to matchapplication scenarios 'in the wild', and make machine learning more trustworthy without sacrificing predictive power, are crucial questions for stable learning to address in the future.

Received: 16 December 2020; Accepted: 12 January 2022;
Published online: 23 February 2022

References

1. Athey, S. C., Bryan, K. A. & Gans, J. S. The allocation of decision authority to human and artificial intelligence. *AEA Papers and Proceedings* **110**, 80–84 (2020).
2. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
3. Corbett-Davies, S. & Goel, S. The measure and mismeasure of fairness: a critical review of fair machine learning. Preprint at https://arxiv.org/abs/1808.00023 (2018).
4. Heinze-Deml, C. & Meinshausen, N. Conditional variance penalties and domain shift robustness. *Mach. Learn.* **110**, 303–348 (2021).
5. Pearl, J. Theoretical impediments to machine learning with seven sparks from the causal revolution. In *Proc. of the Eleventh ACM International Conference on Web Search and Data Mining* (2018).
6. Imbens, G. W. & Rubin, D. B. *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge Univ. Press, 2015).
7. Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).
8. Athey, S. & Imbens, G. A measure of robustness to misspecification. *Am. Econ. Rev.* **105**, 476–480 (2015).
9. Holland, P. W. Statistics and causal inference. *J. Am. Stat. Assoc.* **81**, 945–960 (1986).

10. Xu, R., Cui, P., Shen, Z., Zhang, X. & Zhang, T. Why stable learning works? A theory of covariate shift generalization. Preprint at https://arxiv.org/abs/2111.02355 (2021).
11. Kuang, K., Cui, P., Athey, S., Xiong, R. & Li, B. Stable prediction across unknown environments. In *Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 1617–1626 (2018).
12. Yu, B. et al. Stability. *Bernoulli* **19**, 1484–1500 (2013).
13. Vapnik, V. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems* 831–838 (1992).
14. Pan, S. J. et al. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
15. Shen, Z. et al. Towards out-of-distribution generalization: a survey. Preprint at https://arxiv.org/abs/2108.13624 (2021).
16. Athey, S., Imbens, G. W. & Wager, S. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Series B Stat. Methodol.* **80.4**, 597–623 (2018).
17. Zubizarreta, J. R. Stable weights that balance covariates for estimation with incomplete outcome data. *J. Am. Stat. Assoc.* **110**, 910–922 (2015).
18. Hainmueller, J. Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Political Anal.* **20.1**, 25–46 (2012).
19. Guo, R., Cheng, L., Li, J., Hahn, P. R. & Liu, H. *A Survey of Learning Causality With Data: Problems and Methods* **53.4**, 137 (ACM Computing Surveys (CSUR), 2021).
20. Hicks, R. & Tingley, D. Causal mediation analysis. *Stata J.* **11**, 605–619 (2011).
21. Pearl, J. Direct and indirect effects. In *Proc. of the Seventeenth conference on Uncertainty in Artificial Intelligence* 411–420 (2001).
22. Shen, Z., Cui, P., Kuang, K., Li, B. & Chen, P. Causally regularized learning with agnostic data selection bias. In *Proc. of the 26th ACM International Conference on Multimedia* 411–419 (2018).
23. Bisgaard, T. M. & Sasvári, Z. When does e (xk· yl)= e (xk)· e (yl) imply independence? *Stat. Probabil. Lett.* **76**, 1111–1116 (2006).
24. Kuang, K., Xiong, R., Cui, P., Athey, S. & Li, B. Stable prediction with model misspecification and agnostic distribution shift. In *Proc. of the AAAI Conference on Artificial Intelligence* **34**, No. 04 (2020).
25. Shen, Z., Cui, P., Zhang, T. & Kunag, K. Stable learning via sample reweighting. In *Proc. of the AAAI Conference on Artificial Intelligence* **34**, no. 04, 5692–5699 (2020).
26. Cornelißen, T. & Sonderhof, K. Partial effects in probit and logit models with a triple dummy-variable interaction term. *Stata J.* **9**, 571–583 (2009).
27. Gelman, A. & Hill, J. in *Data Analysis Using Regression and Multilevel/Hierarchical Models* 167–198 (Cambridge Univ. Press, 2007).
28. Holzinger, A., Langs, G., Denk, H., Zatloukal, K. & Müller, H. Causability and explainability of artificial intelligence in medicine. *WIREs Data Min. Knowl. Discov.* **9**, e1312 (2019).
29. Gunning, D. & Aha, D. W. DARPA's explainable artificial intelligence program. *AI Mag.* **40**, 44–58 (2019).
30. Rai, A. Explainable AI: from black box to glass box. *J. Acad. Market. Sci.* **48**, 137–141 (2020).
31. Zhang, X., Cui, P., Xu, R., Zhou, L., He, Y., & Shen, Z. Deep stable learning for out-of-distribution generalization. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 5372–5382 (2021).
32. Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. Fairness through awareness. In *Proc. of the 3rd Innovations in Theoretical Computer Science Conference* 214–226 (2012).
33. Hardt, M. et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* 3315–3323 (2016).
34. Kusner, M. J., Loftus, J., Russell, C. & Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems* 4066–4076 (2017).
35. Kilbertus, N. et al. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems* 656–666 (2017).
36. Adragna, R., Creager, E., Madras, D. & Zemel, R. Fairness and robustness in invariant learning: a case study in toxicity classification. Preprint at https://arxiv.org/abs/2011.06485 (2020).
37. Hashimoto, T. B., Srivastava, M., Namkoong, H. & Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning* 1929–1938 (PMLR, 2018).
38. Roh, Y., Lee, K., Whang, S. E. & Suh, C. FR-Train: a mutual information-based approach to fair and robust training. In *International Conference on Machine Learning* 8147–8157 (PMLR, 2020).

Acknowledgements

Peng Cui's research is supported by National Key R&D Program of China (No. 2018AAA0102004), National Natural Science Foundation of China (No. U1936219), Beijing Academy of Artificial Intelligence (BAAI) and Guoqiang Institute of Tsinghua University.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Peng Cui.

Peer review information *Nature Machine Intelligence* thanks Kush Varshney and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.