Research Statement

Krikamol Muandet

Ultimately, I seek to understand "*how we can build machines that can learn to generalize in the real world from past observations*". Achieving this goal requires inference and learning not only on data points, but also over probability distributions that generate them. To this end, I have worked extensively on a kernel mean embedding (KME) of distributions which offers powerful mathematical tools to work with distributions. In particular, it provides a succinct representation of an interventional distribution which is instrumental for causal reasoning. Understanding of cause-effect relationships enables machines to generalize better in the real world and is a prerequisite for consequential decision making in health care, education, public policy, and justice system. In these areas, collecting experimental data can be expensive, time-consuming, or even unethical, so most algorithmic decisions are made on the basis of non-experimental data alone. As a result, a reliable algorithmic decision remains unattainable. Finally, the omnipresence of data-driven systems and scarcity of resources in socioeconomic systems call for an oversight that will ensure long-term sustainability.

My research aims to combine ideas from machine learning with those from economics such as game theory and mechanism design. The complementary nature of tools from these two mostly disjoint fields will allow me to address the aforementioned challenges. In Section I, I will discuss my scientific achievement followed by vision for future research in Section II. A bird's-eye view of my research is given in Fig. 1.

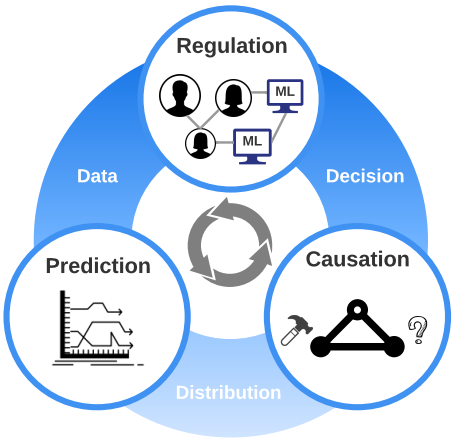

Fig. 1: **Prediction-Causation-Regulation (PCR)** research agenda. (**P**) A predictive modeling requires an accurate estimate of the underlying probability distribution. (**C**) The interventional distribution informs the consequences of algorithmic decisions in the real world. (**R**) Due to the scarcity, these decisions, especially personalized ones, may introduce incentives for individuals to change their behaviour. This in turn creates a feedback loop in the data collection process.

I. Scientific Achievements

My research endeavour has led to several scientific contributions at the flagship conferences and journals in machine learning (JMLR [1–5], FnT ML [6], NeurIPS [7–12], ICML [13–18], UAI [19–21], and AISTATS [22, 23]) as well as other disciplines (CVPR [24], 3DV [25, *best paper award*], Physical Review Research [26], and Journal of Nonlinear Science [27] among others). In Section I-A and I-B, I describe my research that lie on the "prediction–causation" spectrum in Fig. 1.

A. Distributional Prediction with Kernels

To deal with probability distributions, I have worked extensively on a **kernel mean embedding** (KME) of distributions [28, 29] and have published a highly-cited book entitled "Kernel Mean Embedding of Distributions: A Review and Beyond" [6]. The idea of KME is to represent a distribution $P(X)$ over some random variable X as a function in a high-dimensional feature space known as a reproducing kernel Hilbert space (RKHS), i.e., $P \mapsto \mu_P := \mathbb{E}_P[k(X,\cdot)]$ for some positive definite kernel k. It provides a complete representation of distributions, and by means of the KME, one can compare two distributions as well as perform basic operations such as sum rule, product rule, and Bayes rule. Furthermore, owing to kernel function k, one can work with distributions over non-standard data structures such as strings, graphs, and groups in an integrated framework. The key advantage of KME is that it can be estimated consistently using an empirical average over the i.i.d. sample $x_1,\ldots,x_n$ without any parametric assumptions on $P(X)$, i.e., $\hat{\mu}_P := (1/n)\sum_{i=1}^n k(x_i,\cdot)$, making it less susceptible to model mis-specification. In [3], my colleagues and I showed that **this estimator is minimax optimal**. Surprisingly, in another line of research we showed that the kernel mean estimation can still be improved in practice thanks to the Stein's shrinkage phenomenon [2, 8, 15]. That is, shrinkage estimators of the KME outperform the standard one $\hat{\mu}_P$ even when we have no further information about $P(X)$. We proposed a family of estimators called **kernel mean shrinkage estimators** (KMSE) equipped with an efficient cross validation procedure to select an optimal shrinkage parameter. In [12], my colleagues and I overcame the long-standing limitations of the conditional mean embedding (CME)–the KME of the conditional distribution $P(Y|X)$–via an alternative operator-free definition based on a measure-theoretic perspective, i.e., **we define the embedding of $P(Y|X)$ as an X-measurable random variable taking values in the RKHS**. This novel definition eases theo-

The extended version of this research statement is available at http://www.krikamol.org/assets/pdf/krikamol-pcr-research.pdf.

retical analyses associated with the CME, allows us to naturally extend maximum mean discrepancy (MMD) and Hilbert-Schmidt Independent Criterion (HSIC) to the conditional setting, which we call the *maximum conditional mean discrepancy* (MCMD) and the *Hilbert-Schmidt conditional independence criterion* (HSCIC), and also inspires future applications. My works in this area have led to some fundamental results as well as numerous collaborations in deep learning [11, 24], causality [1, 16, 18, 30], molecular dynamic [27], dynamical systems [31], quantum computing [26], and econometrics [9, 21, 32].

I was also involved in the pioneering development of learning algorithms that operate on a collection of distributions. For instance, I developed a framework called **distributional risk minimization** (DRM) that extends the conventional empirical risk minimization (ERM) to a space of distributions [7, 19] and proved a result that generalizes the well-known **representer theorem** [33] to a space of distributions [7, Theorem 1]. Based on this framework, I proposed learning algorithms on distributions called *support measure machine* (SMM) and *one-class support measure machine* (OCSMM) [7, 19] which generalize the support vector machine (SVM) algorithm with extended kernel functions on probability distributions. Learning directly on distributions allows us to better capture the uncertainty and to reason about the aggregate behaviors that arise from multiple collections of data.

My work [13] also laid the foundation for subsequent works in the area of out-of-distribution (OOD) generalization. In [13], my colleagues and I studied the **domain generalization** (DG) problem [34]: *Given data sets from multiple domains, learn a predictor that generalizes well to any previously unseen domains*.[1] Our work was the first to advocate that **learning a domain-invariant representation leads to improved out-of-distribution generalization**, which was justified theoretically by our learning-theoretic analysis [13, Theorem 5]. To allow for better generalization, my intern and I recently combined this idea with deep neural network and applied it to a few-shot learning problem [11]. A novel task representation called *model-aware task embedding* (MATE) incorporates not only the data distributions of different tasks, but also the complexity of these tasks. Since the OOD generalization also bears a resemblance to the social welfare in economics, I am now exploring ideas to solve it from this perspective.

B. Causation and Decision Making

More recently, my research focuses on a synergy between machine learning and economics [1, 9, 18, 21, 32]. I am developing algorithms that help solve problems in economics, while bringing the economic perspective into the design of new learning algorithms. While machine learning embraces the potential of powerful predictive models, economics addresses substantive questions in education, public policy, and social programs. The interface between these two fields will thus reveal the challenges of deploying machine learning in the real world. The nonparametric nature of KME also makes it ideal for economic applications where a distributional assumption is undesirable.

Distributional effects: In econometrics, a prime objective is to predict the effect of a policy intervention or a counterfactual change in economic conditions on some outcome variables. That is, we are concerned with an **interventional distribution** $P(Y|\mathrm{do}(X))$ which describes how the distribution of Y would change (or would have changed) as a result of an intervention on X. The policy intervention generally aims to understand non-trivial effect of change in distribution $P(X)$. However, most of existing works in the literature focus on mean effects such as average treatment effect (ATE) and conditional average treatment effect (CATE), which do not inform changes in higher-order moments. To overcome this limitation, my colleagues and I proposed a *counterfactual mean embedding* (CME) [1] as a Hilbert space representation of the counterfactual distribution that represents **the outcome of hypothetical change which by definition is not observable**. Thanks to our estimators, the CME can be estimated consistently from observable quantities. One of the important applications of our work is **offline policy evaluation** [1, Section 6], which is relevant to a program evaluation in economics. In [17], my colleagues and I extended this idea and proposed the *conditional distributional treatment effect* (CoDiTE), which, in contrast to the more common CATE, is designed to encode a treatment's distributional aspects and heterogeneity beyond the mean. We applied CoDiTE to LaLonde's well-known National Supported Work (NSW) dataset [35] to model the wage distribution. Since wage distributions are known to be skewed, CoDiTE can capture the treatment effect better than CATE [17, Sec. 6.2]. Our works [1, 17] are at the forefront of distributional treatment effect (DTE) estimation with machine learning algorithms.

Instrumental variable (IV): A major obstacle faced by policymakers and decision makers is the presence of unobserved confounders. It jeopardizes the reliability of the decisions and policies. An instrumental variable (IV) is a signature method in econometrics to overcome the endogeneity in data. A nonparametric instrumental variable (NPIV) regression solves a Fredholm integral equation which is an ill-posed inverse problem. Recent approaches in machine learning can be categorized either as a two-stage approach [36, 37] or GMM-based approach [38, 39]. In [9], I showed that the NPIV can be reformulated as **a two-player game based on a convex-concave utility function**. This reformulation, which is called DualIV, offers much simpler algorithms. In particular, when both players are parameterized by the RKHS functions, **the global**

[1] This problem was first studied in [34], but the term "domain generalization (DG)" was coined for the first time in our paper [13].

equilibrium can be obtained in closed-form [9, Sec. 4]. This work also sheds light on the duality between the two-stage and GMM-based approaches, as later pointed out in [40, Appendix F] and our work [18, Sec. 3.3]. Lastly, I find the DualIV fascinating as it elucidates the kind of problems for which a game-theoretic perspective as a search for Nash equilibrium can lead to simpler algorithms than the standard algorithms.

Maximum moment restriction (MMR): Like many problems in econometrics, the NPIV can also be solved via a so-called conditional moment restriction (CMR): *for correctly specified models, the conditional mean of certain functions of data is almost surely equal to zero* [41, 42]. The major challenge when working with the CMR is that it implies an infinite number of unconditional moment restriction (UMR) which make the corresponding inference and estimation intractable. Based on the vector-valued RKHS (vv-RKHS), my colleagues and I proposed a **maximum moment restriction** (MMR) [21] by transforming the original CMR to a maximum of the interaction between the generalized residual function and functions of the conditioning variables that belong to a unit ball of the vv-RKHS. Surprisingly, **the MMR not only provide a tractable form for inference and estimation, but also captures all necessary information about the original CMR.** In other words, no information is lost by the restriction to the RKHS. This work opens up new research avenues that lie at the intersection between econometrics and modern kernels methods in machine learning. In particular, my colleagues and I have applied this framework successfully to the problems of conditional moment (CM) test [21], IV regression [32], and proximal causal learning [18].

II. Vision for Future Research

As I described in Section I-B, modern kernel methods and kernel mean embedding are natural tools for tackling several problems in economics. **To go even further, my vision for future research lies in the challenges that arise from interactions between machine learning systems and real-world environments, especially socioeconomic systems**. To create models that can learn to generalize in the real world, it is imperative to incorporate information about the *data collection process*, effects of the *deployment in complex environments*, and *human behaviour* into the design of learning algorithms. In my opinion, an incorporation of economic thinking will be a game changer for machine learning in dealing with critical challenges such as feedback loop, market equilibrium, non-stationarity, heterogeneity, delayed effect, endogeneity, and strategic response, which fall into the "regulation" spectrum in Fig. 1. I discuss a couple of important challenges below.

Feedback loop: The deployment can create a feedback loop in the data collection process (see Fig. 1). That is, the historical data are used in building the models which are in turn deployed to collect more data. In loan decisions, for example, a bank may decide whether or not to offer a loan based on learned models of the credit default. These decisions generate more data that are used to improve the models. In [23], my colleagues and I analyzed consequential decision making using imperfect predictive models, which are learned from data gathered by potentially biased historical decisions. We articulated that when starting with a non-optimal deterministic policy, this approach fails to optimize utility for sequential decisions. To avoid this failure mode while respecting a common fairness constraint, we suggest to directly learn the decisions with exploring policies. The results of this work highlight the need of regulation that governs the design and applications of predictive models in the real world. One of the questions I am exploring is how to overcome this challenge by leveraging the idea of IV described in Section I-B.

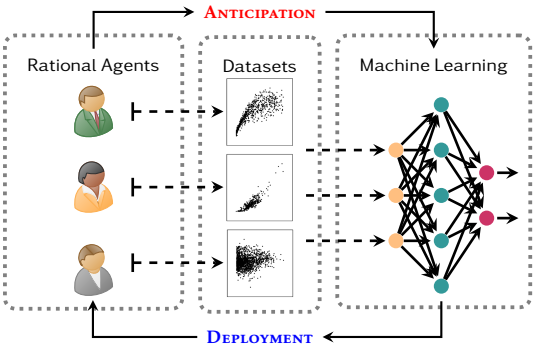

Fig. 2: **Learning as a robust mechanism design**: When data come from self-interest agents (individuals, institutions, or even countries) that can anticipate the outcome of data-driven models, the deployment of static models can be *sub-optimal*. In this setting, the problem is akin to designing a game to be played by the data against the model.

Learning as a robust mechanism design: The deployment may also introduce incentives for individuals to change their behaviour. Hence, the models become *sub-optimal* when observed data are manipulable, e.g., in heterogeneous pricing, individualized credit offer, and target social program, as illustrated in Fig. 2. When the data are subject to manipulation, we must design a mechanism with which the models will be governed to reach a desirable long-term social welfare, for example, by eliciting the right incentives for people [43]. This question leads to a growing interest in strategic classification [44–46] and performative prediction [47] among others. Compared to immediate impacts of machine learning such as unfairness and algorithmic biases, long-term impacts are often neglected as they can be difficult to foresee. Thus, human behaviour must be carefully taken into account when building machine learning systems that will become part of the society. To address this challenge, the key question is how to design learning algorithms that produce the incentive-compatible models, i.e., the data provided are truthful.

References

[1] **Muandet, K.**, Kanagawa, M., Saengkyongam, S., Marukatat, S., "Counterfactual mean embedding," *Journal of Machine Learning Research*, vol. Accepted, 2021. [Online]. Available: https://arxiv.org/abs/1805.08845.

[2] **Muandet, K.**, Sriperumbudur, B., Fukumizu, K., Gretton, A., Schölkopf, B., "Kernel mean shrinkage estimators," *Journal of Machine Learning Research*, vol. 17, no. 48, pp. 1–41, 2016.

[3] Tolstikhin, I., Sriperumbudur, B. K., **Muandet, K.,** "Minimax estimation of kernel mean embeddings," *Journal of Machine Learning Research*, vol. 18, no. 86, pp. 1–47, 2017.

[4] Shah, N. B., Tabibian, B., **Muandet, K.**, Guyon, I., Luxburg, U., "Design and analysis of the nips 2016 review process," *Journal of Machine Learning Research*, vol. 19, no. 49, pp. 1–34, 2018.

[5] Lopez-Paz, D., **Muandet, K.**, Recht, B., "The randomized causation coefficient," *Journal of Machine Learning Research*, vol. 16, no. 90, pp. 2901–2907, 2015.

[6] **Muandet, K.**, Fukumizu, K., Sriperumbudur, B., Schölkopf, B., "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends in Machine Learning*, vol. 10, no. 1-2, pp. 1–141, 2017.

[7] **Muandet, K.**, Fukumizu, K., Dinuzzo, F., Schölkopf, B., "Learning from distributions via support measure machines," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 10–18.

[8] **Muandet, K.**, Sriperumbudur, B., Schölkopf, B., "Kernel mean estimation via spectral filtering," in *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., 2014, pp. 10–18.

[9] **Muandet, K.**, Mehrjou, A., Lee, S. K., Raj, A., "Dual instrumental variable regression," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 2710–2721.

[10] Kübler, J., Jitkrittum, W., Schölkopf, B., **Muandet, K.,** "Learning kernel tests without data splitting," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 6245–6255.

[11] Chen, X., Wang, Z., Tang, S., **Muandet, K.,** "MATE: Plugging in model awareness to task embedding for meta learning," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 11 865–11 877.

[12] Park, J., **Muandet, K.,** "A measure-theoretic approach to kernel conditional mean embeddings," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 21 247–21 259.

[13] **Muandet, K.**, Balduzzi, D., Schölkopf, B., "Domain generalization via invariant feature representation," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, JMLR, 2013, pp. 10–18.

[14] Zhang, K., Schölkopf, B., **Muandet, K.**, Wang, Z., "Domain adaptation under target and conditional shift," in *Proceedings of the 30th International Conference on Machine Learning*, PMLR, 2013, pp. 819–827.

[15] **Muandet, K**, Fukumizu, K, Sriperumbudur, B, Gretton, A, Schölkopf, B, "Kernel mean estimation and Stein effect," in *Proceedings of The 31st International Conference on Machine Learning*, vol. 32, JMLR, 2014, pp. 10–18.

[16] Lopez-Paz, D., **Muandet, K.**, Schölkopf, B., Tolstikhin, I., "Towards a learning theory of cause-effect inference," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, vol. 37, JMLR, 2015, pp. 1452–1461.

[17] Park, J., Shalit, U., Schölkopf, B., **Muandet, K.,** "Conditional distributional treatment effect with kernel conditional mean embeddings and U-statistic regression," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[18] Mastouri, A., Zhu, Y., Gultchin, L., Korba, A., Silva, R., Kusner, M. J., Gretton, A., **Muandet, K.,** "Proximal causal learning with kernels: Two-stage estimation and moment restriction," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, forthcoming, 2021.

[19] **Muandet, K.**, Schölkopf, B., "One-class support measure machines for group anomaly detection," in *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, AUAI Press, 2013, pp. 449–458.

[20] Doran, G., **Muandet, K.**, Zhang, K., Schölkopf, B., "A permutation-based kernel conditional independence test," in *30th Conference on Uncertainty in Artificial Intelligence (UAI2014)*, 2014, pp. 132–141.

[21] **Muandet, K.**, Jitkrittum, W., Kübler, J., "Kernel conditional moment test via maximum moment restriction," in *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, ser. Proceedings of Machine Learning Research, vol. 124, PMLR, 2020, pp. 41–50.

[22] Schuster, I., Mollenhauer, M., Klus, S., **Muandet, K.,** "Kernel conditional density operators," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 108, PMLR, 2020, pp. 993–1004.

[23] Kilbertus, N., Rodriguez, M. G., Schölkopf, B., **Muandet, K.**, Valera, I., "Fair decisions despite imperfect predictions," in *AISTATS*, vol. 108, PMLR, 2020, pp. 277–287.

[24] Zhang, Y., Tang, S., **Muandet, K.**, Jarvers, C., Neumann, H., "Local temporal bilinear pooling for fine-grained action parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[25] Karunratanakul, K., Yang, J., Zhang, Y., Black, M., **Muandet, K.**, Tang, S., "Grasping field: Learning implicit representations for human grasps," in *8th International Conference on 3D Vision*, IEEE, Nov. 2020, pp. 333–344.

[26] Kübler, J. M., **Muandet, K.**, Schölkopf, B., "Quantum mean embedding of probability distributions," *Physical Review Research*, vol. 1, no. 3, p. 033 159, 2019.

[27] Klus, S., Schuster, I., **Muandet, K.,** "Eigendecompositions of transfer operators in reproducing kernel hilbert spaces," *Journal of Nonlinear Science*, vol. 30, no. 1, pp. 283–315, 2020.

[28] Berlinet, A., Thomas-Agnan, C., *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

[29] Smola, A. J., Gretton, A., Song, L., Schölkopf, B., "A Hilbert space embedding for distributions," in *Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT)*, Springer-Verlag, 2007, pp. 13–31.

[30] Schölkopf, B., **Muandet, K.**, Fukumizu, K., Peters, J., "Computing functions of random variables via reproducing kernel Hilbert space representations," *Statistics and Computing*, vol. 25, no. 4, pp. 755–766, 2015.

[31] Zhu, J.-J., **Muandet, K.**, Diehl, M., Schölkopf, B., "A new distribution-free concept for representing, comparing, and propagating uncertainty in dynamical systems with kernel probabilistic programming," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 7240–7247, 2020, 21th IFAC World Congress.

[32] Zhang, R., Imaizumi, M., Schölkopf, B., **Muandet, K.,** "Maximum moment restriction for instrumental variable regression," *ArXiv Preprint*, 2020. [Online]. Available: https://arxiv.org/abs/2010.07684.

[33] Schölkopf, B., Herbrich, R., Smola, A. J., "A generalized representer theorem," in *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, ser. COLT '01/EuroCOLT '01, Springer-Verlag, 2001, pp. 416–426.

[34] Blanchard, G., Lee, G., Scott, C., "Generalizing from several related classification tasks to a new unlabeled sample," in *Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 2178–2186.

[35] LaLonde, R. J., "Evaluating the econometric evaluations of training programs with experimental data," *The American Economic Review*, vol. 76, no. 4, pp. 604–620, 1986.

[36] Hartford, J., Lewis, G., Leyton-Brown, K., Taddy, M., "Deep IV: A flexible approach for counterfactual prediction," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, PMLR, 2017, pp. 1414–1423.

[37] Singh, R., Sahani, M., Gretton, A., "Kernel instrumental variable regression," in *Advances in Neural Information Processing Systems*, 2019, pp. 4595–4607.

[38] Bennett, A., Kallus, N., Schnabel, T., "Deep generalized method of moments for instrumental variable analysis," in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 3564–3574.

[39] Dikkala, N., Lewis, G., Mackey, L., Syrgkanis, V., "Minimax estimation of conditional moment models," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 12 248–12 262.

[40] Liao, L., Chen, Y.-L., Yang, Z., Dai, B., Kolar, M., Wang, Z., "Provably efficient neural estimation of structural equation models: An adversarial approach," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 8947–8958.

[41] Newey, W., "Efficient estimation of models with conditional moment restrictions," in *Handbook of Statistics*, vol. 11, 1993, ch. 16, pp. 419–454.

[42] Ai, C., Chen, X., "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, vol. 71, no. 6, pp. 1795–1843, 2003.

[43] Kleinberg, J., Raghavan, M., "Algorithmic classification and strategic effort," *SIGecom Exch.*, vol. 18, no. 2, pp. 45–52, 2020.

[44] Hardt, M., Megiddo, N., Papadimitriou, C. H., Wootters, M., "Strategic classification," in *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, M. Sudan, Ed., ACM, 2016, pp. 111–122.

[45] Kleinberg, J., Raghavan, M., "How do classifiers induce agents to invest effort strategically?" In *Proceedings of the 2019 ACM Conference on Economics and Computation*, Association for Computing Machinery, 2019, pp. 825–844.

[46] Miller, J., Milli, S., Hardt, M., "Strategic classification is causal modeling in disguise," in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, PMLR, 2020, pp. 6917–6926.

[47] Perdomo, J., Zrnic, T., Mendler-Dünner, C., Hardt, M., "Performative prediction," in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, PMLR, 2020, pp. 7599–7609.