Evaluating and Correcting Performative Effects of Decision Support Systems via Causal Domain Shift

Philip Boeken P.A.BOEKEN@UVA.NL
University of Amsterdam
Booking.com

Onno Zoeter ONNO.ZOETER@BOOKING.COM
Booking.com

Joris M. Mooij J.M.MOOIJ@UVA.NL
University of Amsterdam

Editors: Francesco Locatello and Vanessa Didelez

Abstract

When predicting a target variable Y from features X, the prediction $\hat{Y}$ can be *performative*: an agent might act on this prediction, affecting the value of Y that we eventually observe. Performative predictions are deliberately prevalent in algorithmic decision support, where a Decision Support System (DSS) provides a prediction for an agent to affect the value of the target variable. When deploying a DSS in high-stakes settings (e.g. healthcare, law, predictive policing, or child welfare screening) it is imperative to carefully assess the performative effects of the DSS. In the case that the DSS serves as an alarm for a predicted negative outcome, naive retraining of the prediction model is bound to result in a model that underestimates the risk, due to effective workings of the previous model. In this work, we propose to model the deployment of a DSS as causal domain shift and provide novel cross-domain identification results for the conditional expectation $\mathbb{E}[Y \mid X]$, allowing for pre- and post-hoc assessment of the deployment of the DSS, and for retraining of a model that assesses the risk under a baseline policy where the DSS is not deployed. Using a running example, we empirically show that a *repeated regression* procedure provides a practical framework for estimating these quantities, even when the data is affected by sample selection bias and selective labelling, offering for a practical, unified solution for multiple forms of target variable bias.

Keywords: Performative Prediction, Decision Support Systems, Domain Adaptation, Causal Modelling, Evaluation, Bias Correction.

1. Introduction

When the value of some variable is predicted, this prediction can cause an agent to take action. In the context of linguistics, Austin (1962) coined the term *performative* for utterances that aim at instigating action; in contrast with sentences of a *descriptive* nature. In economics, the concept of performativity has received much attention, and has seen multiple different manifestations. For example, it has been described as a more general concept where the emergence of economic theories legitimize the markets they describe, which caused these markets to become more active. A very concrete type of performativity has been observed in the common use of the Black-Scholes-Merton (BSM) formula for predicting option prices, which in turn affects the price of said options to be close to their predicted value (MacKenzie et al., 2007). Related notions are that of a self-fulfilling prophecy, like the BSM formula, and the self-defeating prophecy, like warning of excessive risk that instigates action to reduce this risk.

In recent machine learning literature, much attention has been given to the performative effects of predictions (Perdomo et al., 2020; Mendler-Dünner et al., 2020; Miller et al., 2021; Pombal et al., 2022; Kim and Perdomo, 2023; Yan et al., 2023, among others). A common goal of these works is to make a prediction that is close to the value that will be observed, taking into account the effect that the prediction has on this target variable. Here, a core concept is the minimization of the *performative risk*: the risk of a prediction model, evaluated on the data distribution it entails.

In this work we place the problem of performative prediction in the light of human-algorithmic decision making, where predictions are deliberatively of a performative nature, but do not necessarily have to be close to the eventually observed target variable. For example, algorithms that warn of excessive risk (e.g. in churn prediction, predictive policing, or patient monitoring in the ICU) aim at instigating an action that will reduce the predicted risk and thus aim at invalidating the prediction that they make. Such models can be considered to predict risk under the baseline policy where the decision support system (DSS) is not deployed (Coston et al., 2020). Naive retraining of such prediction models can suffer from a bias that is induced in the training data by the previous prediction algorithm, a concept that we refer to as *performative bias*. In this work, we show how to correct for performative bias by explicitly modelling the deployment of the DSS, and treating the estimation of the *baseline predictor* as a domain adaptation problem.

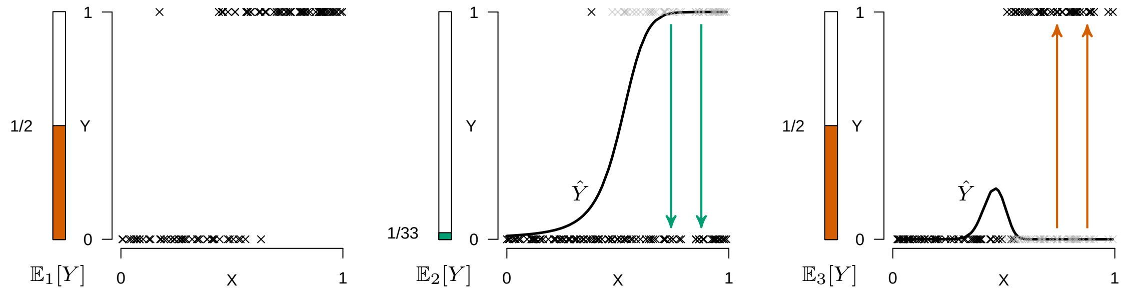
In aforementioned high-stakes environments, proper evaluation of the DSS is crucial. Over the years many decision support systems have been deployed in high stakes environments, but not all to great success (Coston et al., 2023). These events motivate thorough testing of any DSS prior to deployment and thorough examination of the system during deployment, to foresee and monitor any undesirable performative effects of the DSS. Despite the urgency of proper continuous assessment of decision support systems, Wu et al. (2021) show that among all medical AI devices that are approved by the FDA between January 2015 and December 2020, most evaluations of those devices are pre-deployment studies, and hardly any post-deployment evaluations have been performed.[1] To address the need of evaluation of DSSes, we propose and investigate the estimability of the *deployment effect*, i.e. the effect of the deployment of the DSS on the target variable, and of the *retraining effect*, i.e. the effect of a new prediction model on the target variable, compared to the average outcome under the previous prediction model. In practise it can be unfeasible or unethical to perform randomized controlled trials with the deployment of a DSS, which makes the estimability of these evaluation metrics a domain adaptation problem.

In the following numerical example we further demonstrate the manifestation of performative bias after naive retraining of a prediction model, and with it the need of its evaluation, e.g. by analysis of the retraining effect. This example is inspired by a real-world scenario where in the training data, high-risk individuals receive a treatment that effectively lowers the risk of a negative outcome, inducing a bias in the training data (Caruana et al., 2015).

Example 1 *Let* $X \sim \mathrm{Unif}[0,1], \hat{Y} = f(X)$ *for some function* f*, and* $Y \sim \mathrm{Ber}(\sigma(X - 1/2)\mathbb{1}\{\hat{Y} < 1/2\})$ *with* $\sigma(x) = (1 + e^{-x})^{-1}$*. Three 'epochs' (samples) of this data generating process are shown in Figure 1. In the first epoch, the DSS is not deployed, so we let* $\hat{Y} \equiv 0$*. In the second epoch, a*

[1] Although not *all* medical AI devices that are considered by Wu et al. (2021) provide explicit decision support, many can be interpreted to do so. For example, image classification techniques for detecting tumours can be seen as providing decision support, and evaluation of the performative effects of the deployment of such AI devices is likely of importance.

DSS that is trained on data from epoch one is deployed, so we let $\hat{Y} = \hat{\mathbb{E}}_1[Y|X]$,[2] estimated from $\mathbb{P}_1(X,Y)$. To units where the predicted risk $\hat{Y}$ exceeds the threshold $1/2$, action is taken to greatly reduce this risk, effectively setting $\mathbb{P}(Y = 1 \mid X, \{\hat{Y} > 1/2\}) = 0$ and thus preventing the outcome $Y = 1$. The green arrows signify this positive effect, marking the grey-coloured counterfactual observations that would have had the value $Y = 1$ if no prediction had been made, and which have the value $Y = 0$ now that the prediction is made. This improvement is also indicated by the bar charts, showing $\mathbb{E}_2[Y] \approx 1/33 < 1/2 \approx \mathbb{E}_1[Y]$. In the third epoch, a DSS is naively trained on data from the second epoch, resulting in a model that underestimates the risk, due to effective workings of the DSS in the previous epoch. Hence, the re-training has a negative effect on the average outcome, as indicated by the red arrows and by the bar chart for $\mathbb{E}_3[Y]$.



(*a*) $t = 1$, DSS off (*b*) $t = 2, \hat{Y} = \mathbb{E}_1(Y \mid X)$ (*c*) $t = 3, \hat{Y} = \mathbb{E}_2(Y \mid X)$

Figure 1: In epoch $t = 1$ the DSS is not deployed. In $t = 2$ a DSS $\hat{Y}$ is deployed that is trained on data from $t = 1$, effectively reducing the mean of Y. In $t = 3$, a DSS $\hat{Y}$ that is naively retrained on data from $t = 2$ is deployed, increasing the mean of Y.

Contributions In this work, we model the deployment of decision support systems as causal domain shift, and we investigate two applications of this causal model. The first application is the evaluation of whether a novel DSS should be deployed, an existing DSS should be taken offline, or whether a retrained version should be deployed. We define the *deployment effect* and *retraining effect* as suitable evaluation metrics, and we show that the estimation of these evaluation metrics constitutes two domain adaptation tasks (T1 and T2). The second application concerns the estimation of a prediction model to be used by the DSS. We show that naive retraining of such prediction models gets affected by *performative bias* that is induced by the previous prediction model; correcting for this bias constitutes another domain adaptation problem (T3). We show that these domain adaptation tasks are mathematically equivalent, and that they are not solvable (without additional assumptions besides the causal model) when one cannot perform randomized experiments with the deployment of the DSS. We define a *domain pivot* as a set of variables that, when measured in both the source- and target domain of the domain adaptation problem, provides a solution to the domain adaptation problems T1–3, and hence to the evaluation and bias correction applications. We employ the *repeated regression* estimator from Boeken et al. (2023) for estimating the quantities of interest. As this

[2] We let $\hat{\mathbb{E}}[Y \mid X]$ denote an estimate of the conditional expectation $\mathbb{E}[Y \mid X]$. This should interchangeably be interpreted as the function $x \mapsto \hat{\mathbb{E}}[Y \mid X = x]$ or as the evaluation $\hat{\mathbb{E}}[Y \mid X = X]$.

estimator has originally been devised to deal with selection bias, we generalise the identifiability and estimation results to settings that are subject to selection bias and/or selective labelling (missing response). Efficacy of these methods is subsequently shown using Example 1.

1.1. Related work

A line of work following from Perdomo et al. (2020) considers the general setting where model parameters θ for making a prediction $\hat{\mathbb{E}}_\theta[Y \mid X]$ induce a shift of the distribution of (X, Y). This dependence can be made explicit by writing $\mathbb{P}(X, Y \mid \theta)$. Similar to Mendler-Dünner et al. (2022) and Kim and Perdomo (2023), we consider the specific setting of *outcome performativity* where the parameters don't affect the features X but only the outcome Y, so where the distribution factorizes according to $\mathbb{P}(X, Y \mid \theta) = \mathbb{P}(Y \mid X, \theta)\mathbb{P}(X)$, and conditional on the parameters the (X, Y) pairs are drawn i.i.d. This is a different setup than e.g. Chen et al. (2023) consider, as they allow effects like $\hat{Y}_i \to X_j$ and $\hat{Y}_i \to Y_j$, where $i \neq j$ are sample indicators. We extend the setting of Mendler-Dünner et al. (2022) and Kim and Perdomo (2023) by explicitly considering the domain where the DSS is not deployed, allowing for the formulation of the domain adaptation task that we consider. For more details we refer to Appendix A.1.

The task of transporting a statistical relation $\mathbb{E}[Y \mid X]$ over such domains is considered in the line of work on *transportability*. Similar to Pearl and Bareinboim (2011) and Magliacane et al. (2018) we leverage sets of variables that render a target variable independent from a domain indicator (which we refer to as *domain pivots*) to transfer statistical relations over domains. Sound and complete algorithms for transporting statistical relations are for example given by Correa and Bareinboim (2019) and Lee et al. (2020). However, these algorithms make weaker assumptions than we do (in terms of available data), which makes them unable to identify the target quantities that we consider.

The work of Coston et al. (2020) considers risk estimation under binary treatment, similar to how we estimate the effect of deployment on an outcome variable Y. However, their estimation method requires stronger assumptions on the available data than we do, making it unsuitable for the setting that we consider.

In special cases where the utilized *domain pivot* consists of a context X and an *action variable* (with a finite state space) that the agent controls to optimize a reward Y, our proposed evaluation method can be interpreted as a form of off-policy evaluation for contextual bandits, as investigated by Dudík et al. (2014); Wang et al. (2017). We elaborate on this connection in Appendix A.2.

To estimate our quantities of interest, we employ the *repeated regression* estimator from Boeken et al. (2023). This estimator is originally proposed to correct for selection bias. In this work, we show that this estimator can simultaneously correct for selection bias *and* performative bias. The repeated regression estimator bears resemblance to the work on surrogate indices by Athey et al. (2019), as both methods consider the use of a conditional expectations as pseudo-labels in the estimation procedure. However, translating it to our setting, the work on surrogate indices operates under a different set of assumptions than we do, as it requires the target variable Y to be measured under both deployment and non-deployment. More details are provided in Appendix A.3.

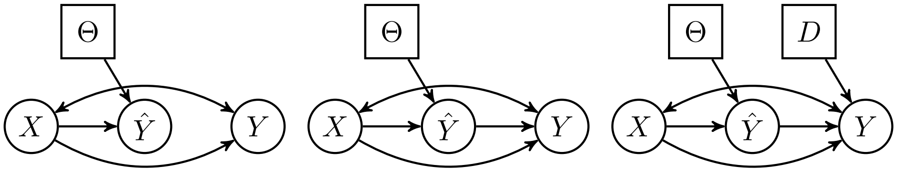
2. Causal modelling of decision support systems

We consider the setting with multidimensional covariates X, univariate target variable Y, and a prediction $\hat{Y}$ of Y that is a function of X and parameters Θ, denoted with $\hat{Y} = f_{\hat{Y}}(X, \Theta)$. We allow for hidden confounding between X and Y. We assume the state space of any variable V to be (a subset of) $\mathbb{R}^{d_V}$ for some $d_V \in \mathbb{N}$, equipped with its standard topology and the Borel sigma-algebra. Variables, their state space, and their values are indicated with uppercase, calligraphic and lowercase letters $(V, \mathcal{V}, v)$ respectively.

To distinguish between pre- and post-deployment settings of the decision support system, we introduce a domain indicator D which represents a context-specific dependency: $D = 0$ indicates the domain where $\hat{Y}$ does not affect Y (i.e. the prediction is not published), and $D = 1$ indicates the domain where $\hat{Y}$ affects Y. Formally, the structural causal model[3] to describe this data generating process is

$$X = f_X(E_X, E_{XY}), \quad \hat{Y} = f_{\hat{Y}}(X, \Theta), \quad Y = \begin{cases} f_{Y,0}(X, E_{XY}, E_Y) & \text{if } D = 0 \\ f_{Y,1}(X, \hat{Y}, E_{XY}, E_Y) & \text{if } D = 1 \end{cases} \tag{1}$$

with independent exogenous distributions $\mathbb{P}(E_X), \mathbb{P}(E_Y)$ and $\mathbb{P}(E_{XY})$ and measurable functions $f_X, f_{\hat{Y}}, f_{Y,0}, f_{Y,1}$. The Acyclic Directed Mixed Graph (ADMG) of this SCM in the different domains D is depicted in Figures 2(*a*) and 2(*b*), and the causal graph of the joint model is depicted in Figure 2(*c*), with explicit domain indicator D.



(*a*) $D = 0$ (*b*) $D = 1$ (*c*) Causal graph with domain indicator D.

Figure 2: Modelling the deployment of the DSS with prediction $\hat{Y}$ as domain shift.

As we don't assume an a-priori distribution for the variables D and Θ, these variables are graphically indicated with squared nodes and formally referred to as *input nodes* of the SCM, following Forré (2021). For given values of D and Θ, this SCM gives rise to the Markov kernel $\mathbb{P}(X, \hat{Y}, Y \mid \text{do}(D, \Theta))$, defined as the pushforward of the exogenous distribution $\mathbb{P}(E_X) \otimes \mathbb{P}(E_Y) \otimes \mathbb{P}(E_{XY})$ through the structural equations.[4] We assume that data will be sampled over epochs, indicated by t, where for given values θ_t, d_t we sample $(X_{t,i}, \hat{Y}_{t,i}, Y_{t,i}) \sim \mathbb{P}(X, \hat{Y}, Y \mid \text{do}(D = d_t, \Theta = \theta_t))$ i.i.d. for $i = 1, ..., n_t$ and some $n_t \in \mathbb{N}$. Note that this implies that D and Θ are neither influenced by, nor confounded with the variables $(X_{t,i}, \hat{Y}_{t,i}, Y_{t,i})$ of the current epoch. Denoting measurements $(V_{t,1}, ..., V_{t,n_t})^T$ of a variable V with $\boldsymbol{V}_t$, the values θ_t and d_t can be determined by data from past epochs $\{(\boldsymbol{X}_s, \hat{\boldsymbol{Y}}_s, \boldsymbol{Y}_s, \theta_s, d_s) : s < t\}$.

When considering SCMs with more variables than $\{X, \hat{Y}, Y, D, \Theta\}$, we require the latent projection onto $\{X, \hat{Y}, Y, D, \Theta\}$ to be (a subgraph of) the graph from Figure 2(*c*) for it to appropriately represent the deployment of a DSS.

Assumption 1 *We consider the set $\mathcal{M}$ of SCMs with endogenous variables $V \supseteq \{X, \hat{Y}, Y\}$, input variables $\{D, \Theta\}$ and graph G such that $\mathrm{Pa}_G(\hat{Y}) = \{X, \Theta\}$ and $\mathrm{Ch}_G(D) = \mathrm{Ch}_G(\hat{Y})$, and such*

[3] We will use many concepts from this causal framework: parents, children, ancestors, d-separation, the Markov property, etc. For more information, we refer to Pearl (2009) and Bongers et al. (2021).

[4] We define $\mathbb{P}(X, Y \mid \text{do}(D = 0))$ (without dependence on Θ) similarly but with the structural equation for Y evaluated at $D = 0$, for which we have $\mathbb{P}(X, Y \mid \text{do}(D = 0)) = \mathbb{P}(X, Y \mid \text{do}(D = 0, \Theta = \theta))$ for all θ.

that the graph of the latent projection of G onto $\{X, \hat{Y}, Y, D, \Theta\}$ is a subgraph of the ADMG in Figure 2(c).

As alluded to in Section 1, we are interested in the evaluation of the DSS prior to- and during its deployment, and in correcting for the bias that is induced by a previous deployment of the DSS when retraining the prediction model. Having explicitly defined the deployment indicator D and parameters Θ, we are enabled to make these estimation tasks more precise.

2.1. Application A: Evaluation

When a DSS with some parameter value θ and prediction model $\hat{Y} = f_{\hat{Y}}(X, \theta)$ has been developed, the intention behind this DSS is to improve the value of some outcome metric Y. As motivated in the introduction, human usage of a newly developed DSS can involve errors that have a negative impact on Y. To evaluate this, we define the *deployment effect* of a DSS with parameters θ.

Definition 2 (Deployment effect) *The* deployment effect *of a DSS with parameters θ is defined as the average causal effect of the deployment of the DSS on the target variable, i.e.*

$$\tau(\theta) := \mathbb{E}[Y \,|\, \mathrm{do}(D = 1, \Theta = \theta)] - \mathbb{E}[Y \,|\, \mathrm{do}(D = 0)]. \tag{2}$$

Prior to deployment of the DSS we are interested in estimating $\tau(\theta)$ from data sampled from $\mathbb{P}(X, Y \,|\, \mathrm{do}(D = 0))$. Since $\mathbb{E}[Y \,|\, \mathrm{do}(D = 0)]$ is directly estimable, the challenge lies in estimating $\mathbb{E}[Y \,|\, \mathrm{do}(D = 1, \Theta = \theta)]$. We refer to this domain adaptation task as **T1.a**. After deployment, we are interested in estimating $\tau(\theta)$ from $\mathbb{P}(X, Y \,|\, \mathrm{do}(D = 1, \Theta = \theta))$, e.g. to monitor the correct usage of the DSS: if the mean value of Y is estimated to be worse for $(D = 1, \Theta = \theta)$ than for $D = 0$, it could be better to turn off the DSS, and further investigate why it has a negative effect on the outcome. Since $\mathbb{E}[Y \,|\, \mathrm{do}(D = 1, \Theta = \theta)]$ is directly estimable, the challenge then lies in estimating $\mathbb{E}[Y \,|\, \mathrm{do}(D = 0)]$. We refer to this domain adaptation task as **T1.b**.

When in epoch t a DSS with parameters θ_t is deployed one might further develop the DSS, resulting in parameters θ_{t+1}. Before deploying this 'retrained' DSS, one might want to evaluate the impact that these new parameters will have on Y.

Definition 3 (Retraining effect) *The* retraining effect *is defined as the average causal effect of the deployment of a retrained DSS on the target variable, i.e.*

$$\rho(\theta_{t+1}, \theta_t) := \mathbb{E}[Y \,|\, \mathrm{do}(D = 1, \Theta = \theta_{t+1})] - \mathbb{E}[Y \,|\, \mathrm{do}(D = 1, \Theta = \theta_t)]. \tag{3}$$

In the setting described above, we aim at estimating $\rho(\theta_{t+1}, \theta_t)$ from $\mathbb{P}(X, Y \,|\, \mathrm{do}(D = 1, \Theta = \theta_t))$. Since $\mathbb{E}[Y \,|\, \mathrm{do}(D = 1, \Theta = \theta_t)]$ is directly estimable, the challenge lies in estimating $\mathbb{E}[Y \,|\, \mathrm{do}(D = 1, \Theta = \theta_{t+1})]$. We refer to this domain adaptation task as **T2**. We have summarised these domain adaptation tasks in Table 1.

2.2. Application B: Bias correction

Let Y be an outcome whose expected value we want to minimize, e.g. a cost, negative utility, or negative reward. Prior to deployment, data is generated from $\mathbb{P}(X, Y \,|\, \mathrm{do}(D = 0))$, and the average outcome Y is related to features X via $\mathbb{E}[Y \,|\, X, \mathrm{do}(D = 0)]$. This could be considered to be a

	Metric	Source domain	Target domain	Target quantity
T1.a	$\tau(\theta)$	$D = 0$	$D = 1, \Theta = \theta$	$\mathbb{E}[Y \mid \mathrm{do}(D = 1, \Theta = \theta)]$
T1.b	$\tau(\theta)$	$D = 1, \Theta = \theta$	$D = 0$	$\mathbb{E}[Y \mid \mathrm{do}(D = 0)]$
T2	$\rho(\theta_{t+1}, \theta_t)$	$D = 1, \Theta = \theta_t$	$D = 1, \Theta = \theta_{t+1}$	$\mathbb{E}[Y \mid \mathrm{do}(D = 1, \Theta = \theta_{t+1})]$
T3	–	$D = 1, \Theta = \theta$	$D = 0$	$\mathbb{E}[Y \mid X, \mathrm{do}(D = 0)]$

Table 1: Domain adaptation tasks for evaluation (T1, T2) and performative bias correction (T3).

'baseline policy'. It might be the case that a DSS is developed to identify cases that have (under this baseline policy) a high risk of seeing an outcome that is to be prevented, like a patients death, a customer churning, or a crime to be committed in a particular neighbourhood (Coston et al., 2020). In this setting, a sensible predictor $\hat{Y}$ would be the following:

Definition 4 (Baseline predictor) *We are interested in estimating the* baseline predictor

$$\hat{Y}_{bp} = \mathbb{E}[Y \mid X, \mathrm{do}(D = 0)]. \tag{4}$$

As demonstrated in Example 1, naive regression of Y on X to retrain the model for $\hat{Y}$ from $\mathbb{P}(X, Y \mid \mathrm{do}(D = 1, \Theta))$ would yield a predictor $\hat{Y} = \mathbb{E}[Y \mid X, \mathrm{do}(D = 1, \Theta)]$, and hence is biased when the DSS is supposed to make the baseline prediction.

Definition 5 (Performative bias) *The* performative bias *is defined as the bias that the deployment of the DSS induces on the statistical relation* $\mathbb{E}[Y \mid X]$*, i.e.*

$$\mathbb{E}[Y \mid X, \mathrm{do}(D = 1, \Theta = \theta)] - \mathbb{E}[Y|X, \mathrm{do}(D = 0)]. \tag{5}$$

Estimating the baseline predictor from the domain $(D = 1, \Theta = \theta)$, and thus correcting for performative bias, is a domain adaptation task that we refer to as **T3**.

If we let Y be binary, the baseline predictor is indeed the optimal prediction function $\hat{Y} : \mathcal{X} \to [0, 1]$ if it is a risk assessment for the event $Y = 1$ (given features X) and serves as an 'alarm' to identify 'risky cases', based on which an agent can take an action A which surely decreases the risk to a known level, but which one also wants to use sparingly. The action A could for example be to operate a patient with features X to minimize the probability of death Y, or the offering of a discount A to a customer X to minimize the probability of churning Y. Clearly, one wants to use these actions sparingly. This type of optimality of the baseline predictor is formalised as follows:

Proposition 6 *Given a Markov kernel* $\mathbb{P}(Y = 1|X, A)$*, consider the SCM* $X \sim \mathbb{P}(X), A = D \cdot \mathbb{1}\{\hat{Y} > \varepsilon(X)\}, \hat{Y} = \hat{y}(X), Y \sim \mathbb{P}(Y = 1 \mid X, A)$ *with* $\varepsilon(x) := \mathbb{P}(Y = 1 \mid X = x, A = 1)$ *and some function* $\hat{y} : \mathcal{X} \to [0, 1]$*. The baseline predictor* $\hat{Y}_{bp}$ *solves the following bilevel optimization problem:*

$$\begin{aligned} &1.\ H := \underset{\hat{y} \in [0,1]^{\mathcal{X}}}{\arg\min}\, \mathbb{P}(Y = 1 \mid X = x, \mathrm{do}(D = 1, \hat{Y} = \hat{y}(x))) \\ &2.\ \hat{Y}_{bp} \in \underset{\hat{y} \in H}{\arg\min}\, \mathbb{P}(A = 1 \mid X = x, \mathrm{do}(D = 1, \hat{Y} = \hat{y}(x))) \end{aligned} \tag{6}$$

for $\mathbb{P}(X)$*-almost all* $x \in \mathcal{X}$*.*

2.3. Equivalence of T1–3 and their non-identifiability

Having these two applications in mind, our goal is to estimate the deployment effect $\tau(\theta)$, the retraining effect $\rho(\theta_{t+1}, \theta_t)$, and the baseline predictor $\hat{Y}_{bp}$ from varying source domains, and thus solving domain adaptation tasks T1–3 as displayed in Table 1. A prerequisite for estimation is the *identifiability* of these quantities: whether there exists a mathematical operation on the source distribution that yields the target quantity. This concept is formally defined as follows:

Definition 7 (Identifiability) *Given a set of SCMs $\mathcal{M}$, a target quantity $t(M)$ (some function of $M \in \mathcal{M}$) is* identifiable *in $\mathcal{M}$ from a set $s(M) := \{s_1(M), ..., s_n(M)\}$ of (marginal, conditional and/or interventional) distributions induced by M if $s(M_1) = s(M_2) \implies t(M_1) = t(M_2)$ for all $M_1, M_2 \in \mathcal{M}$.*

Throughout, we will consider $\mathcal{M}$ as defined in Assumption 1. Given source distribution(s) $s(M)$ and target quantity $t(M)$, identifiability means that for all $M \in \mathcal{M}$ the map $s(M) \mapsto t(M)$ is well defined, which can be shown by explicitly providing it. Non-identifiability can be shown by providing SCMs $M_1, M_2 \in \mathcal{M}$ for which $s(M_1) = s(M_2)$ but $t(M_1) \neq t(M_2)$. Throughout, our target $t(M)$ will be a conditional expectation, and we will specify $s(M)$ and $t(M)$ without explicit dependence on M.

We first consider the graph depicted in Figure 2(*c*) and we show that tasks T1–3 are equivalent to a single domain adaptation task:

Lemma 8 *Identifiability of each of the target quantities of the domain adaptation tasks T1–3 is equivalent to identifiability of the conditional expectation $\mathbb{E}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)]$ from $\mathbb{P}(X, Y | \mathrm{do}(D = d', \Theta = \theta'))$ for some $(d', \theta') \neq (d, \theta)$ with $d' = d = 1 \iff \theta' \neq \theta$.*

The last condition on the domains ensures that we don't consider $d = d' = 0$ and $\theta \neq \theta'$, which is trivially excluded as the distribution of (X, Y) would then be the same in the source and target domains.

The following proposition shows that measuring a single source distribution is not sufficient for identification of the target quantities as specified above, and with that, that the tasks T1–3 cannot be solved without imposing additional assumptions.

Proposition 9 *The target quantity $\mathbb{E}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)]$ is not identifiable in $\mathcal{M}$ from $\mathbb{P}(X, Y \mid \mathrm{do}(D = d', \Theta = \theta'))$ if $(d', \theta') \neq (d, \theta)$ with $d' = d = 1 \iff \theta' \neq \theta$.*

It is immediately clear that if (X, Y) pairs are measured in both the source and target domains, then $\mathbb{E}[Y|X, \mathrm{do}(D = d, \Theta = \theta)]$ would be identifiable, and hence also $\tau(\theta), \rho(\theta_{t+1}, \theta_t)$ and $\hat{Y}_{bp}$. However, in practice it might not be feasible to gather labels from the target domain. Alas, in high-stakes settings, deploying a DSS without knowing what effect it will have on the outcome Y can be undesirable. So, to be able to identify these target quantities without requiring measurements of the outcome variable Y in the target domain we will leverage additional assumptions, as demonstrated in the next section.

2.4. Domain pivots: mediators of the prediction and outcome

When a prediction $\hat{Y}$ affects the value Y, this might happen through an *action* A that perfectly mediates $\hat{Y}$ and Y, i.e. we have $\hat{Y} \to A \to Y$ and there is no edge $\hat{Y} \to Y$. In this case, only A is directly affected by the deployment of the DSS, so we have $D \to A$ and not $D \to Y$. If this action A is unconfounded with Y, we have the independence $Y \perp\!\!\!\perp D, \Theta \mid X, A$. If A and Y are confounded (by C, say), we have $Y \perp\!\!\!\perp D, \Theta \mid X, Z$ where $Z = \{A, C\}$. A graphical depiction of this setting is provided in Figure 3. As we will see later, finding a set of variables Z for which this conditional independence is satisfied is instrumental for the domain adaptation task that we have in mind. Note that Figure 3 is not the only graph that exhibits $Y \perp\!\!\!\perp D, \Theta \mid X, C, A$, as we can add multiple instances of latent confounding (bidirected edges) to this graph and maintain the required independence.

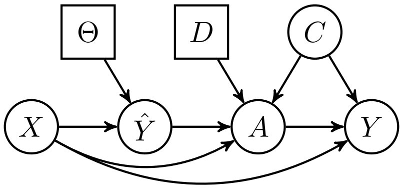


Figure 3: Performative prediction through a mediator A, with an observed common cause C.

Definition 10 (Domain pivot) *Given domain indicator D, features X, target variable Y and estimand $F(\mathbb{P}(Y|X, D = d))$ with F a statistical functional*[5] *and $d \in \mathcal{D}$, we call $\{X, Z\}$ a* domain pivot *for $F(\mathbb{P}(Y|X, D = d))$ if $Y \perp\!\!\!\perp D \mid X, Z$.*

Our main solution for T1–3 assumes that the domain pivot $\{X, Z\}$ can be measured in the target domain. Sampling from $\mathbb{P}(X, Z \mid \mathrm{do}(D = 1, \Theta))$ prior to deployment of the DSS might seem like an unreasonable assumption, but there can be practical ways to do so. Consider the example of a patient with features X, and a doctor having to decide treatment A, where $\{X, A\}$ is a domain pivot. Here, $\mathbb{P}(A|X, \hat{Y}, \mathrm{do}(D = 1, \Theta))$ can be measured by showing the doctor the prediction $\hat{Y}$, and measuring the treatment that the doctors prescribe for this patient after seeing this prediction. Measuring such intended actions is also leveraged by Stensrud et al. (2023) to improve treatment regimes. Practically, one would require the availability of another doctor who has not seen the prediction of the DSS to prescribe the treatment that will actually be carried out.

Similarly, if a DSS is currently deployed one can sample from $\mathbb{P}(X, A \mid \mathrm{do}(D = 0))$ without taking the DSS offline, by measuring an intended action A without revealing the prediction $\hat{Y}$ to the agent. After having made this measurement, one can reveal the prediction $\hat{Y}$, and the agent can proceed with taking actions.

If A and Y are confounded by a common cause C (so $Y \not\perp\!\!\!\perp D, \Theta \mid X, A$ and $Y \perp\!\!\!\perp D, \Theta \mid X, A, C$) then it is instrumental to also measure this confounding information, i.e. measure $\mathbb{P}(X, A, C| \mathrm{do}(D = d, \Theta = \theta))$ for target domain (d, θ). This is a restrictive, but common assumption in causal inference. In automated decision making, it is not uncommon for a decision algorithm to heuristically combine a prediction $\hat{Y}$ and additional covariates C that were not used for making the prediction (so they are not part of features X), in which case this confounding information might be readily available.

Note that prior to deployment, one cannot test for the required independence due to absence of labels Y from the domain $D = 1$. Instead, one could motivate this independence assumption by causal modelling of the data generating process.

[5] F is a statistical functional if it is a function $F : \mathcal{P}(\mathcal{Y}) \to \mathbb{R}^d$, with $\mathcal{P}(\mathcal{Y})$ the space of probability distributions on $\mathcal{Y}$, and $d \in \mathbb{N}$. For more information, see Shao (2003).

Our main result is that when it is unfeasible to measure labels Y in the target domain, but when we are able to measure variables $\{X, Z\}$ in the target domain, our target quantities can be identified if and only if $\{X, Z\}$ is a domain pivot.

Proposition 11 *Let $(d', \theta') \neq (d, \theta)$ be given with $d' = d = 1 \iff \theta' \neq \theta$. The target quantity $\mathbb{E}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)]$ is identifiable in $\mathcal{M}$ from*

$$\{\mathbb{P}(X, Y, Z \mid \mathrm{do}(D = d', \Theta = \theta')), \mathbb{P}(X, Z \mid \mathrm{do}(D = d, \Theta = \theta))\} \tag{7}$$

if $\mathbb{P}(X, Z \mid \mathrm{do}(D = d, \Theta = \theta)) \ll \mathbb{P}(X, Z \mid \mathrm{do}(D = d', \Theta = \theta'))$,[6] *and if and only if $Y \perp\!\!\!\perp D, \Theta \mid X, Z$, in which case*

$$\mathbb{E}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)] = \int \mathbb{E}[Y \mid X, Z, \mathrm{do}(D = d', \Theta = \theta')]\mathrm{d}\mathbb{P}(Z \mid X, \mathrm{do}(D = d, \Theta = \theta)) \tag{8}$$

$\mathbb{P}(X | \mathrm{do}(D = d, \Theta = \theta))$-a.e.

The absolute continuity ensures that the conditional expectation $\mathbb{E}[Y \mid X, Z, \mathrm{do}(D = d', \Theta = \theta')]$ is $\mathbb{P}(X, Z| \mathrm{do}(D = d, \Theta = \theta))$-a.e. well-defined, and hence that the integral in (8) is well-defined.

Via Lemma 8 and Proposition 11, we have solvability of domain adaptation tasks T1–3 under the assumption that a domain pivot is measured in the source and target domain. For completeness, we provide an overview of these implied identifiability results:

Corollary 12 *In the subset of SCMs of $\mathcal{M}$ that have a domain pivot $\{X, Z\}$ for $\mathbb{E}[Y|X, \mathrm{do}(D, \Theta)]$ and for which $\mathbb{P}(X, Z \mid \mathrm{do}(D = d, \Theta = \theta)) \ll \mathbb{P}(X, Z \mid \mathrm{do}(D = d', \Theta = \theta'))$ for all d, d', θ, θ', we have that*

T1. the deployment effect $\tau(\theta)$ is identifiable from $\{\mathbb{P}(X, Y, Z \mid \mathrm{do}(D = 0)), \mathbb{P}(X, Z \mid \mathrm{do}(D = 1, \Theta = \theta))\}$ via

$$\tau(\theta) = \mathbb{E}[\mathbb{E}[Y \mid X, Z, \mathrm{do}(D = 0)] \mid \mathrm{do}(D = 1, \Theta = \theta)] - \mathbb{E}[Y| \mathrm{do}(D = 0)] \tag{9}$$

and from $\{\mathbb{P}(X, Y, Z \mid \mathrm{do}(D = 1, \Theta = \theta)), \mathbb{P}(X, Z \mid \mathrm{do}(D = 0))\}$ via

$$\tau(\theta) = \mathbb{E}[Y \mid \mathrm{do}(D = 1, \Theta = \theta)] - \mathbb{E}[\mathbb{E}[Y \mid X, Z, \mathrm{do}(D = 1, \Theta = \theta)]| \mathrm{do}(D = 0)]; \tag{10}$$

T2. the retraining effect $\rho(\theta_{t+1}, \theta_t)$ is identifiable from $\{\mathbb{P}(X, Y, Z \mid \mathrm{do}(D = 1, \Theta = \theta_t)), \mathbb{P}(X, Z \mid \mathrm{do}(D = 1, \Theta = \theta_{t+1}))\}$ via

$$\rho(\theta_{t+1}, \theta_t) := \mathbb{E}[\mathbb{E}[Y \mid X, Z, \mathrm{do}(D = 1, \Theta = \theta_t)]| \mathrm{do}(D = 1, \Theta = \theta_{t+1})] - \mathbb{E}[Y \mid \mathrm{do}(D = 1, \Theta = \theta_t)]; \tag{11}$$

T3. the baseline predictor $\mathbb{E}[Y \mid X, \mathrm{do}(D = 0)]$ is identifiable from $\{\mathbb{P}(X, Y, Z \mid \mathrm{do}(D = 1, \Theta = \theta)), \mathbb{P}(X, Z \mid \mathrm{do}(D = 0))\}$ via

$$\mathbb{E}[Y \mid X, \mathrm{do}(D = 0)] = \mathbb{E}[\mathbb{E}[Y \mid X, Z, \mathrm{do}(D = 1, \Theta = \theta)]|X, \mathrm{do}(D = 0)]. \tag{12}$$

We note that the assumption in Proposition 11 of availability of measurements of a domain pivot $\{X, Z\}$ from the target domain is necessary for solving T1–3. Indeed, Lemma 8, Proposition 9 and Proposition 11 together show the necessity of these measurements to solve these estimation tasks, if one is not willing to make further assumptions on the causal model.

[6] For two distributions $\mathbb{P}(X)$ and $\tilde{\mathbb{P}}(X)$, $\mathbb{P}(X) \ll \tilde{\mathbb{P}}(X)$ denotes absolute continuity of $\mathbb{P}$ with respect to $\tilde{\mathbb{P}}$, i.e. $\mathbb{P}(X \in B) > 0 \implies \tilde{\mathbb{P}}(X \in B) > 0$ for all measurable sets B.

3. Estimation

The identification result for the quantity $\mathbb{E}[Y \mid X, \mathrm{do}(D, \Theta)]$ expresses the target quantity as an integral of a conditional expectation; this expression does not indicate how to *estimate* the quantity of interest. When X and Z have finite sample spaces, one can estimate the conditional expectation (the integrand) with maximum likelihood, and compute the integral as a finite sum. However, when there are continuous variables involved, one has to tend to regression methods. In this section we expand on the suitable *repeated regression* procedure, as proposed by Boeken et al. (2023).

To estimate $\mathbb{E}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)]$ we can express equation (8) as

$$\mathbb{E}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)] = \mathbb{E}[\mathbb{E}[Y \mid X, Z] \mid X, \mathrm{do}(D = d, \Theta = \theta)], \tag{13}$$

where we used $Y \perp\!\!\!\perp D, \Theta \mid X, Z$ to remove the conditioning on (D, Θ) in the inner expectation.[7] We formulate an estimation procedure based on this expression by estimating both conditional expectations on the right-hand-side with a regression model. More explicitly, given a sample $(X_i, Y_i, Z_i) \sim \mathbb{P}(X, Y, Z \mid \mathrm{do}(D = d', \Theta = \theta'))$ with indices i in index set $\mathcal{I}_S$ (source) and $(X_i, Z_i) \sim \mathbb{P}(X, Z \mid \mathrm{do}(D = d, \Theta = \theta))$ with indices i in index set $\mathcal{I}_T$ (target), the *repeated regression estimator* $\hat{\mathbb{E}}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)]$ is defined by estimating $\hat{\mathbb{E}}[Y \mid X, Z]$ from $\mathcal{I}_S$, augmenting the target dataset with pseudo-labels $\widetilde{Y}_i := \hat{\mathbb{E}}[Y \mid X = X_i, Z = Z_i]$ for all $i \in \mathcal{I}_T$, and estimating $\hat{\mathbb{E}}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)] := \hat{\mathbb{E}}[\widetilde{Y} \mid X]$ on the augmented target dataset.

For estimation of the expectation $\mathbb{E}[Y| \mathrm{do}(D = d, \Theta = \theta)]$, we can estimate a regression model $\hat{\mathbb{E}}[Y \mid X, Z]$ on the source dataset $\mathcal{I}_S$, and directly compute $\hat{\mathbb{E}}[Y| \mathrm{do}(D = d, \Theta = \theta)] := |\mathcal{I}_T|^{-1} \sum_{i \in \mathcal{I}_T} \hat{\mathbb{E}}[Y \mid X = x_i, Z = z_i]$.

The repeated regression procedure only requires measurements of the variable Z to be available during training and not during deployment. Hence, this estimation procedure falls under the Learning using Privileged Information paradigm Vapnik and Vashist (2009); Vapnik and Izmailov (2015). This is a convenient property, as it might be costly or even impossible to measure the covariates Z at test time. We leave the choice of the regression method up to the practitioner, but we remark that these methods typically impose further assumptions on the sample spaces, exogenous distributions and structural equations.

Using Example 1, we demonstrate how these methods can be used to evaluate the deployment of the DSS.

Example 1 (Application A: Evaluation) *Recall the data generating process $X \sim \mathrm{Unif}[0, 1], \hat{Y} = f(X)$ for some function f, but now with intermediary variables $A = D \cdot \mathbb{1}\{\hat{Y} > 1/2\}$ with $Y \sim Ber(\sigma(X - 1/2) \cdot (1 - A))$. Recalling Figure 1, we will compute τ or ρ between the epochs to see whether deployment of a new model would be the right choice. Since $Y \perp\!\!\!\perp D, \Theta \mid X, A$, we justifiably use $\{X, A\}$ as domain pivot for estimating τ and ρ. Between $t = 1$ and $t = 2$, for a model with parameter value θ_2 to be deployed in epoch 2, we can sample $\mathbb{P}(X, A| \mathrm{do}(D = 1, \Theta = \theta_2))$, and estimate $\tau(\theta_2) \approx -0.47$ using equation (9), with the iterated expectation computed with repeated (polynomial logistic) regression. The lower the value of Y the better, so we decide to deploy this model in epoch $t = 2$, as displayed in Figure 1(b). Between epochs $t = 2$ and $t = 3$, we have retrained a*

[7] Formally, the independence $Y \perp\!\!\!\perp D, \Theta \mid X, Z$ between random variable Y and input variables D, Θ is to be interpreted as transitional conditional independence (Forré, 2021). It implies $\mathbb{E}[Y|X, Z] = \mathbb{E}[Y|X, Z, \mathrm{do}(D = d, \Theta = \theta)]$ $\mathbb{P}(X, Z| \mathrm{do}(D = d, \Theta = \theta))$-a.e., so we may pool data over multiple epochs when estimating $\mathbb{E}[Y|X, Z]$.

model with parameter value θ_3. Before deploying it, we can sample $\mathbb{P}(X, A|\operatorname{do}(D = 1, \Theta = \theta_3))$, and estimate $\rho(\theta_3, \theta_2) \approx 0.47$ using equation (11), based on which we can decide not to deploy it.

Continuing with Example 1, we demonstrate that the repeated regression procedure correctly estimates the baseline predictor $\mathbb{E}[Y \mid X, \operatorname{do}(D = 0)]$ from $\mathbb{P}(X, A \mid \operatorname{do}(D = 0))$ and $\mathbb{P}(X, A, Y \mid \operatorname{do}(D = 1, \Theta))$, and thereby provides a stable estimation procedure for retraining prediction models.

Example 1 (Application B: Bias correction) *As mentioned above, a model with parameter value θ_3 that is naively trained on data from epoch $t = 2$ suffers from performative bias, as is found by estimating $\rho(\theta_3, \theta_2)$. Instead of deploying the naively retrained model, we leverage the domain pivot $\{X, A\}$ to estimate the baseline predictor $\hat{Y}_{bp} = \mathbb{E}[Y|X, \operatorname{do}(D = 0)]$ using repeated (polynomial logistic) regression. This model is displayed in Figure 4.*

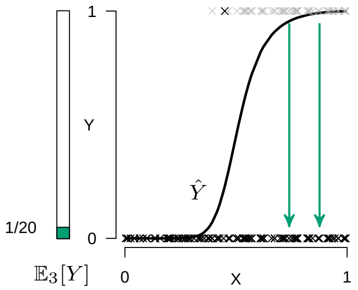


Figure 4: $t = 3, \hat{Y} = \hat{\mathbb{E}}[Y|X, \operatorname{do}(D = 0)]$

4. Sample selection bias and selective labelling

When dealing with real-world data, it is not uncommon that the data suffers from some form of *sample selection bias*: that units are filtered before they are being measured, rendering the sample unrepresentative of the population. Another, related form of bias is when units are selectively labelled, i.e. when the label Y can be missing. This is explicitly prevalent in human-algorithmic decision making, when based on some prediction $\hat{Y}$ the unit can be dismissed, and the outcome Y is not measured; see also Guerdan et al. (2023). These selection mechanisms can be causally modelled by including a binary sample selection indicator S^s and binary labelling indicator S^ℓ in the SCM, where $S^s = 1$ indicates that the unit is included in the dataset, and $S^\ell = 1$ indicates that the label Y is observed.

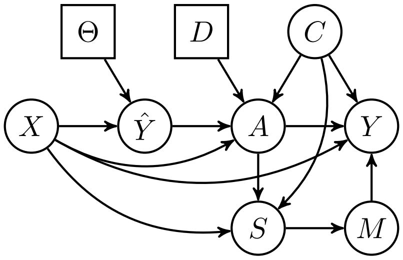


Figure 5: A causal graph with selection variable S.

For ease of notation, we let $S = S^s \wedge S^\ell$, so that in the source domain we measure data from $\mathbb{P}(X, Y, Z \mid S = 1, \mathrm{do}(D, \Theta))$, with $Z \subseteq V$ a set of variables. If one wants to correct for selection bias, the target quantity becomes $\mathbb{E}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)]$, so without the conditioning on $S = 1$. Similar to the treatment in Section 2 it can be shown that this target quantity is not identifiable, but one can extend Proposition 11 to settings where selection bias is in play by considering (D, Θ, S) to be the domain indicator and $\{X, Z\}$ a domain pivot, so with $Y \perp\!\!\!\perp D, \Theta, S \mid X, Z$. For example, in Figure 5 we can let $Z = \{A, C, M\}$, where A can for example be an action, C a confounder, and M a mediator of the selection variable and the outcome. The target quantity $\mathbb{E}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)]$ is then identifiable in $\mathcal{M}$ from $\{\mathbb{P}(X, Y, Z \mid S = 1, \mathrm{do}(D = d', \Theta = \theta')), \mathbb{P}(X, Z \mid \mathrm{do}(D = d, \Theta = \theta))\}$ if also $\mathbb{P}(X, Z \mid \mathrm{do}(D = d, \Theta = \theta)) \ll \mathbb{P}(X, Z \mid S = 1, \mathrm{do}(D = d', \Theta = \theta'))$, in which case

$$\mathbb{E}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)] = \mathbb{E}[\mathbb{E}[Y \mid X, Z, S = 1] \mid X, \mathrm{do}(D = d, \Theta = \theta)] \tag{14}$$

$\mathbb{P}(X \mid \mathrm{do}(D = d, \Theta = \theta))$-a.e. This iterated expectation can be estimated using repeated regression.

A similar identification result has been given in Boeken et al. (2023), but solely for selection bias or missing response. The above identification result shows the ability of repeated regression to correct for multiple forms of domain shift simultaneously, provided that a domain pivot can be measured in the target domain. For more information on selection bias and missing response, we refer to Boeken et al. (2023) and references therein.

5. Discussion

In this work, we modelled the deployment of a decision support system as causal domain shift, introduced evaluation and bias correction as two applications of this causal model, and have shown how certain estimands in these applications can be only be estimated under the availability of measurements of a *domain pivot* in the target domain. We have demonstrated how *repeated regression* is a suitable estimation procedure for evaluation and bias correction, even if the measured labels are subject to selection bias and/or selective labelling.

Sensitivity analysis with respect to the conditional independence assumption, such as the estimation of bounds of the quantity of interest when this independence does not hold, might be a promising direction for future work. Constructing doubly robust and efficient estimators for the deployment effect, retraining effect, and baseline predictor, e.g. using influence functions (Dudík et al., 2014; Athey et al., 2019), would also be of interest.

An important assumption for the relevance of our identifiability results is that labels are never observed in the target domain of the domain adaptation problems. Estimating the deployment effect, retraining effect, and baseline predictor can be done directly on available data if labels are measured in the target domain, e.g. through A/B testing of the deployment of the DSS. However, in high stakes environments it is often neither desired nor ethical to carry out such experiments. If these labels are not measured in the target domain, our proposed methods heavily depend on the availability of measurements of a domain pivot in the target domain, without actually deploying the DSS. This can be a restrictive assumption in practice. If such data is available, our evaluation metrics rely on an estimated model; a practice that requires caution.

Nevertheless, we hope that the proposed causal model, evaluation metrics and the concept of performative bias will be useful tools for responsible applications of AI systems in high-stakes settings.

Acknowledgments

We thank the anonymous reviewers, Stephan Bongers and Sourbh Bhadane for their feedback, which considerably helped improving the manuscript. This work is supported by Booking.com.

References

Susan Athey, Raj Chetty, Guido Imbens, and Hyunseung Kang. The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely. Technical Report w26463, National Bureau of Economic Research, Cambridge, MA, November 2019. URL http://www.nber.org/papers/w26463.pdf.

John Langshaw Austin. *How to do things with words*. Harvard University Press, 1962.

Philip Boeken, Noud de Kroon, Mathijs de Jong, Joris M. Mooij, and Onno Zoeter. Correcting for selection bias and missing response in regression using privileged information. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 195–205. PMLR, July 2023. URL https://proceedings.mlr.press/v216/boeken23a.html.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730, Sydney NSW Australia, August 2015. ACM. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2788613. URL https://dl.acm.org/doi/10.1145/2783258.2788613.

Yatong Chen, Zeyu Tang, Kun Zhang, and Yang Liu. Model Transferability with Responsive Decision Subjects. In *Proceedings of the 40th International Conference on Machine Learning*, pages 4921–4952. PMLR, July 2023. URL https://proceedings.mlr.press/v202/chen23y.html.

Juan D. Correa and Elias Bareinboim. From Statistical Transportability to Estimating the Effect of Stochastic Interventions. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 1661–1667, Macao, China, August 2019. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-4-1. doi: 10.24963/ijcai.2019/230. URL https://www.ijcai.org/proceedings/2019/230.

Amanda Coston, Edward Kennedy, and Alexandra Chouldechova. Counterfactual Predictions under Runtime Confounding. In *Advances in Neural Information Processing Systems*, volume 33, pages 4150–4162. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/2b64c2f19d868305aa8bbc2d72902cc5-Abstract.html.

Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 690–704, February

2023. doi: 10.1109/SaTML54575.2023.00050. URL https://ieeexplore.ieee.org/abstract/document/10136159.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly Robust Policy Evaluation and Optimization. *Statistical Science*, 29(4), November 2014. ISSN 0883-4237. doi: 10.1214/14-STS500. URL http://arxiv.org/abs/1503.02834.

Patrick Forré. Transitional Conditional Independence, August 2021. URL http://arxiv.org/abs/2104.11547.

Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein. Ground(less) Truth: A Causal Framework for Proxy Labels in Human-Algorithm Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 688–704, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594036. URL https://dl.acm.org/doi/10.1145/3593013.3594036.

Michael P. Kim and Juan C. Perdomo. Making Decisions under Outcome Performativity, January 2023. URL http://arxiv.org/abs/2210.01745.

Sanghack Lee, Juan Correa, and Elias Bareinboim. General transportability–synthesizing observations and experiments from heterogeneous domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10210–10217, 2020.

Donald A MacKenzie, Fabian Muniesa, Lucia Siu, et al. *Do economists make markets?: on the performativity of economics*. Princeton University Press, 2007.

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/39e98420b5e98bfbdc8a619bef7b8f61-Abstract.html.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic Optimization for Performative Prediction. In *Advances in Neural Information Processing Systems*, volume 33, pages 4929–4939. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/33e75ff09dd601bbe69f351039152189-Abstract.html.

Celestine Mendler-Dünner, Frances Ding, and Yixin Wang. Anticipating Performativity by Predicting from Predictions. *Advances in Neural Information Processing Systems*, 35:31171–31185, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ca09b375e8e2b2c789698c079a9fc51c-Abstract-Conference.html.

John P. Miller, Juan C. Perdomo, and Tijana Zrnic. Outside the Echo Chamber: Optimizing the Performative Risk. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7710–7720. PMLR, July 2021. URL https://proceedings.mlr.press/v139/miller21a.html.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl and Elias Bareinboim. Transportability of Causal and Statistical Relations: A Formal Approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25(1):247–254, August 2011. ISSN 2374-3468. doi: 10.1609/aaai.v25i1.7861. URL https://ojs.aaai.org/index.php/AAAI/article/view/7861.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative Prediction. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7599–7609. PMLR, November 2020. URL https://proceedings.mlr.press/v119/perdomo20a.html.

José Pombal, Pedro Saleiro, Mário A. T. Figueiredo, and Pedro Bizarro. Prisoners of Their Own Devices: How Models Induce Data Bias in Performative Prediction, June 2022. URL http://arxiv.org/abs/2206.13183.

Jun Shao. *Mathematical Statistics*. Springer Texts in Statistics. Springer, New York, 2nd ed edition, 2003. ISBN 978-0-387-95382-3.

Mats J. Stensrud, Julien Laurendeau, and Aaron L. Sarvet. Optimal regimes for algorithm-assisted human decision-making, April 2023. URL http://arxiv.org/abs/2203.03020.

Vladimir Vapnik and Rauf Izmailov. Learning Using Privileged Information: Similarity Control and Knowledge Transfer. *Journal of Machine Learning Research*, 16(61):2023–2049, 2015. ISSN 1533-7928. URL http://jmlr.org/papers/v16/vapnik15b.html.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, July 2009. ISSN 0893-6080. doi: 10.1016/j.neunet.2009.06.042. URL https://www.sciencedirect.com/science/article/pii/S0893608009001130.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and Adaptive Off-policy Evaluation in Contextual Bandits. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3589–3597. PMLR, July 2017. URL https://proceedings.mlr.press/v70/wang17a.html.

Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E. Ho, and James Zou. How medical AI devices are evaluated: Limitations and recommendations from an analysis of FDA approvals. *Nature Medicine*, 27(4):582–584, April 2021. ISSN 1546-170X. doi: 10.1038/s41591-021-01312-x. URL https://www.nature.com/articles/s41591-021-01312-x.

Tom Yan, Shantanu Gupta, and Zachary Lipton. Discovering Optimal Scoring Mechanisms in Causal Strategic Prediction, February 2023. URL http://arxiv.org/abs/2302.06804.

Appendix A. Relation to existing literature

A.1. Performative prediction

Translated to our notation, Perdomo et al. (2020) introduced the performative risk:

$$\mathbb{E}[\ell(\hat{Y}, Y) |\operatorname{do}(D = 1, \Theta = \theta)]. \tag{15}$$

where ℓ is a loss function. They define a parameter θ_t to be performatively stable if it constant under retraining, so if we get

$$\theta_t \approx \arg\min_{\theta} \mathbb{E}[\ell(\hat{Y}, Y) |\operatorname{do}(D = 1, \Theta = \theta_t)]. \tag{16}$$

They don't consider the case where $D = 0$, so the deployment effect $\tau(\theta)$ can for example not be defined using the existing framework.

The baseline predictor $\hat{Y} = \mathbb{E}[Y|X, \operatorname{do}(D = 0)]$ does not minimize performative risk, but in the setting of Proposition 6, it is a performatively stable predictor: if we parametrise $\hat{Y}_{\theta_t} = \hat{\mathbb{E}}_{\theta_t}[Y|X, \operatorname{do}(D = 0)]$, then estimating $\hat{Y}_{\theta_{t+1}} = \hat{\mathbb{E}}_{\theta_{t+1}}[Y|X, \operatorname{do}(D = 0)]$ from $\mathbb{P}(X, Y, Z| \operatorname{do}(D = 1, \Theta = \theta_t))$ and $\mathbb{P}(X, Z| \operatorname{do}(D = 0))$ will yield $\theta_t \approx \theta_{t+1}$. A performatively stable parameter $\theta_t \approx \theta_{t+1}$ will have as retraining effect $\rho(\theta_t, \theta_{t+1}) \approx 0$.

A.2. Off-policy evaluation

In contextual bandits, one considers a context X, an action $A \sim \mathbb{P}(A|X, \operatorname{do}(\Theta = \theta))$ (where $\mathbb{P}(A|X, \operatorname{do}(\Theta))$ is referred to as a *policy* with parameters Θ), and a reward $Y \sim \mathbb{P}(Y|X, A)$. This gives rise to a joint distribution $\mathbb{P}(X, A, Y| \operatorname{do}(\Theta))$. When one has measured data from a policy with parameters θ, the problem of *off-policy evaluation* is that of estimating for a new set of parameters $\theta' \neq \theta$ the reward $\mathbb{E}[Y| \operatorname{do}(\Theta = \theta')]$ from $\mathbb{P}(X, A, Y| \operatorname{do}(\Theta = \theta))$.

If the relation between X, A and Y is such as described above, Proposition 11 combined with the repeated regression procedure says that we can estimate

$$\mathbb{E}[Y| \operatorname{do}(\Theta = \theta')] = \mathbb{E}[\mathbb{E}[Y|X, A]| \operatorname{do}(\Theta = \theta')], \tag{17}$$

which is known as the *direct method* in contextual bandit literature Dudík et al. (2014). Note that our results from Section 2 consider a rather intricate policy, namely one that factorizes according to $\mathbb{P}(A|X, C, \operatorname{do}(D = 1, \Theta = \theta)) = \mathbb{P}(A|X, C, \hat{Y}, \operatorname{do}(D = 1))\mathbb{P}(\hat{Y}|X, \operatorname{do}(\Theta = \theta))$. Note that typically, for given parameters θ, we don't *know* the policy $\mathbb{P}(A|X, C, \operatorname{do}(D = 1, \Theta = \theta))$ (contrary to when one considers a setting of automated decision making) but we can merely sample from it, as is explained in Section 2.4.

A.3. Surrogate indices

Athey et al. (2019) consider the estimation of a causal effect with a similar technique as repeated regression. For estimating a causal effect

$$\mathbb{E}[Y| \operatorname{do}(D = 1)] - \mathbb{E}[Y| \operatorname{do}(D = 0)], \tag{18}$$

they consider the case where one has two samples: an observational sample with measurements of covariates X, target variable Y and so-called *surrogates* Z (so not of D), and an experimental sample with measurements of X, D and Z (so not of Y). For a set of variables Z to be surrogates, they require the independence $Y \perp\!\!\!\perp D \,|\, X, Z$[8], and that $\{X, Z\}$ is a valid *adjustment set* for estimating the causal effect of D on Y, i.e. $\mathbb{E}[Y \,|\, \mathrm{do}(D)] = \mathbb{E}[\mathbb{E}[Y|X, Z]|D]$. Their identification strategy is built on the equation

$$\mathbb{E}[Y \,|\, \mathrm{do}(D = 1)] - \mathbb{E}[Y \,|\, \mathrm{do}(D = 0)] = \mathbb{E}[\mathbb{E}[Y|X, Z]|D = 1] - \mathbb{E}[\mathbb{E}[Y|X, Z]|D = 0]. \tag{19}$$

Note that we consider a different setup. Instead of having ‘observational’ and ‘experimental’ samples, alternating the measurements of treatment D or outcome Y, we consider a setting where for one value of the treatment ($D = 0$, say) we observe Y, and for the other value of the treatment we don’t observe Y. An overview of these different assumptions is given in Table 2. Our identification result is similar to that of Athey et al. (2019), but can be interpreted as intervention extrapolation, instead of causal effect estimation using surrogate outcomes.

	Sample	X	$D=0$	$D=1$	Z	Y
Athey et al. (2019)	Observational	✓	×	×	✓	✓
	Experiment	✓	✓	✓	✓	×
This work	Source ($D=0$)	✓	✓	×	✓	✓
	Target ($D=1$)	✓	×	✓	✓	×

Table 2: A comparison of the setting in Athey et al. (2019) and this work.

Appendix B. Proofs

Proof [Proposition 6] Define $G := \{x : \mathbb{P}(Y = 1 \,|\, X = x, A = 0) > \varepsilon(x)\}$. We have the unique optimal policy $A^*(X) := \mathbb{1}_G(X)$ for $\min_{a \in \{0,1\}^{\mathcal{X}}} \mathbb{P}(Y = 1 \,|\, X, A = a)$, and thus we have the set of minimizers

$$\begin{aligned} H &:= \operatorname*{arg\,min}_{\hat{y} \in [0,1]^{\mathcal{X}}} \mathbb{P}(Y = 1 \,|\, X = x, \mathrm{do}(\hat{Y} = \hat{y}(x))) \\ &= \operatorname*{arg\,min}_{\hat{y} \in [0,1]^{\mathcal{X}}} \mathbb{P}(Y = 1 \,|\, X = x, A = \mathbb{1}\{\hat{y}(x) > \varepsilon\}) \\ &= \{\hat{y} \in [0,1]^{\mathcal{X}} \ : A^*(x) = \mathbb{1}\{\hat{y}(x) > \varepsilon(x)\} \forall x \in \mathcal{X}\} \\ &= \{\hat{y} \in [0,1]^{\mathcal{X}} \ : \hat{y}(x) \geq \mathbb{P}(Y = 1 \,|\, X = x, A = 0)\}. \end{aligned}$$

Clearly we have

$$\begin{aligned} \hat{Y}^* &:= \operatorname*{arg\,min}_{\hat{y} \in H} \mathbb{P}(A = 1 \,|\, X = x, \mathrm{do}(\hat{Y} = \hat{y}(x))) \\ &= \operatorname*{arg\,min}_{\hat{y} \in H} \mathbb{1}\{\hat{y}(x) \geq \varepsilon(x)\} \\ &= \mathbb{P}(Y = 1 \,|\, X = x, A = 0), \end{aligned}$$

[8] This is the same conditional independence that we require for the domain pivot, but we are reluctant to call a domain pivot a surrogate, as we don’t restrict the domain shift to be the value of an intervention but also other types of domain shift, like selection bias.

and since $Y \perp\!\!\!\perp D \mid X, A$ and $D = 0 \implies A = 0$ we further get

$$\mathbb{P}(Y = 1 \mid X = x, A = 0) = \mathbb{P}(Y = 1 \mid X = x, A = 0, D = 0) = \mathbb{P}(Y = 1 \mid X = x, D = 0),$$

and thus $\hat{Y}^* = \mathbb{E}[Y \mid X, D = 0]$. ■

Proof [Lemma 8] If $\mathbb{E}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)]$ is identifiable from $\mathbb{P}(X, Y \mid \mathrm{do}(D = d', \Theta = \theta'))$, then since $X \perp\!\!\!\perp D, \Theta$ we have $\mathbb{P}(X \mid \mathrm{do}(D = d', \Theta = \theta')) = \mathbb{P}(X \mid \mathrm{do}(D = d, \Theta = \theta))$, so $\mathbb{E}[Y \mid \mathrm{do}(D = d, \Theta = \theta)] = \mathbb{E}[\mathbb{E}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)] \mid \mathrm{do}(D = d', \Theta = \theta')]$, so $\mathbb{E}[Y \mid \mathrm{do}(D = d, \Theta = \theta)]$ is identifiable as well.

If $\mathbb{E}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)]$ is not identifiable from $\mathbb{P}(X, Y \mid \mathrm{do}(D = d', \Theta = \theta'))$, then there exist M_1 and M_2 such that $\mathbb{P}_{M_1}(X, Y \mid \mathrm{do}(D = d', \Theta = \theta')) = \mathbb{P}_{M_2}(X, Y \mid \mathrm{do}(D = d', \Theta = \theta'))$ and $\mathbb{E}_{M_1}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)] \neq \mathbb{E}_{M_2}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)]$. Let x' be such that $\mathbb{E}_{M_1}[Y \mid X = x', \mathrm{do}(D = d, \Theta = \theta)] \neq \mathbb{E}_{M_2}[Y \mid X = x', \mathrm{do}(D = d, \Theta = \theta)]$ and let $\tilde{M}_1, \tilde{M}_2$ be equal to the SCMs M_1, M_2, except for the structural equation for X, which is set to $X = x'$ in both $\tilde{M}_1$ and $\tilde{M}_2$. Then we still have $\mathbb{P}_{\tilde{M}_1}(X, Y \mid \mathrm{do}(D = d', \Theta = \theta')) = \mathbb{P}_{\tilde{M}_2}(X, Y \mid \mathrm{do}(D = d', \Theta = \theta'))$, and $\mathbb{E}_{\tilde{M}_1}[Y \mid \mathrm{do}(D = d, \Theta = \theta)] = \mathbb{E}_{\tilde{M}_1}[Y \mid X = x', \mathrm{do}(D = d, \Theta = \theta)] \neq \mathbb{E}_{\tilde{M}_2}[Y \mid X = x', \mathrm{do}(D = d, \Theta = \theta)] = \mathbb{E}_{\tilde{M}_2}[Y \mid \mathrm{do}(D = d, \Theta = \theta)]$, so $\mathbb{E}[Y \mid \mathrm{do}(D = d, \Theta = \theta)]$ is not identifiable from $\mathbb{P}(X, Y \mid \mathrm{do}(D = d', \Theta = \theta'))$. ■

Proof [Proposition 9] Let $\theta, \theta' \in \mathbb{R}, d, d' \in \{0, 1\}$ be given, and consider for $i = 1, 2$ the SCM M_i given by $X \sim \mathcal{N}(0, 1), \hat{Y} = \Theta \cdot X + i \cdot \mathbb{1}\{\Theta \neq \theta'\}, Y = X + \mathbb{1}\{D = 1\} \cdot \hat{Y} + i \cdot \mathbb{1}\{D \neq d'\}$. One can readily verify that $\mathbb{P}_{M_1}(X, Y \mid \mathrm{do}(D = d', \Theta = \theta')) = \mathbb{P}_{M_2}(X, Y \mid \mathrm{do}(D = d', \Theta = \theta'))$, but that $\mathbb{E}_{M_1}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)] = 1 + X \neq 2 + X = \mathbb{E}_{M_2}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)]$ if $d = 0$, that $\mathbb{E}_{M_1}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)] = (\theta + 1)X + 1 \neq (\theta + 1)X + 2 = \mathbb{E}_{M_2}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)]$ if $d = 1$ and either $d = d'$ or $\theta = \theta'$, and that $\mathbb{E}_{M_1}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)] = (\theta + 1)X + 2 \neq (\theta + 1)X + 4 = \mathbb{E}_{M_2}[Y \mid X, \mathrm{do}(D = d, \Theta = \theta)]$ if $d = 1$ and both $d \neq d'$ and $\theta \neq \theta'$. ■

Proof [Proposition 11] That $Y \perp\!\!\!\perp D, \Theta \mid X, Z$ implies identifiability (under the required positivity assumption) is immediate from equation (8).

If $Y \not\perp\!\!\!\perp D \mid X, Z$, we note that for all $Z' \in Z$, we cannot have $X \to Z' \to \hat{Y}$ as this violates Assumption 1. If there is an edge $\hat{Y} \to Y$, we can augment the constructed M_1, M_2 from the proof of Proposition 9 where we let all $Z' \in Z$ be independent variables having the same distribution in both models, which proves non-identifiability in that setting. The last case to check is where every directed path from $\hat{Y}$ to Y contains at least one element from Z. Since we have $Y \not\perp^{d}_{G'} D \mid X, Z$, there is at least one such a path $\pi = \hat{Y} \to Z_1 \to ... \to Z_n \to Y$ for some $n \in \mathbb{N}$ with $Z_1, ..., Z_n \in Z$ and $Z_1 \leftrightarrow Y$ in G'. We define M_1 by letting $X \sim \mathrm{Ber}(1/2), \hat{Y} = X, E_{Z_1} \sim \mathrm{Ber}(1/2), E_{Z_1 Y} \sim \mathrm{Ber}(1/2), Z_1 = \mathbb{1}_{\{D \neq d\}} \cdot \mathrm{XOR}(D, E_{Z_1 Y}) + \mathbb{1}_{\{D = d\}} \cdot \mathrm{XOR}(D, E_{Z_1}), Z_{i+1} = Z_i$ for $i = 1, ..., n-1$, and $Y = \mathrm{XOR}(Z_n, E_{Z_1 Y})$. We let all other variables in Z be independent. Define M_2 to be equal to M_1, with the only difference that $Z_1 = \mathrm{XOR}(D, E_{Z_1 Y})$. Then indeed $\mathbb{P}_1(X, Y, Z \mid \mathrm{do}(D = d', \Theta)) = \mathbb{P}_2(X, Y, Z \mid \mathrm{do}(D = d', \Theta))$, $\mathbb{P}_1(X, Z \mid \mathrm{do}(D = d, \Theta)) = \mathbb{P}_2(X, Z \mid \mathrm{do}(D = d, \Theta))$ and $\mathbb{E}_1[Y \mid X, \mathrm{do}(D = d, \Theta)] = d \neq 1/2 = \mathbb{E}_2[Y \mid X, \mathrm{do}(D = d, \Theta)]$, proving non-identifiability.
■