Lechner
Swiss Journal of Economics and Statistics (2023) 159:8
https://doi.org/10.1186/s41937-023-00113-y

Swiss Journal of
Economics and Statistics

CONFERENCE KEY NOTE

Open Access

Causal Machine Learning and its use for public policy

Michael Lechner[1]*

Abstract

In recent years, microeconometrics experienced the 'credibility revolution', culminating in the 2021 Nobel prices for David Card, Josh Angrist, and Guido Imbens. This 'revolution' in how to do empirical work led to more reliable empirical knowledge of the causal effects of certain public policies. In parallel, computer science, and to some extent also statistics, developed powerful (so-called Machine Learning) algorithms that are very successful in prediction tasks. The new literature on *Causal Machine Learning* unites these developments by using algorithms originating in Machine Learning for improved causal analysis. In this non-technical overview, I review some of these approaches. Subsequently, I use an empirical example from the field of active labour market programme evaluation to showcase how Causal Machine Learning can be applied to improve the usefulness of such studies. I conclude with some considerations about shortcomings and possible future developments of these methods as well as wider implications for teaching and empirical studies.

Keywords Causal analysis, Machine Learning, Econometric evaluation

1 Introduction

Arguably, in 1983 (applied) econometrics reached its low point in what Ed Leamer called the credibility crisis in his famous article in the American Economic Review (Leamer, 1983). The quote from his paper 'hardly anyone takes anyone else's data analyses seriously' (p. 37) described rather accurately the state of (applied) econometrics in those times. Obviously, once empirical results lose their credibility by the standards of our own discipline, they are worthless as a tool for evaluating and improving public policy.

After this near death experience in the 1980s, the field changed dramatically. Researchers became much more aware of the limits of econometric empirical analyses and of their dependence on crucial (identifying) assumptions. Since such assumptions can dramatically impact the conclusions to be drawn from the empirical studies, they must be credible. Credibility means that these assumptions are reasonable approximations of the empirical reality analysed. Angrist and Pischke (2010) coined the term 'credibility revolution' for this period. Novel methods and a better understanding of old methods, like Instrumental Variables (Heckman, 1997; Imbens & Angrist, 1994) and regression, as well as a focus on parameters that could be 'credibly identified' changed the way empirical analyses were performed. Although this change was certainly most pronounced in microeconometrics, it proliferated and still proliferates across all fields of applied econometrics, although at different speeds.

This progress was honoured by the Nobel committee in Stockholm by giving the 'The Sveriges Riskbank Prize in Economic Sciences in Memory of Alfred Nobel' to Jim Heckman in 2000 and to Joshua Angrist, David Card, and Guido Imbens in 2021.[1] It is now best practice in empirical work to clearly state the assumptions needed for

*Correspondence:
Michael Lechner
Michael.Lechner@unisg.ch
[1] Swiss Institute for Empirical Economic Research (SEW), University of St. Gallen, Varnbüelstrasse 14, 9000 St. Gallen, Switzerland

[1] The Nobel lectures of Angrist (2022) and Imbens (2022) contain nice recaps of these developments.

estimating (causal) effects, discuss their credibility, and, finally, acknowledge that these causal effects may be heterogeneous across the units analysed.

Other fields also saw massive improvements in conducting credible empirical analyses. For example, in statistics Donald Rubin (Imbens & Rubin, 2015; Rubin, 1974) formalized the potential outcome approach, in epidemiology Robins (1986) proposed a dynamic approach based on potential outcomes, and in computer science Judea Pearl (Pearl, 2000; Pearl & Mackenzie, 2018) introduced a graph-based approach to causality, i.e. the directed acyclic graphs (DAG).

In summary, around 2010 we had new best practices on how to estimate *average* effects of what was now called 'treatments' (policies, actions, decisions, etc.). By using the term 'quasi-experiments' or 'natural experiments' for such methods, we rather explicitly claim that such observational studies are almost as credible as if the data came from an experiment. Major shortcomings were still the difficulty of estimating effect heterogeneity at a fine-grained level as well as the frequent use of parametric statistical models that impose additional restrictions like specific functional form assumptions that cannot be justified other than by computational (and intellectual) convenience. Nevertheless, the value of empirical economic studies for public policy had increased substantially.[2]

In addition to improvements in the methodology and application of econometrics mentioned before, we saw major technological advances that also affected econometrics. Computing power increased and is still increasing exponentially at constant or even decreasing costs.[3] At the same time, costs of data storage fell, and data collection became much easier and socially more acceptable. Although the latter trend was mainly driven by private companies (just think about the data our smartphones are continuously collecting and transferring to various tech companies), the digitalization of the state also increased the volume of the data collected by public offices and eased the technical access and use for research and policy consulting. Simultaneously, computer scientists developed new and powerful, but computationally demanding, algorithms that turned out to be very successful at prediction tasks. These *Machine Learning* algorithms tend to be highly nonlinear, very flexible, and easily beat common econometric algorithms at most prediction tasks.

Causal Machine Learning unites these trends: it adapts Machine Learning methods to answer well identified causal questions using large and informative data (e.g. Athey, 2017; Athey & Imbens, 2019). Credible answers to causal questions are of course not only of interest to economists, but also to researchers in other fields as well as to the private sector. Therefore, it is not surprising that methods development and applications appear almost simultaneously in many research fields, such as computer science, econometrics, epidemiology, marketing, and statistics, as well as in the private sector. In the remainder of this non-technical paper, I will describe the main methods of Causal Machine Learning that became popular in econometrics as well as showcasing their use with a particular empirical example.

In the next section, I compare a special type of Machine Learning, namely Supervised Statistical Learning, to classical econometrics. In Sect. 3, I zoom in on a special case, namely a static binary treatment model identified by so-called selection-on-observables assumptions. In this context, I discuss some important estimation approaches for various aggregated and disaggregated causal effects. Subsequently, in Sect. 4, an evaluation study of training programmes for unemployed in Flanders is used to show some of the potential of these methods in practice. This application is followed in Sect. 5 by more general considerations about the usefulness and limits of Causal Machine Learning and how its usefulness depends on various features of the study and the chosen research design, i.e. the specific set of identifying assumptions entertained by the researcher. The last section concludes and points to some of the many issues that are still unresolved and can hopefully be resolved in future research.

Many topics are omitted from this brief overview. For example, I will not discuss the use of Causal Machine Learning (CML henceforth) in the private sector, although many firms, particular in the tech industry, currently build up substantial capacity for development and application of these methods (e.g. Athey & Luca, 2019). Furthermore, I will discuss a generic but simple causal set-up, namely the static binary treatment models, thus excluding for example the important field of continuous treatments (e.g. Klosin, 2021). Furthermore, I will ignore more complicated causal designs and mechanism, such as mediation (e.g. Farbmacher et al., 2022), moderation (e.g. Bansak, 2021), networks (e.g. Graham, 2020), dynamic sequences of treatments (Bodory et al., 2022b; Lewis & Syrgkanis, 2020), as well as dynamic learning (bandits, reinforcement learning, e.g. Kasy & Sautmann, 2021; Kock et al., 2022). Furthermore, I will not discuss the *discovery* of causal structures (e.g. Soleymani et al., 2022). I will also not discuss the use of Machine Learning either to generate variables (e.g. from text, pictures,

[2] Admittedly, despite these encouraging trends, there is still substantial room for improvement as the current debate about a replication crises and p-hacking exemplifies (e.g. Brodeur et al., 2020).

[3] Moore's law stating that the number of transistors on a microchip roughly doubles every two years still holds at least approximately (e.g. https://www.britannica.com/technology/Moores-law).

or natural language) that can subsequently be used in causal analysis or for prediction purposes, or both. Some of these topics are already discussed in the recent excellent surveys by Athey and Imbens (2019) and Mullainathan and Spiess (2017), among others.[4]

2 Machine Learning and classical econometrics

Machine Learning is a vast and not very well specified field that had its origins in computer science. The relevant subfield for econometric predictions (and the basis of Machine Learning empowered causal analysis) is *Statistical Learning*. For example, the classical textbook by Hastie et al. (2009)[5] discusses the major methods in this field. The main idea of statistical learning methods is to use flexible specifications to find structure in the data. In *Unsupervised Statistical Learning*, all variables stand on the same footing for that task. Thus, the structure has to be found from all variables simultaneously by exploiting their associations in some way, for example as in cluster analysis.[6] In *Supervised Statistical Learning* (SSL), we know more in the sense that we have a variable we want to predict (outcome variable, y) and other variables that will be used for this prediction (covariates, x).[7] Therefore, (implicitly) the task is to find structures in the covariates space that lead to good predictions of the outcome variable. Within SSL, the specific methods also differ related to what a 'prediction' is: while for continuous outcomes, a conditional expectation is a natural target, for discrete outcomes one may want to focus either on the conditional probability of a specific event happening (*Regression*), or explicitly predict specific values of the discrete outcome variable (*Classification*). CML methods have their origins in both types of SSL.

For regression SSL, the goal is to find a function $f(x)$ such that it approximates the true conditional expectation of y given x well. As in nonparametric econometrics, there are two ways to approach this problem. The first approach consists in globally approximating $f(x)$ with a flexible function, like some polynomial. For example, series estimation in nonparametric econometrics, and ordinary least squares (OLS) and logit in parametric econometrics belong to this class of estimators. The alternative is a local approach: for each specific value of x of interest, say $\underline{x}$, take the mean of y for observations that have values close to $\underline{x}$ and use these means as estimator of $f(x)$. In nonparametric econometrics, kernel regression belongs to this class of estimators.

What are examples of important SSL methods? Let us start with an extension of the most generic econometric method, namely OLS. In this extension, the linear specification of ($f(x)=x\beta$) is kept, but the objective function of OLS is amended. Instead of minimizing the average squared difference of the realized and predicted values of y (i.e. the mean squared error) only, a penalty term is added. This penalty term increases with the absolute magnitude of the coefficients. Such methods are called shrinkage methods as they shrink coefficients relative to OLS and were known before SSL existed in its current form. Their names and properties depend on the type of penalty. For example, *LASSO* (Least Absolute Shrinkage and Selection Operator; Tibshirani, 1996) penalizes the absolute values of the coefficients. If the squared values instead of the absolute values enter the penalty, then this is *Ridge Regression* (Hoerl & Kennard, 1970). If both penalties are combined, we obtain *Elastic Net* (Zou & Hastie, 2005). These estimators share some properties: (1) from a computational point of view, the dimension of x can be larger than the sample size; (2) they are very likely to do better than OLS in predicting y in an out-of-sample mean squared error sense (if the penalty is well chosen); and (3) the estimates of the coefficients of $x\beta$ are usually biased and inconsistent. LASSO has the additional property that it may set some of the estimated coefficients explicitly to zero. Thus, in this sense LASSO may also perform variable selection. How good or bad these methods perform in specific situations depends on properties of the data generating process (DGP). However, a more detailed discussion is beyond the scope of this introductory paper.[8]

Neural networks for regressions are based on connected systems of such shrinkage methods. Technically speaking, they have not much to do with how the brain works but approximate $f(x)$ very flexibly, using a huge number of parameters and nonlinear functions. Subsequently, they regularize heavily to obtain good predictions. Currently, deep neural networks (*deep* implies a network architecture that leads to a very high degree of flexibility) are among the stars of the prediction scene in the sense that they may be able to predict (or classify) y very precisely.

As already mentioned, the local approach to prediction and classification is based on using x to find neighbourhoods in which observations have similar values of

[4] Recently, several books and survey articles appeared that discuss many aspects of Causal Machine Learning, like, for example, Chernozhukov et al. (2023), Huber (2023), Kreif and DiazOrdaz (2019), Lieli et al. (2022), and Shah et al. (2021).

[5] A less technical alternative is the textbook by James et al. (2013).

[6] In other words, all variables are on the same footing and the classical (in econometrics) y–x-distinction does not exist.

[7] In many cases, y will we one dimensional and x will be multidimensional.

[8] An important role plays the so-called sparsity of the true DGP. There are different versions of sparsity that all imply that a low number of variables (out of the possibly very large number of variables available) is sufficient to predict y well.

y, and use the mean of *y* in this region as predictor for the *y*'s of observations with values of *x* that belong to this neighbourhood as well. The key question to answer with this type of approach is on how to form such neighbourhoods such that the resulting estimators have good predictive properties as well as remain computationally tractable. A Classification And Regression Tree (CART, Breiman et al., 1984) fulfils these criteria. The main idea of a CART is to create a sequence of binary splits until the remaining stratum satisfies some condition, like a minimum size of the leaf (leaf = stratum). The splits are created by considering all possible binary splits that can be created by each variable separately. Of all possible binary splits, the one is chosen that minimizes some criteria, like the mean square error reduction resulting from the particular split. The combination of all chosen splits forms the 'tree'. There is bias-variance trade-off which has to be addressed when building such trees: the smaller the single leaf, the more homogenous the observations inside the leaf will be, but the larger the variance of the sample mean of *y* coming from this leaf.

Due to the sequential nature on how they are formed, trees may be unstable. A small change in the data may lead to a different first split and thus a different tree. Furthermore, predictions from trees are non-smooth functions of *x*. Random forests (RF; Breiman, 2001) overcome these two problems and improve substantially on the prediction power. The main idea is to average predictions from many different trees. To obtain different regression trees, the different trees are built on random samples from the original data (bootstrapped or subsampled). Furthermore, for each single splitting decision, a random sample subset of the splitting variables is considered only. Each tree used for a RF is 'deep' (i.e. small), implying that the prediction from the leaf has very low bias. The variance is reduced by averaging over the decorrelated trees. Built in this way, *random forests* turned out to be a very powerful, and yet simple, prediction methods.

What are the main differences between econometrics (EM) and SSL? It begins with the object of interest: While EM is typically interested in estimating low-dimensional parameters (e.g. causal effects, demand elasticities, regression coefficients that have some deeper economic meaning), SSL is optimized towards getting a good prediction of the outcome variable, *y*. Why does this difference matter? It matters mainly because SSL can check how good the estimator performs by comparing the predictions of *y* with the realizations of *y*. There is no chance to do this with a parameter that is unknown. At best, we may have a prior about plausible ranges, but we never know the true value. Therefore, statistical theory is much more important for EM estimators. It acts as a main tool to judge the quality of an estimator in a specific application, for example, by considering its location and uncertainty. From this perspective, it may not be surprising that for some very popular SSL methods that are known 'to work well' for many DGPs, deriving the (asymptotic) statistical properties turned out to be difficult, if not (so far) impossible. This need for inference has another implication on whether an estimator is considered as being acceptable for empirical analysis by econometricians: With increasing sample size (asymptotically), the squared bias of such an estimator must vanish faster than the variance, otherwise the confidence intervals will not be centred around the truth, and inference will be misleading. Such restrictions play no role in SSL. Thus, an MSE minimal estimator may be perfectly fine for SSL, while it may not be attractive for EM. This loss of predictive power is the implicit price to pay for obtaining inference.

While predicting-a-parameter vs. predicting-some-observable-quantity is a major difference of EM and SSL, there are other differences as well. SSL procedures tend to be considerably more flexible than EM models. They may allow for many nonlinearities, lots of unknown coefficients as well as for far more variables (even exceeding the number of observations that are available for estimation). In contrast, in EM students are trained to avoid too flexible models as they tend to overfit and thus mislead the researcher and to use parsimonious specifications instead (e.g. the famous Occam's Razor).

For the very flexible models in SSL, overfitting the model to the data at hand is a potentially very dangerous threat. Thus, all predictions are assessed on data that were not used for estimation (call 'learning' in SSL terminology), i.e. in an out-of-sample assessment. Such samples are generated either by a single random sample split or by cross-validation. EM usually does not make use of such data splits and relies mainly on in-sample assessments.

A final (subjective) difference is that the more illustrative names of SSL estimators are likely to be much more appealing to a non-specialist audience (e.g. *neural network*, *random forest*) than the technical terms used typically by EM (e.g. *ordinary least squares*).

3 Methodology for the prototype model

3.1 Notation and identification

In this section, we focus on the simplest causal model to show the main ideas of important CML approaches. We use Rubin's (1974) potential outcome language to describe the so-called static binary treatment model under unconfoundedness (e.g. Imbens, 2004).

Let D denote the treatment indicator, which is either 0 (control) or 1 (treated). The (potential) outcome of interest that realizes under treatment d is denoted by Y^d. For each member of the population, we observe only the potential outcome related to the received treatment, $Y = (1-D)Y^0 + DY^1$.[9] There are two groups of variables to condition on, $\tilde{X}$ and Z. $\tilde{X}$ contains those covariates that may be needed to correct for selection bias (confounders), while Z contains variables that define subpopulations for which an average causal effect estimate is desired.[10] $\tilde{X}$ and Z may overlap in any way. Denote the union of the two groups of variables by X, $X = \{\tilde{X}, Z\}$. The dimension of X equals p, which is treated as fixed (potentially large).

Next, we define three types of average causal effects at different levels of granularity:

$$\mathrm{IATE}(x) = E(Y^1 - Y^0|X = x),$$

$$\mathrm{GATE}(z) = E(Y^1 - Y^0|Z = z) = \int \mathrm{IATE}(x) f_{X|Z=z}(x)\mathrm{d}x,$$

$$\mathrm{ATE} = E(Y^1 - Y^0) = \int \mathrm{IATE}(x) f_X(x)\mathrm{d}x.$$

First, the individualized average treatment effects (IATEs), IATE(x), measure the mean impact of the treatment for units with specific covariate values x. The IATEs represent the causal parameters at the finest aggregation level of the features available. On the other extreme, the average treatment effect (ATE) represents the population average at large. The ATE is (or was?) the classical parameter investigated in econometric causal studies. The group average treatment effect (GATE) parameters are in between those two with respect to granularity. The analyst preselects the variables Z before estimation according to her policy interests. The IATEs and the GATEs are special cases of the so-called conditional average treatment effects (CATEs).[11]

In our setting of unconfoundedness, a necessary condition for the identification of these causal parameters is that all variables that jointly influence treatment and potential outcomes are observable (conditional independence). Such confounders must not be influenced by the treatment (exogeneity). Furthermore, for all values of such confounders, there must be a non-trivial probability of becoming treated or non-treated (common support). Finally, the *potential* outcomes must not be influenced by the treatment allocation (stable unit treatment value assumption, SUTVA). An important implication of these assumptions is that exogenous unobservables have the same distribution for treated and controls conditional on the observed confounders. Clearly, the credibility of such an identification strategy depends on the specific application. From now on, we suppose that all these conditions are met. In this case, the causal parameters of interest are identified, i.e. they can be expressed in terms of random variables, for which we sample data (i.e. for $i = 1, \ldots, N$, we observe x_i, y_i, d_i). For the IATE(x), we obtain the following such expression (*estimand*):

$$\begin{aligned}\mathrm{IATE}(x) &= E(Y|X = x, D = 1) \\ &\quad - E(Y|X = x, D = 0) \\ &= g(x, 1) - g(x, 0),\end{aligned}$$

Since GATEs and ATE are just averaged versions of the IATEs, and since realizations of X and Z are observable, they are identified as well.

Before discussing the estimation of these parameters, it is worth pointing out that Machine Learning methods only have an indirect role at the identification stage of the empirical analysis. First, they may be useful to generate additional variables from other sources, such as texts or figures. Second, their use in the estimation stage may allow to consider a larger number of actual variables (such as different measurements of distances) than in a conventional analysis that may require to keep the number of variables small.

3.2 Estimation

Once identification is credibly achieved, estimation follows. Assume for simplicity that we have a sample of N independent realizations from (Y, D, X), $\{y_i, d_i, x_i\}_{i=1}^N$. The last equation shows that the estimation of these causal parameters is a combination of prediction problems (i.e. estimating $g(x, 0)$ and $g(x, 1)$ which are conditional expectations of observable variables). This insight may suggest using the following (naïve) Causal Machine Learning estimator: (1) estimate $\hat{g}(x, 1)$ with your favourite SSL method among the treated, (2) estimate $\hat{g}(x, 0)$ with your favourite SSL method among the controls, and finally, (3) estimate the IATEs as their difference, $\widehat{\mathrm{IATE}}(x) = \hat{g}(x, 1) - \hat{g}(x, 0)$. ATEs and GATEs are obtained from the respective sample averages, i.e. $\widehat{\mathrm{ATE}} = \frac{1}{N}\sum_{i=1}^{N} \widehat{\mathrm{IATE}}(x_i)$ and $\widehat{\mathrm{GATE}}(z) = \frac{1}{N^z}\sum_{i=1}^{N} \underline{1}(z_i = z)$ $\widehat{\mathrm{IATE}}(x_i)$ (for discrete Z), where N^Z is the number of observations with observed value z and $\underline{1}(\cdot)$ denotes the indicator function.

[9] If not obvious otherwise, capital letters denote random variables, and small letters realized values. Small letters subscripted by 'i' denote the value of the respective variable for observation 'i'.

[10] $\tilde{X}$ may also contain variables that are predictors of effect heterogeneity only.

[11] Similarly, we can also define these parameters for different treatment groups to obtain, for example, average and group average treatment effects for the treated. Beyond this, there are more parameters used in the literature, like quantile treatment effects. However, for the sake of brevity we focus on ATE, GATE(z), and IATE(x) only.

Although such an approach may provide a good starting point, it has some drawbacks. The first potential problem is bias. Standard SSLs that minimize the mean squared error may be biased asymptotically, despite being consistent, in the sense that with increasing sample size the (squared) bias does not disappear faster than the variance. In this case, confidence intervals coming from the limiting distribution of such an estimator will not be centred around the true values, making valid inference difficult to impossible. Second, even without considering inference, the problem of estimating a difference is different from estimating its components.[12] Therefore, typical CML methods account for the specific structure of the causal estimation problem.

Next, we discuss more sophisticated estimation principles. Since CML is popular in many fields, and different fields may have different standards of what they consider as being a 'good' estimator, in both theoretical and practical terms, many different CML estimators for the different parameters appeared in the literature. A full review goes beyond this overview. Instead, we focus on two general, comprehensive estimation principles that gained popularity in econometrics. They are 'comprehensive' in sense that they allow the estimation of all parameters of interest either in one estimation step or in few tightly related estimation steps that use Machine Learning. These two comprehensive estimation principles are double/debiased Machine Learning (DML) and causal forests (CF). While DML uses predictive Machine Learning (SSL) inside a special moment condition, CF changes the standard random forest algorithm so that it can be used to estimate causal effects.[13] Let us briefly discuss the main ideas of these two approaches.

Chernozhukov et al. (2018) introduced double or debiased Machine Learning.[14] The main idea is that we need to estimate two types of parameters. The first type is made of low-dimensional, structural parameters (such as ATEs or GATEs) which we deeply care about. However, there are also additional potentially high-dimensional parameters (nuisance parameters), which we do not care about. The estimation of the structural parameters is based on moment conditions that depend on the nuisance parameters as well as on the structural parameters. Chernozhukov et al. (2018) show that if we choose specific moment functions that fulfil the so-called Neyman orthogonality condition, we can use a specific two-step procedure. Neyman orthogonality implies that small deviations of the nuisance parameters from their true values do not matter for the estimation of the structural parameter. Thus, under certain conditions, for the purpose of inference for the structural parameters, estimates of the nuisance parameters can be treated as true values.

In the first step of the estimator, the nuisance parameters are estimated ('learned') by 'good enough' predictive Machine Learning (SSL). The main condition for being 'good enough' is consistency and a convergence rate at least close to $N^{1/4}$, which several SSL methods fulfil (e.g. certain random forests and neural networks).[15] In this case, the estimation error from the first step SSL estimation can be ignored when solving the moment condition for the structural parameters. The resulting estimator will be $N^{1/2}$-consistent and asymptotically normal. If we use the influence function as the basis for these moment functions, DML is likely to result in efficient estimators.

How do such moment conditions look like for the estimation of the ATE under unconfoundedness? It turns out that they correspond to moment conditions identical to those of so-called doubly robust estimators (e.g. Robins et al., 1994). Denote the conditional treatment probability, i.e. the propensity score, by $p(x) = P(D = 1|X = x)$, then an alternative estimand for the ATE can be obtained that it is instructive for doubly robust and DML estimation[16]:

$$\begin{aligned}\text{ATE} =& E_X E\Bigg[g(X,1) + \frac{D(Y-g(X,1))}{p(X)} \\ & -g(X,0) - \frac{(1-D)(Y-g(X,0))}{1-p(X)}\Bigg|X=x\Bigg]\end{aligned}$$

Thus, a DML estimator for the ATE is the following:

$$\begin{aligned}\widehat{\text{ATE}} =& \frac{1}{N}\sum_{i=1}^{N}\Bigg[\hat{g}(x_i,1) + \frac{d_i(y_i-\hat{g}(x_i,1))}{\hat{p}(x_i)} \\ & -\hat{g}(x_i,0) - \frac{(1-d_i)(y_i-\hat{g}(x_i,0))}{1-\hat{p}(x_i)}\Bigg]\end{aligned}$$

[12] To see this, e.g. consider the case when both $g(d,x)$'s are highly nonlinear and thus difficult to estimate, while the treatment effect is constant. Since any estimation error of $g(1,x)$ and $g(0,x)$ that cancels due to differencing does not matter, clever joint estimators of $g(1,x)$ and $g(0,x)$ may lead to more precise estimators than treating the estimation of $g(1,x)$ and $g(0,x)$ as independent estimation problems.

[13] Lechner and Mareckova (2023a, 2023b) provide a more detailed overview over these comprehensive CML methods.

[14] DML is closely related to target maximum likelihood (van der Laan and Rubin, 2006) combined with Machine Learning which is popular, for example, in biostatistics.

[15] If some SSLs converge faster, other SSLs may be allowed to converge slower.

[16] Double robustness in its classical use means a misspecification of the parametric model of either $g(.)$ *or* $p(x)$ does not matter for consistency if at least one of the two models is correctly specified.

The potentially high-dimensional nuisance functions $\hat{g}(x_i, 1)$, $\hat{g}(x_i, 0)$, and $\hat{p}(x_i)$ are estimated by suitable SSL methods. This estimator is $N^{1/2}$-consistent, asymptotically normal, and asymptotically efficient.[17] Inference is also straightforward and computationally inexpensive.

Knaus (2022) shows how to use this approach in computationally efficient ways in empirical applications to estimate ATEs, GATEs, and low-dimensional parametric approximations of IATEs (and more). He also discusses further subtleties of the estimation which are beyond the scope of this paper. In summary, the advantage of the DML approach for low-dimensional parameter estimation is that we can leverage the power of flexible off-the-shelf Machine Learning methods while retaining a complete asymptotic distribution theory. Even when the dimension of the parameter of interest increases, like in a GATE with one or more continuous variables, DML may still have good properties, but the resulting estimators may have lower convergence rates (i.e. Zimmert & Lechner, 2019).

Meanwhile, many researchers work on extending DML. There are general theoretical extensions (e.g. Chernozhukov et al., 2022a, 2022b) as well as extensions to other treatment types and models, like continuous treatments (e.g. Klosin, 2021), dynamic models (e.g. Bodory et al., 2022b; Lewis & Syrgkanis, 2020), quantile treatment effects (Kallus et al., 2020), or mediation analysis (Farbmacher et al., 2022), to mention only a few important ones.

While DML inserts standard predictive Machine Learning estimators into moment-condition-based estimators, causal trees (CT, Athey & Imbens, 2016) and causal forests (CF, Wager & Athey, 2018) adapt Machine Learning algorithms to the causal question. This works particularly well for experiments and unconfoundedness, because in these cases the effect estimates are based on treated and controls with similar values of the covariates. This similarity of covariate values of different observations is also a defining feature of a (final) leaf of a CART. Thus, the main difference between a CART and a CT is that the latter computes average outcome differences of treated and controls (with or without propensity score weighting) in the final leaves and uses a splitting criterion adapted to causal analysis. This adapted splitting rule is based on maximizing treatment effect heterogeneity instead of minimizing the (squared) prediction error. The variance-bias trade-off in a CT also requires finding an optimal leaf size that is small enough to make the bias small but not so small that the variance of the estimator become too large.

However, CTs are rarely used in applications for the same reason as RF may be preferred to CARTs for prediction tasks. As in a RF, final leaves in a CF are small and, thus, bias is low. This is possible because the variance of the prediction from a single leaf is reduced by averaging over such predictions from many randomized trees. As in RF, randomization of these (deep) trees is done by randomly selecting splitting variables and by inserting randomness via the data used in building the tree. However, while trees in RF are typically estimated on bootstrap samples, the theory of CF requires to use subsampling (i.e. sampling without replacement) instead. Another important concept used is 'honesty', i.e. the data used to build the CT is different from the data to compute the effects given the CT. This is achieved by sample splitting. Under various additional regularity conditions, estimated IATEs from such CF's converge to a normal distribution centred at the true IATEs. As usual GATEs and IATEs are then obtained by averaging.

The modified version of the CF (MCF) proposed by Lechner (2018) and theoretically analysed by Lechner and Mareckova (2023a, 2023b) uses a different splitting rule. The MCF also exploits the fact that CT and CF estimators can be expressed as weighted sum of the outcomes, where the weights directly follow from the algorithm. Thus, aggregating IATEs to GATEs and ATE amounts to aggregating the weights directly. The MCF exploits this feature of CF by using a weight-based inference procedure that allows one-step estimation and inference for ATEs, GATEs, and IATEs which turns out to be very convenient for applied work.[18]

Another popular approach that is somewhat in-between CF- and DML-based estimators is the generalized random forest (GRF, Athey et al., 2019). The main idea of GRFs is to obtain the parameters of interest from local maximum likelihood estimation, where specifically designed random forests are used to provide the local weighting scheme. As with DML, this approach is also directly applicable to settings other than unconfoundedness, to different treatment parameters, and to more complex causal structures.

It is a practically important feature of the literature that the authors of the method papers provide users with implementations of their estimators in Python or R, or both (and sometimes even in Stata). Fairly user-friendly packages are, for example, described (1) in Bach et al.

[17] An additional condition for this approach to have good properties is that the observations used to evaluate the nuisance functions should not be used to estimate them (also called cross-fitting). This is achieved by using some form of sample splitting, e.g. cross-validation. Kennedy (2022) provides a very comprehensive as well as very instructive discussion of the theory of DML and how it can be applied to various models and parameters.

[18] Lechner and Mareckova (2023a, 2023b) show (under certain regularity conditions) consistency and asymptotic normality of IATEs, GATEs, and ATEs, as well as $N^{1/2}$-convergence of ATEs.

(2022) for DML, (2) in Bodory et al. (2022a) for the MCF, and (3) in Athey and Wager (2019) for the GRF.

3.3 Decision-making: optimal policy

One of the main advantages of CML estimators that appeared recently in the literature is that we obtain a much better picture of the underlying heterogeneity of the impact of a policy. While ATEs are informative on how the target population will gain or lose on average, GATEs might reveal subgroups, e.g. based on education or gender, that benefit more than others, or might even be hurt by participating in the policy. This information will subsequently help policy makers to reconsider the target population, redesign the allocation rules, or change the policy more fundamentally.

Even when the policy works as intended for such broadly defined target populations, there may still be room for potential improvements by using the fine-grained information that is potentially contained in the IATEs. The subfield of CML that deals with such tasks comes under the headings of *optimal policy*, *policy learning*, or *statistical assignment rules*. The goal is to build an algorithm, or in more fancy language, an artificial intelligence system (AI), that suggests if a specific person is participating in the policy. This decision is based solely on the observable information contained in x (or a subset of x). This section attempts to show some important ideas and issues in this field. However, it will not be comprehensive at all (even less so than the previous sections), will remain on a very non-technical level (even though most of this literature is very technical), and will ignore most of this rapidly developing literature.

The key ingredients into such a stochastic decision algorithm are the following: (1) the objective function of the decision-maker; (2) estimates of individual policy impacts (which in many cases will be IATEs) for the relevant population; and (3) possible constraints. Let us consider them in turn.

The objective function is a function of outcomes. Usually, the policy maker is assumed to be risk neutral, meaning that her objective (welfare) function is a weighted sum of the expected outcomes under a specific policy. However, this objective function may be adapted, for example, to allow for risk aversion, i.e. to integrate equity concerns, etc. For example, nonlinear transformations that give more weight to specific parts of the outcome distribution may be attractive (e.g. a certain amount of additional earnings of a poor person may be more important to the policy maker than if a rich person gains the same amount).

The second ingredients are estimates of the policy effects as a function of observable features. When these policy effects correspond to the IATEs, it might appear natural to use the estimators discussed before. However, the fact that an estimator has good properties for the estimation of the IATE per se, like consistency or asymptotic normality, does not necessarily mean that it also has good properties for the estimation of decision rules, which is a different statistical problem. In the language of SSL, this is a classification problem, not a regression problem (as for effect estimation).[19] Furthermore, as the key information for these algorithms comes from the IATEs, IATEs must be identified in the first place, usually ruling out research designs that identify effects only for specific subgroups, such as difference-in-differences (identification only for the treated), instrumental variables (only for compliers), or regression discontinuity designs (only for the population around the discontinuity).[20]

Furthermore, there may be constraints that need to be incorporated into the algorithm. Such constraint might relate to the resources available. Furthermore, fairness and antidiscrimination considerations may play a role. Such issues could sometimes be tackled by omitting some critical values from the covariates used to decide the allocation (e.g. by not using information on race or gender). However, as such variables may also be correlated with many other variables, this does not always solve the problem. In addition, there may be certain restrictions peculiar to the specific application that should be considered. Finally, there may be computational constraints. While efficient classification-type algorithms are available for certain standard situations with a binary treatment, this is not the case for multiple treatments or even more complex treatment models.

The seminal paper for the current literature in this field is Manski (2004). He systematically investigates how to build and theoretically evaluate algorithms that perform decision making under uncertainty. Hirano and Porter (2020) provide a nice overview over this literature. Some of the methods proposed in this literature link this decision problem to DML in the sense that they are using the same ingredients as DML uses for point estimation, i.e. the doubly robust score. Two examples of such papers are Athey and Wager (2021) for binary treatments and Zhou et al. (2022) for multiple treatments. These two papers also address another issue that is important in practice. If such an AI system is used together with human decision-makers, it is important that humans accept it as a helpful tool. For this to happen, it is helpful if the human can understand on which criteria the recommendation of the AI is based upon. Using low-dimensional (policy) trees

[19] Fernández-Loría and Provost (2022) explain the issues in more detail in a non-technical way.

[20] This problem can be overcome if such designs are amended by (plausible) effect homogeneity assumptions that allow the extrapolation to the population of interest.

could be a good way to communicate the results from the AI to the human, as suggested, for example, in those two papers.[21]

4 An example: active labour market policies in Flanders

In this section, we illustrate the usefulness of CML methods with an empirical example. This example is an evaluation study of the effects of active labour market programmes (ALMP) in Flanders. ALMP mainly aim at improving reemployment chances of individuals who became unemployed. The programmes mainly consist of different types of training courses, subsidized employment opportunities in a protected sector, and some types of support for private firms hiring such unemployed. Many countries run such programmes, and there is a large literature on their effects on earnings and reemployment chances (e.g. Card et al., 2018). The main questions that these papers attempt to answer can be simplified to '*what works and for whom?*' and '*who should we send to which programme*?'. If we find good answers to these questions, we might improve the life situation of many unemployed as well as devise ways to use public resources more efficiently.

In Cockx et al. (2023), we investigate the effects of the training part of the Flemish ALMP based on administrative data from the Flemish employment services covering a recent inflow of about 60,000 formerly employed individuals into unemployment. The field developed some standards on which types of control variables are needed (e.g. Lechner & Wunsch, 2013) to remove confounding in such application of European ALMP using administrative data. Based on these insights combined with our institutional knowledge about the caseworker-based selection process of unemployed into specific programmes in Flanders, we argue that the data is rich enough to be able control for all major confounders such that a selection-on-observables strategy is plausible. For three of the four training types considered, this claim could not be rejected by a placebo study (see paper for details). Therefore, the summary of the results here is based on these three programme types only (short vocational training, long vocational training, and so-called orientation training). By concentrating on training programmes only, we are not able to give a full answer to the questions posed above. Instead, we investigate the more restricted choice between different types of training programmes (and no participation in any programme in the first 9 months of this unemployment spell). All results in Cockx et al. (2023) are based the Python version of the MCF algorithm.[22]

The first set of (selected) results relate to the average performance of the programmes and how this performance evolves over time, in comparison with not participating in any programme in a certain period. For this purpose, we estimate the average treatment effect (ATE). Since we have monthly observations of labour market states several years before and up to 30 months after programme participation, we estimate an ATE for each of these 30 months separately. The corresponding results in Fig. 1 relate to the probability of being employed in first labour market in a particular month after the start of the programme. We find that after the usual lock-in effects in the first few months after programme start, all training programmes lead to positive effects, and such effects are largest for the short training programmes. This overall positive assessment is however subject to the caveat that the positive medium-term effects do not necessarily mean that the programmes are cost-effective, as cost data is not available in the data used.

Next, we are interested whether the effectiveness of the training programmes depend on the command of the local language. In other words, we want to estimate a GATE for the four different levels of proficiency of Dutch (which is the local language in Flanders) that are observable in our data. We could perform this analysis for every post-treatment month, but to ease the presentation of the results we focus on an aggregate outcome measure instead. It is constructed by summing up the post-treatment months in employment over the 30 months available. As a further difference to the previous figure, the results and their confidence intervals are presented as deviations from the ATE. In Fig. 2, we show the results for short vocational training compared to non-participation.

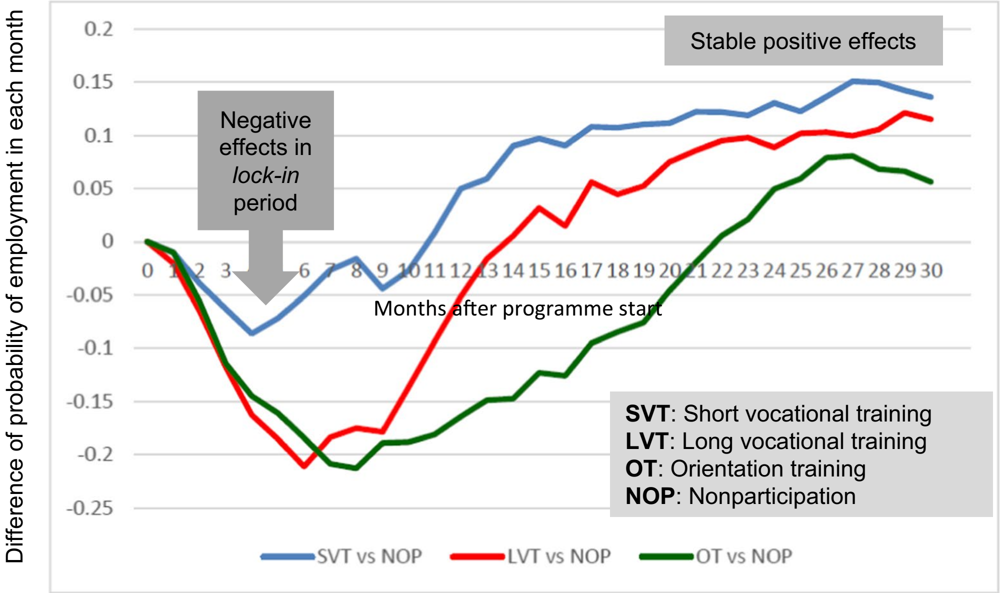
The GATEs shown in Fig. 2 indicate that effectiveness of the programmes declines with an increased proficiency in Dutch. Although this result appears to be clear-cut on its own, its correct interpretation is that the effectiveness of the training programme *correlates* with local language proficiency. The differences of the effects may or may be *caused* by language proficiency. The reason is the usual one, i.e. language proficiency itself correlates with variables that are not held constant for computing the different 4 GATEs, like migration background, human capital, for example.[23,24]

[21] A related approach from the Machine Learning literature is Amram et al. (2022).

[22] It is freely available on PyPI.

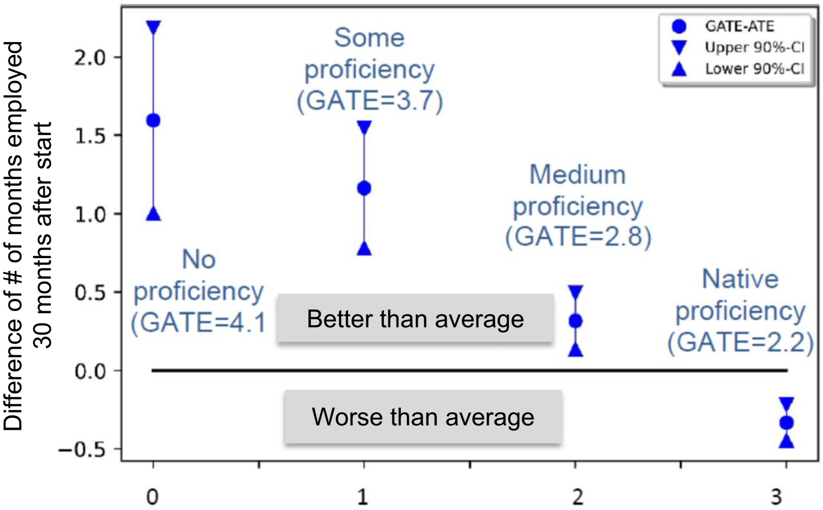
[23] Bansak (2021) addresses this issue for experiments while Bansak and Nowacki (2022) focus on regression discontinuity designs. Bearth and Lechner (2023) introduce a new parameter that keeps the other 'background' variables constant and call it MGATE (moderated GATE). They propose double Machine Learning estimation under unconfoundedness for the special case of a discrete mediator.

[24] To avoid ex-post-data mining that would invalid any inference, the variables for which to compute the GATEs should be selected in advance (ideally derived from theory), and they should be very few.



Confidence bands are not shown. Effects are significant, with the exceptions of the months for which they are close to 0.

Fig. 1 Evolvement of the ATE over time



Dutch proficiency displayed on horizontal axis. Vertical axis denotes difference of respective GATE with ATE. (GATE-ATE) and its 90% confidence interval shown. Dutch proficiency varies between no proficiency (0) and native proficiency (3).

Fig. 2 GATE for language proficiency for short training vs. non-participation

The parameters at the finest level of granularity are the IATEs. They may also be used for an explorative analysis of the structure of the causal heterogeneity. One way is to compute the IATEs for each value of x observed in the sample, cluster them into few homogeneous groups, for example, with a k-means++ clustering algorithms (Arthur & Vassilvitskii, 2007), as done in this empirical study, sort the groups according to their mean value of the IATEs, and then analyse moments or quantiles of the covariates in the respective cluster. This way, we may *find* additional patterns of the covariates. However, this simple IATE-based data mining procedure is explorative in nature, and inference is not available.[25] Of course, there are many alternative procedures that can be used to better understand which variables are correlated with IATEs (like any regression method or the more sophisticated approach of Chernozhukov & Fernandez-Val, 2022). With respect to the findings in this application, the just described cluster analysis reflected only those differences that appeared already in the comparisons of the preselected GATEs. Thus, no new causal heterogeneity was discovered.

Finally, the estimated IATEs (more precisely the underlying potential outcomes in the case of multiple treatment like in this study), or (asymptotically) unbiased estimates thereof, could be used to investigate the quality of the observed allocation of unemployed. Here, we investigated several allocations based on the estimated IATEs. The goal was to compare the actual allocation to the 'best possible' allocation, a purely random allocation, as well as an 'explainable' allocation (in addition to several other allocation schemes). In these simulations, we used the observed programme shares as upper bounds. This was to ensure that we investigated the gains through redistribution, instead of programme expansion. Of course, we might be able to obtain interesting insights by simulations based on specific expansions or reductions of the programmes. However, without having cost information, this appeared to be less attractive. Except for the random allocation, all allocations investigated improved upon the observed allocation. The 'best' allocation has been computed by allocating the treatment that has the highest potential outcome for the specific value of x. However, such a black-box-based AI may not be acceptable for caseworkers in the field; therefore, we investigated also an allocation based on a low-level policy tree (as proposed by Zhou et al., 2022). This easy-to-understand but less efficient rule suggested to allocate specific types of unemployed migrants to short vocational courses, unemployed with specific longer employment histories to long vocational courses, and not to use orientation training at all (for details, see Cockx et al., 2023).

In summary, this exercise led to useful information about the ALMP in Flanders that might be helpful in futures redesigns of the ALMP. Of course, this is just an example, and there are many other policies (or decision situations in general) for which such fine-grained heterogeneity and allocation analysis is very useful.

5 The promise of Causal Machine Learning and some of its limits

The first step in any causal analysis is to find a credible research design, i.e. a set of identifying assumptions such that the causal effects of interest are credibly identified for the population of interest. Does CML help with this task? Strictly speaking, and as already mentioned above, not really. In practical terms, it is still helpful, because (1) it may allow to include more covariates in the estimation and thus more confounders can be controlled for and (2) there is no need to impose additional unjustifiable parametric assumptions on all or parts of the estimation problem as it is necessary when using conventional parametric or semiparametric estimators.

The part where CML really shines is the flexible estimation of causal effects at various aggregation levels, as well as employing the information on the fine-grained causal heterogeneity (IATEs) to obtain 'good' allocation schemes and decision rules. While the ATE is informative about the effectiveness of the policy at large, the GATEs help to better understand for which groups the policy is effective. The IATE-based optimal policy algorithms help to find better ways of targeting policies such that (ideally) the overall goal of the policy/decision-maker can be fulfilled in an optimal way. It is important to note that the literature has developed such that we now know key (asymptotic) statistical properties for most of the estimators and for some allocation algorithms. In certain circumstances, there is also the possibility to use the insights from double/debiased Machine Learning to develop estimators that are asymptotically efficient by combining standard Supervised Statistical Learning procedures with a specific choice of moments for low-dimensional causal effects. These statistical guarantees provide a large sample safeguard against getting 'crazy' (i.e. too noisy, or inconsistent) results from these new and sometimes very complex nonparametric methods.

The focus on unconfoundedness as identifying assumption in Sects. 3 and 4 was on purpose because most of the literature has proposed methods for this case (and for experiments, which can be seen as a special case of unconfoundedness). The reason why this happened is likely related to the fact that the assumptions underlying experiments and unconfoundedness are strong enough to identify the marginal distribution of the potential outcomes conditional on covariates. Thus, in this case, we have enough information to go the full way from the estimation of effects at the different aggregation levels to devising allocation algorithms.

However, the literature is full of empirical studies with observational data for which unconfoundedness does not appear to be plausible. Thus, alternative identification strategies are pursued. They may be less powerful,

[25] Chernozhukov and Fernandez-Val (2022) propose a more sophisticated procedure for this idea and derive valid inference procedures.

but in the specific situations more credible, than unconfoundedness. We will consider the most important of such strategies (in their canonical form), i.e. differences-in-differences (DiD), instrumental variables (IV), and the regression discontinuity design (RDD) in turn.[26]

DiD designs usually identify treatment effects for the treated subpopulation. As treated and non-treated observations may systematically differ with respect to unobservables (which is the justification for using DiD in the first place), this implies that assignment algorithms with are based on X only (as the treatment status is unknown before the assignment) are of little use. However, CML helps to estimate ATEs and GATEs efficiently by double Machine Learning (e.g. Chang, 2020; Zimmert, 2019).

In instrumental variable estimation (IV), the estimated causal effect is valid only for the subpopulation that potentially reacts to a (potential) change of the value of the instrument with a change of the treatment status. Since these so-called compliers are unobservable, it appears impossible to use such complier-specific effects to optimally assign treatments without further assumptions.[27] A very powerful instrument that leads to 100% compliance could of course solve this issue. Such instruments are rare, though. While CML is of limited use for finding optimal allocations, CML is useful for estimating ATE and GATE-like effects for the respective complier subpopulation. CML may be particularly helpful when there is the need to control for many covariates to ensure the validity of the instrument, or when there are many instruments. Indeed, there is a large literature on using CML methods in IV. This literature started with a constant effect model (Belloni et al., 2012), but the GRF and DML methodologies could also be applied for IVs allowing for effect heterogeneity, as well as for estimating heterogeneous effects (Athey et al., 2019; Syrgkanis et al., 2019).

In regression discontinuity designs (RDDs), we either obtain identification of causal effects for the population local to the cut-off that provides the identification results (sharp RDD), or for the compliers among this local-to-the-cut-off population (fuzzy RDD). If this cut-off is not of policy interest, allocation rules valid for such a population appear to be less valuable. The estimation task in a RDD setting is usually to adjust for different values of a single so-called running variable. For this case, well-established nonparametric procedures are available, and thus, the potential of CML appears to be low. Nevertheless, Kreiß and Rothe (2023) show that the information in other covariates that are predictive for the outcome can be used to improve the precision of estimator the causal effect.[28]

5.1 Conclusion and the road ahead

The new developments at the intersection of modern Machine Learning methods and causal analysis of policies, or decisions more generally, provide empirical researchers with a new toolkit that is substantially more powerful than what we had in the past. It does not only allow getting more robust and more precise estimates of average effects but also to systematically investigate the underlying fine-grained heterogeneity. This heterogeneity may then in turn be used to investigate and find allocation/treatment rules that decision-makers find optimal given their objective functions and constraints. This additional information should be very valuable to every decision-maker. In the sphere of public policy, they should help to improve the specific policies to reach their goals as well as leading to a more efficient use of public resources.

One might even go one step further and imagine using these tools to build a semi-autonomous process in which policies will be evaluated continuously and their content and allocation rules adapted accordingly. Whether the resulting artificial intelligence (AI) is allowed to make the decision autonomously or serves as additional information for human decision-makers will then depend of course on the quality of its decisions as well as on the preferences of those in charge of and subject to the policy.

Until we reach this utopian state sometime in the future (if ever), there are still many issues that require further research. Let me mention just a few of them: One open issue concerns the finite sample properties of the algorithms. The statistical guarantees we have so far for the CML methods are asymptotic. There are only a limited number of studies that have systematically investigated the finite sample properties of various CML estimators.[29] This may be a particular concern for CML, as the estimators allow for very flexible estimation approaches (such a CF or DML) which are likely to require larger samples than conventional parametric or semiparametric methods. This may or may not be a particular issue for the estimation of the allocation rules.

The next issue relates to tuning parameter choice. As most CML parameters depend on several tuning parameters, choosing them in a ‘good’ way may be of particular concern since the theoretical guidance on how to

[26] We consider the cases without any additional homogeneity assumptions. Such assumptions might allow to extrapolate the effects of the population for which they are identified to other populations, but they are usually difficult to justify in specific applications.

[27] Qiu et al. (2021) and Cui and Tchetgen Tchetgen (2021) provide such approaches.

[28] CML can be used to obtain similar improvements in the experimental context as well (Chiang et al., 2023).

[29] One example is the paper by Knaus et al. (2021) that investigate some estimators for the IATEs.

determine the tuning parameters for CML, in contrast to conventional ML, is more limited.

Another open question is the relevance of common support issues and how to deal with them. Are different CML methods more, or even less sensitive to these issues than conventional methods? The suspicion is that the answer may be method specific, but systematic studies on this topic are still lacking. The practical issue is then how to best deal with situations in which we have limited (weak) or no common support? Is the choice of a robust estimator already (almost) enough (if it exists at all), or do we need to explicitly limit the scope of the analysis to a smaller population, and if so, how do we determine such subpopulation?[30]

The development of optimal allocation schemes also brings a couple of additional open issues. To start with, there are still many objective functions and types of constraints for which there are no or limited statistical guarantees and efficient computation is unclear. Computation is also an issue on its own. Many of these allocation algorithms are computational very complex and lack of efficient computation certainly restricts their use in practice. Another topic are the statistical guarantees that are different from the ones we are used to when considering point estimates. I believe, it is still an open issue how relevant they are in practical applications. Inference for such allocation algorithms also appears to be a particular difficult problem, where more research is certainly needed. Finally, all the topics that are relevant for AI in general are important here as well. These begin with bias (not the statistical one), discrimination, and fairness of the algorithms and end with the need to prevent misuse by intentional manipulation by backdoors, hacking, or other criminal activities.

In this paper, I have discussed the advantages and the potential of the CML toolkit. This leaves us with the question of what will happen with the (still) classical toolkit, especially in microeconometrics. Think about matching or regression analysis. It is my guess that most of these methods will simply become irrelevant, at least when we analyse large data causally. For large data, the new CML methods, skilfully applied and well understood, are superior. If true, this has important implications for the teaching of econometrics that go beyond the now almost common sense that students of applied econometrics need much better, or at least different coding skills than in the past. It will require a substantial transformation of the econometrics curriculum.

[30] For a limited range of estimators and parameters, Ma et al. (2023) discuss theoretically appealing ways to address issues of weak common support by trimming observations together with a subsequent bias correction to mitigate the possibly bias-increasing effect trimming may have.

Abbreviations

AI	Artificial intelligence
ALMP	Active labour market programmes
ATE	Average treatment effect
CART	Classification And Regression Tree
CF	Causal forests
CML	Causal Machine Learning
DAG	Directed acyclic graphs
DGP	Data generating process
DiD	Differences-in-differences
DML	Double/debiased machine learning
EM	Econometrics
GATE	Group average treatment effect
IATE	Individualized average treatment effect
IV	Instrumental variables
LASSO	Least Absolute Shrinkage and Selection Operator
MCF	Modified causal forest
MSE	Mean squared error
OLS	Ordinary least squares
RDD	Regression discontinuity design
RF	Random forests
SSL	Supervised Statistical Learning

Acknowledgements

I thank the audience in Basel for the very interesting discussions during and after the presentation. Hannah Busshoff, the editor and a referee, provided valuable comments on a previous draft of this paper.

I am also affiliated with CEPR, London, CESIfo, Munich, IAB, Nuremberg, IZA, Bonn, and RWI, Essen. This paper is based on my keynote lecture at the annual meeting of the German Economic Association (Verein für Socialpolitik) in Basel, September 2022. I thank the organizers, in particular Joachim Winter, for inviting me. The slides from this presentation can be downloaded from https://drive.switch.ch/index.php/s/vFheIKERQqoyH5c.

Author contributions

I fully contributed to my paper. The author read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

Not applicable.

Declarations

Competing interests

None.

Received: 20 February 2023 Accepted: 26 April 2023
Published online: 08 May 2023

References

Amram, M., Dunn, J., & Zhuo, Y. D. (2022). Optimal policy trees. *Machine Learning, 111*, 2741–2768.

Angrist, J. D. (2022). Empirical strategies in economics: Illuminating the path from cause to effect. *Econometrica, 90*, 2509–2539.

Angrist, J. D., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives, 24*(2), 3–30. https://doi.org/10.1257/jep.24.2.3

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms* (pp. 1027–1035). Society for Industrial and Applied Mathematics Philadelphia, PA, USA.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science, 355*, 483–485.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America, 113*(27), 7353–7360.

Athey, S., & Imbens, G. (2019). Machine learning methods economist should know about. *Annual Review of Economics, 11*, 685–725.

Athey, S., & Luca, M. (2019). Economists (and economics) in tech companies. *Journal of Economic Perspectives, 33*(1), 209–230.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized Random Forests. *Annals of Statistics, 47*(2), 1148–1178.

Athey, S., & Wager, S. (2019). Estimating treatment effects with Causal Forests: An application. *Observational Studies, 5*, 37–51.

Athey, S., & Wager, S. (2021). Policy learning with observational data. *Econometrica, 89*, 133–161.

Bach, P., Chernozhukov, V., Kurz, M., & Spindler, M. (2022). DoubleML—An object-orientated implementation of double machine learning in Python. *Journal of Machine Learning Research, 23*, 1–6.

Bansak, K. (2021). Estimating causal moderation effects with randomized treatments and non-randomized moderators. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 184*(1), 65–86.

Bansak, K., & Nowacki, T. (2022). *Effect heterogeneity and causal attribution in regression discontinuity designs*. MIMEO.

Bearth, N., & Lechner, M. (2023). *Double/debiased Machine Learning for moderation analysis*. MIMEO.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica, 80*(6), 2369–2429.

Bodory, H., Busshoff, H., & Lechner, M. (2022a). High resolution treatment effects estimation: Uncovering effect heterogeneities with the Modified Causal Forest. *Entropy, 24*, 1039.

Bodory, H., Huber, M., & Laffers, L. (2022b). Evaluating (weighted) dynamic treatment effects by double machine learning. *Econometrics Journal, 25*(3), 628–648.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth.

Brodeur, A., Cook, N., & Heyes, A. (2020). Methods matter: p-hacking and publication bias in causal analysis in economics. *American Economic Review, 110*(11), 3634–3660.

Card, D., Kluve, J., & Weber, A. (2018). What works? A meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association, 16*(3), 894–934.

Chang, N.-C. (2020). Double/debiased Machine Learning for difference-in-differences models. *The Econometrics Journal, 23*(2), 177–191.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased Machine Learning for treatment and structural parameters. *Econometrics Journal, 21*, C1–C68.

Chernozhukov, V., Escanciano, J. D., Ichimura, H., Newey, W. K., & Robins, J. M. (2022a). Locally robust semiparametric estimation. *Econometrica, 90*, 1501–1535.

Chernozhukov, V., & Fernandez-Val, I. (2022). The sorted effects methods: Discovering heterogeneous effects beyond their averages. *Econometrica, 86*(6), 1911–1938.

Chernozhukov, V., Hansen, C., Spindler, M., & Syrgkanis, V. (2023). *Applied causal inference powered by ML and AI*. MIMEO.

Chernozhukov, V., Newey, W. K., & Singh, R. (2022b). Automatic debiased machine learning of causal and structural effects. *Econometrica, 90*, 967–1027.

Chiang, H., Matsushita, Y., & Otsu, T. (2023). *Regression adjustment in randomized controlled trials with many covariates*. arXiv: https://arxiv.org/abs/2302.00469

Cockx, B., Lechner, M., & Bollens, J. (2023). Priority to unemployed immigrants? A Causal Machine Learning evaluation of training in Belgium. *Labour Economics, 80*, 102306.

Cui, Y., & Tchetgen Tchetgen, E. (2021). A semiparametric instrumental variable approach to optimal treatment Regimes under endogeneity. *Journal of the American Statistical Association, 116*(533), 162–173.

Farbmacher, H., Huber, M., Laffers, L., Langen, H., & Spindler, M. (2022). Causal mediation analysis with double machine learning. *Econometrics Journal, 25*(2), 277–300.

Fernández-Loría, C., & Provost, F. (2022). Causal decision making and causal effect estimation are not the same … and why it matters. *INFORMS Journal of Data Science, 1*(1), 4–16.

Graham, B. S. (2020). Network data. In B. S. Graham (Ed.), *Handbook of Econometrics, Volume 7A, Chapter 2*. Amsterdam: Elsevier.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer (10th printing with corrections, 2013).

Heckman, J. J. (1997). Instrumental Variables. *Journal of Human Resources, 32*, 441–462.

Hirano, K., & Porter, J. R. (2020). Asymptotic analysis of statistical decision rules in econometrics. In S. N. Durlauf, L. P. Hansen, J. J. Heckman, & R. L. Matzkin (Eds.), *Handbook of econometrics, Vol 7A*. Amsterdam: Elsevier.

Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation of non-orthogonal problems. *Technometrics, 12*, 55–67.

Huber, M. (2023). *Causal analysis: Impact evaluation and Causal Machine Learning with applications in R*. MIT Press.

Imbens, G. W. (2004). Nonparametric estimation of Average Treatment Effects under exogeneity: A review. *The Review of Economics and Statistics, 86*, 4–29.

Imbens, G. W. (2022). Causality in econometrics: Choice vs chance. *Econometrica, 90*, 2541–3266.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local Average Treatment Effects. *Econometrica, 62*, 446–475.

Imbens, G., & Rubin, D. (2015). *Causal inference for statistics, social, and biomedical sciences*. Cambridge University Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning (4th printing)*. New York: Springer.

Kallus, N., Mao, X., & Uehara, M. (2020). *Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond*. arXiv: https://arxiv.org/abs/1912.12945.

Kasy, M., & Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica, 89*(1), 113–132.

Kennedy, E. (2022). *Semiparametric doubly robust targeted double machine learning*. arXiv: https://arxiv.org/abs/2203.06469.

Klosin, S. (2021). *Automatic double machine learning for continuous treatment effects*. arXiv: https://arxiv.org/abs/2104.10334.

Knaus, M. (2022). Double machine-learning-based programme evaluation under unconfoundedness. *The Econometrics Journal, 25*, 602–627.

Knaus, M., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal, 24*, 134–161.

Kock, A. B., Preinerstorfer, D., & Veliyev, B. (2022). Functional sequential treatment allocation. *Journal of the American Statistical Association, 117*(539), 1311–1323.

Kreif, N., & DiazOrdaz, K. (2019). *Machine learning in policy evaluation: New tools for causal inference*. arXiv: https://arxiv.org/abs/1903.00402.

Kreiß, A., & Rothe, C. (2023). Inference in regression discontinuity designs with high-dimensional covariates. *The Econometrics Journal* **(forthcoming)**.

Leamer, E. (1983). Let's take the con out of econometrics. *American Economic Review, 73*(1), 31–43.

Lechner, M. (2018). *Modified Causal Forests for estimating heterogeneous causal effects*. arXiv: https://arxiv.org/abs/1812.09487.

Lechner, M., & Mareckova, J. (2023a). Causal Machine Learning in economics: An applied perspective. In K.F. Zimmermann (Ed.), *Handbook of labor, human resources and population economics*. Springer **(forthcoming)**.

Lechner, M., & Mareckova, J. (2023b). *Comprehensive Causal Machine Learning*. mimeo.

Lechner, M., & Wunsch, C. (2013). Sensitivity of matching based program evaluations to the availability of control variables. *Labour Economics, 21*, 111–121.

Lewis, G., & Syrgkanis, V. (2020). *Double/debiased Machine Learning for dynamic treatment effects*. arXiv: https://arxiv.org/abs/2002.07285.

Lieli, R. P., Hsu, Y.-C., & Reguly, A. (2022). The use of machine learning in treatment effect estimation. In F. Chan & L. Mátyás (Eds.), *Econometrics with machine learning, advanced studies in theoretical and applied econometrics*, Vol. 53, Chapter 3. Springer.

Ma, Y., Sant'Anna, P. H., Sasaki, Y., & Ura, T. (2023). *Doubly robust estimators with weak overlap*. arXiv: https://arxiv.org/abs/2304.08974.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica, 72*, 1221–1246.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An applied econometric approach. *Journal of Economic Perspectives, 31*(2), 87–106.

Pearl, J. (2000). *Causality—Models, reasoning, and inference*. Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). *The book of why*. Allen Lane.

Qiu, H., Carone, M., Sadikova, E., Petukhova, M., Kessler, R. C., & Luedtke, A. (2021). Optimal individualized decision rules using instrumental variable methods. *Journal of the American Statistical Association, 116*(533), 174–191.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect. *Mathematical Modelling*, *7*, 1393–1512, with 1987 Errata to A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect. *Computers and Mathematics with Applications*, *14*, 917–921; 1987 Addendum to A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect. *Computers and Mathematics with Applications*, *14*, 923–945; and 1987 Errata to Addendum to 'A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect'. *Computers and Mathematics with Applications, 18*, 477.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association, 89*(427), 846–866.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688–701.

Shah, V., Kreif, N., & Jones, A. M. (2021). Machine learning for causal inference: estimating heterogeneous treatment effects. In N. Hashimzade & M. A. Thornton (Eds.), *Handbook of research methods and applications in empirical microeconomics*, Chap. 16. Edward Elgar Publishing.

Soleymani, A., Raj, A., Bauer, S., Scholkopf, B., & Besserve, M. (2022). Causal feature selection via orthogonal search. *Transactions on Machine Learning Research*, 08/2022.

Syrgkanis, V., Lei, V., Oprescu, M., Hei, M., Battocchi, K., & Lewis, G. (2019). Machine learning estimation of heterogeneous treatment effects with instruments. *Advances in Neural Information Processing Systems, 32*, 1–10.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B, 58*, 267–288.

Van der Laan, M. J., & Rubin, D. B. (2006). Targeted maximum likelihood learning. *International Journal of Biostatistics*, *2*(1), Article 11.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using Random Forests. *Journal of the American Statistical Association, 113*(523), 1228–1242.

Zhou, Z., Athey, S., & Wager, S. (2022). Offline multi-action policy learning: generalization and optimization. *Operations Research* **(forthcoming)**.

Zimmert, M. (2019). *Efficient difference-in-differences estimation with high-dimensional common trend confounding*. arXiv: https://arxiv.org/abs/1809.01643.

Zimmert, M., & Lechner, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *Econometrics Journal* (forthcoming). arxiv: https://arxiv.org/abs/1908.08779.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B, 67*, 301–320.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.