Journal of MANAGEMENT JOM

UNVEILING UNEXPLORED INTERACTIONS IN ORGANIZATIONAL PHENOMENA: CAUSAL MACHINE LEARNING FOR ABDUCTIVE THEORIZING

Journal:	*Journal of Management*
Manuscript ID	Draft
Manuscript Type:	Original Research
Keywords:	Artificial intelligence (AI) < Technology/Future of work < MICRO, Quantitative < METHODS, Pay transparency < Compensation & Benefits < MICRO, Strategic Human Capital (from macro list) < Human Resources (other, incl strategic/macro) < MICRO
Abstract:	In traditional moderation analysis, the estimation of interaction effects is typically limited to those interactions that are theoretically justified to limit the risk of overfitting and multicollinearity. However, such manual pre-selection of interaction terms may hinder a more nuanced understanding of complex phenomena. We develop a novel combination of double machine learning (DML) and directed acyclic graphs (DAG) as an empirical strategy to robustly explore and estimate interaction effects. The paper contributes to the rapidly developing methods for abductive theory elaboration, allowing researchers to unveil relevant moderators within explicit theory-informed bounds. By algorithmically supporting the handling of interaction terms, DML addresses limitations of traditional moderation analysis relying on iterative, ad-hoc inclusion of interaction terms. We demonstrate the benefits of our empirical strategy by elaborating on existing human capital theory explaining gender wage gaps. While robustly addressing overfitting and multicollinearity concerns, our methodology unveils that investments into human capital enable women to decrease the gender wage gap. Also, we systematically point to additional moderating effects related to an individual's socio-economic background. For example, higher levels of parental human capital attainment further reduce the gender wage gap. Overall, we contribute to the growing literature on applying machine learning methods in management research.

UNVEILING UNEXPLORED INTERACTIONS IN ORGANIZATIONAL PHENOMENA: CAUSAL MACHINE LEARNING FOR ABDUCTIVE THEORIZING

INTRODUCTION

Machine learning (ML) methods are often useful for developing and elaborating theory from empirical data on organizational phenomena (von Krogh, Roberson, & Gruber, 2023). Such methods complement human sense making and theorizing when the vast available data challenges scholars' cognitive ability to identify and reason about patterns in the data. Abductive reasoning—the generation of the best possible explanation for an observed phenomenon (Bamberger, 2019; Elkjaer & Simpson, 2011; Peirce, 1878; Sætre & Van De Ven, 2021; Shani, Coghlan, & Alexander, 2020)—lies at the core of theory elaboration. Patterns recognized with the support of conventional ML methods are an important input to abductive reasoning for building and developing theories that explain them (Shrestha, He, Puranam, & von Krogh, 2021). Yet, even with the support of ML-based methods, abductive reasoning can be challenging. While conventional ML methods allow researchers to examine how a very large set of independent variables correlates with a dependent variable (He, Puranam, Shrestha, & von Krogh, 2020) and thus detect patterns that might otherwise have been missed, they do not inherently provide explanations for these patterns, since they are fundamentally designed to optimize predictive accuracy (Shrestha et al., 2021). In other words, they do not allow researchers to consider "what-if" analysis—exploring multiple alternative interactions between variables in a phenomenon—which is at the core of abductive reasoning (Behfar & Okhuysen,

2018; Kistruck & Slade Shantz, 2022). Promising recent developments in causal inference and ML may allow researchers to overcome this limitation.

Despite these limitations, pioneering research leveraging ML algorithms to perform advanced pattern recognition demonstrated the value of identifying complex patterns in quantitative datasets, supporting the inductive generation of theoretical insights and overcoming some constraints of linear models (Choudhury, Allen, & Endres, 2021; Shrestha et al., 2021). These early approaches made use of the non-parametric nature of many ML models, allowing them to capture non-linearities and non-additivity within the mechanisms of interest (Shrestha et al., 2021). Non-additivity refers to situations in which the combined effect of two or more independent variables on the dependent variable is not equal to the sum of their individual effects. This is typically operationalized via interaction terms in moderated regression models, which rely on ordinary least squares (OLS). Such analysis is often restricted to a small subset of interaction terms, since the number of all possible interaction terms can quickly approach the number of samples in the dataset (Aguinis & Gottfredson, 2010). Moreover, only a few covariates tend to influence an outcome in a model. When the number of covariates is large relative to the number of observations, but only a small subset of covariates has a significant effect on the dependent variable, this scenario is commonly known as a "sparse high-dimensional setting" (Ren & Zhou, 2024). Using OLS in such settings is problematic for three reasons: The model may overfit the data, leading to poor generalization; multicollinearity may inflate the variance of coefficient estimates, making them highly sensitive to small changes in the data; and the sparsity of data in high dimensions may complicate the distinction between meaningful patterns and noise (Cortina 1993, Tibshiranit 1996). To ameliorate the situation, scholars often rely on an ex-ante selection of interactions terms, which implies that only a subset of all potentially significant interaction terms can be considered (Aguinis & Gottfredson, 2010).

Causal ML refers to the integration of ML techniques with causal inference methods. Unlike traditional ML, these methods aid in the identification and estimation of cause-and-effect relationships in data by means of conditioning on a set of control variables (Chernozhukov, Hansen, Kallus, Spindler, & Syrgkanis, 2025). Although conditioning on a set of control variables necessitates strong assumptions about the absence of unmeasured confounders, it remains an essential methodological approach for organizational scholars, given their common reliance on observational data. A particular benefit of applying ML methods (e.g., LASSO[1], decision trees) compared to traditional regression models (e.g., OLS) when conditioning on control variables is their scalability in modeling effect heterogeneity, i.e., the estimation of how causal effects vary across individuals or subgroups within a population. We make use of this attribute of causal ML to algorithmically explore previously unknown heterogeneity within organizational phenomena[2] following an abductive reasoning approach. We do so by allowing the model to incorporate all combinatorically possible interaction terms[3] and then assessing their significance values, using a method known as "doubly robust estimation". This approach uses orthogonalization[4] to effectively reduce bias and improve estimation accuracy (Chernozhukov et al., 2018) alongside causal sensitivity analysis to assess the robustness of the results (Bach et al., 2024a; Cinelli & Hazlett, 2020).

Our proposed algorithm-supported abduction process extends ML methods in organization science in two directions (Choudhury et al., 2021; Shrestha et al., 2021). First, we

[1] Least Absolute Shrinkage and Selection Operator

[2] Our proposed theory elaboration workflow is particularly suitable for well-documented organizational phenomena, as establishing causal claims relies heavily on the accuracy of the assumptions underlying the specified causal structure. For less-explored organizational phenomena, accurately specifying this causal structure is likely to be more challenging.

[3] While the term "interaction" is generally used to describe statistical relationships based on associations, the term "effect modifier" is often used to refer to an interaction that is of a causal nature. However, for simplicity, we remain with the more general term "interaction" throughout the paper

[4] Orthogonalization is well known to organization scholars in the context of dealing with collinearity (Barreto, 2012; Liang et al., 2019)

introduce a statistical framework called "double machine learning" (DML) to organization science, which allows for the reduction of bias introduced by the regularization procedure of ML models (Chernozhukov et al., 2018; Vanneste & Gulati, 2022). DML uses two distinct ML models on the same dataset to separately estimate the parts of the cause and the parts of the effect that are influenced by confounding factors (hence the name "double" ML). The differences (i.e., residuals) between actual and estimation values for each model are thus free from the influence of confounding factors. These residuals are then used to estimate the direct causal effect. Second, we complement DML with a graphical approach for control variable selection based on the fast-growing literature on a causal inference framework developed in computer science (Durand & Vaara, 2009; Ellsaesser, Tsang, & Runde, 2014; Pearl, 2009). Graphical models encode assumptions regarding the data-generating process, and we demonstrate how to systematically prevent estimation bias through the introduction of "bad controls" (i.e., mediators or colliders) (Hünermund, Louw, & Rönkkö, 2024). Our two-pronged approach allows researchers to consider a significantly larger number of interaction terms in their models as potential sources of complexity by eliminating overfitting and effectively handling multicollinearity while systematically avoiding bias introduced by bad controls. While the proposed algorithm-supported abduction process is neither designed to replace a researcher's intuition nor to automate theory elaboration, we argue that robust inclusion of all possible interaction terms (instead of a pre-selected subset) can uncover complex patterns in empirical data that might otherwise go unnoticed.

To showcase the value of our algorithm-supported abduction process, we draw on the example of the well-documented gender wage gap—the persistent disparity between women and men's earnings. Given the substantial implications for management research, practice, and policy making, the topic has received significant attention in the management field (Belliveau, 2012; Blevins, Sauerwald, Hoobler, & Robertson, 2019; Castilla, 2015; Cowgill, Agan, & Gee,

2024). Understanding and addressing the gender wage gap enhances an organization's reputation for fairness and inclusivity, thereby attracting top talent and potentially improving financial performance through greater diversity (Abraham, 2017; Blau & Kahn, 2007; Sitzmann & Campbell, 2021). Human capital theory (HCT) has traditionally explained some of the earning disparities by pointing to variations in men and women's investments in human capital via wage regressions (Blau & Kahn, 2017). We explore and estimate the significance and magnitude of interaction terms to grasp the moderating effects of a large set of covariates. Our algorithm-supported abduction process thereby unveils subtle interactions where returns on human capital investments differ based on an individual's gender and socio-economic context (e.g., wealth or parental education levels) and suggests that a more nuanced understanding of human capital dynamics might support pay equity efforts within organizations. While these insights are primarily aimed at demonstrating the utility of our workflow, they contribute to the broader discussion of pay equity as a key managerial concern.

This paper is motivated by the potential of causal ML techniques to enhance theory elaboration, coupled with the need to address common challenges and limitations associated with their application. While the foundations of these methodologies (i.e., causal identification based on directed acyclic graphs (DAGs)) have been introduced in organization research (Durand & Vaara, 2009), unlike adjacent fields, such as marketing (Overgoor, Chica, Rand, & Weishampel, 2019; Ren & Zhou, 2024), organizational scholars have yet to fully leverage their potential. Our proposed systematic workflow builds on these existing foundations, offering a structured approach to help organizational researchers navigate challenges and generate meaningful value from these advanced methods. In the following sections, we provide a short review of ML-based methods for theorizing in organization science and provide an accessible introduction to the foundational elements necessary for robust application of causal ML for

theory elaboration. To the best of our knowledge, an actionable description of causal ML for theory elaboration has not yet been covered by organization scholarship.

CAUSAL DIAGRAMS AND DOUBLE MACHINE LEARNING

Machine Learning Methods in Organization Science

Table 1 provides an overview and categorization of prior studies demonstrating the use of ML to build and elaborate organizational theory. Note that researchers have also applied combinations of the various ML approaches listed in Table 1 (Choudhury, Wang, Carlson, & Khanna, 2019). Generative artificial intelligence, a class of ML-based technologies designed to generate text, image, audio, and video content that resembles human-created output, is not considered, because it draws on large language models and other models that are not built specifically for the purpose of analyzing a dataset collected by the researcher (Grimes, Von Krogh, Feuerriegel, Rink, & Gruber, 2023).

Insert Table 1 about here

The ways in which ML methods support abductive reasoning and advance theory elaboration in organization science can be categorized into three groups. The first group is concerned with applying ML methods to operationalize constructs. Hannigan et al. (2019), for example, provide an extensive review on how ML-based topic modeling enables the identification of latent patterns in unstructured data, facilitating the creation of new variables from text. More recently, Luo et al. (2024) have demonstrated how other data modalities, such as images, video, and audio, can be used for *t* variable construction using ML. While such methods provide researchers with a scalable approach to incorporate novel data sources, they might produce outputs that lack interpretability, making it difficult to align them with established theoretical constructs (Choi et al., 2021). Chang et al. (2009) show that topic

models—an unsupervised ML technique that identifies latent themes or topics within a collection of text documents by analyzing word co-occurrence patterns—frequently produce topics that humans perceive as less semantically meaningful. Moreover, despite extensive progress in the field, the results of such approaches remain sensitive to model specifications and parameter choices, which can lead to instable results (Grimmer & Stewart, 2013).

The second group is exploratory data analysis, for which ML techniques such as random forests and LASSO (least absolute shrinkage and selection operator) regression have proven effective for identifying nonlinear and high-dimensional relationships between independent and dependent variables (Choudhury et al., 2021; Shrestha et al., 2021). A key contribution of these studies lies in sensitizing organizational scholars to the tradeoff between prediction performance and interpretability. In the process of applying ML methods for exploratory data analysis, researchers will necessarily face decreasing interpretability when trying to maximize predictive performance, since more complex model architectures (e.g., neural network) provide fewer prediction errors at the expense of interpretability (Shrestha et al., 2021). In the absence of further assumptions (e.g., instrumental variables), interpretability in this context always refers to purely correlational relationships. Therefore, while these methods are useful in settings with large datasets where traditional techniques may not scale well, the focus on prediction (rather than causality) necessitates further analysis (e.g., experiments) (Shrestha et al., 2021).

Unlike the first two categories of applying ML for theorizing, the third group applies methods like ML-based propensity score matching, DML and causal forests that provide a framework for conditioning on high-dimensional confounders, improving the robustness of causal estimates in observational studies, which is useful for exploring heterogeneous causal effects (Chernozhukov et al. 2018, Athey and Imbens 2019, Rathje et al. 2024). Compared to traditional conditioning models (e.g., based on OLS), a key benefit of using ML-based models

is the robust estimation of interaction effects (Feuerriegel et al., 2024). Despite their advantages, ML methods for causal effect estimation depend on strong assumptions, including correctly specifying confounders, including all relevant confounders and avoiding bad controls, which can be difficult to achieve in practice (Hünermund et al., 2024). The methodological approach introduced and illustrated in this paper aligns with this third category.

Causality and Causal Diagrams

While causality is fundamental to the process of elaborating management theory (Bamberger, 2019; Tsang, 2022), organization scholars often face considerable challenges in theoretically proposing and empirically establishing causality, since study settings are frequently constrained to observational data. Randomized controlled trials would allow for isolating and manipulating variables of interest, but are often dismissed due to feasibility, costs, or on ethical grounds. ML methods are by definition associative, meaning that by themselves, they cannot substitute for randomization in making causal inference (Shrestha et al., 2021). Under the standard assumption used with any statistical adjustment method—the exogeneity of control variables—ML methods can, however, be a powerful tool to estimate causal effects (Athey and Imbens 2019).

Endogeneity—where an explanatory variable is correlated with the error term in a regression model—arises due to, among other reasons, omitted variable bias (OVB) (Antonakis et al. 2010). OVB occurs when a model fails to include one or more relevant variables that influence both the independent and dependent variables, leading to invalid causal effect estimates. This bias arises because the effects of the omitted variables are mistakenly attributed to the included variables, distorting the estimated relationships. For our algorithm-supported abduction process, we focus exclusively on endogeneity introduced by OVB, as it is one of the most pervasive sources of endogeneity in organization studies (Antonakis et al. 2010,

Busenbark et al. 2022). Also, unlike other sources of endogeneity (e.g., sample selection bias), OVB is important for theory generation, because it highlights the need to identify and include theoretically relevant variables to accurately establish causal relationships, ensuring that the theory reflects the true underlying mechanisms.

The causal inference framework developed in computer science by Pearl (2009) is a landmark in the study of causality, providing a rigorous mathematical language for systematically handling endogeneity. At its core are graphical causal diagrams, known as DAGs. A DAG is a visual representation of the researchers' assumptions regarding the causal relationships between variables, depicted as nodes and connected by directed edges (i.e., arrows) that indicate the direction of causation. Each edge has a non-cyclical direction, which means that no path leads back to the same node it started from (Durand & Vaara, 2009). A DAG thereby enables the accurate identification of causal effects via conditioning on confounders, which are the primary source of endogeneity issues related to OVB (Hünermund & Bareinboim, 2023). An example of a simple DAG is provided in Figure 1: When studying the impact (i.e., causal effect) of R&D expenditure on firm performance, the observed relationship might be confounded by market conditions and cultural aspects of the firm. We represent these assumptions in the DAG by adding the appropriate nodes (i.e., "Market" and "Culture") and respective edges (i.e., arrows). There are two further assumptions implicitly encoded in this DAG. First, we do not allow for cycles[5] (e.g., there is no arrow from "Firm Performance" to "Culture"), since this would lead to an ambiguous specification of the causal effect (it would become unclear whether "Culture" causes "Firm Performance" or vice versa). Second, we assume that no other confounding factors exist between the relationship of interest.

[5] Note that we only rule out instantaneous feedback loops, which do not include dynamic cyclic relationships over time.

Insert Figure 1 about here

Within the causal inference framework, DAGs support "causal identification"—the process of determining whether a causal effect can be estimated from observed data, given the assumptions encoded in the causal model. Causal identification relies on the so-called "backdoor criterion," which requires controlling for all confounding variables between the treatment and the outcome (Pearl, 2009). In the example provided in Figure 1, the backdoor variables are "Market" and "Culture." It is important to distinguish this step from the estimation step. In other words, DAGs are agnostic to the statistical model used for estimation; they can be applied with OLS as well as more complex ML modeling approaches (e.g., DML). Note that in the presence of unobserved confounding variables, the causal inference framework also covers causal identification via other identification strategies such as instrumental variables (see Durand & Vaara, 2009; Hünermund et al., 2024). However, due to our focus on OVB, we do not explicitly cover these.

Double Machine Learning

Scholars often encounter high-dimensional settings when performing regression analysis. In macro research, this is a common situation, because many covariates may be deployed to characterize a rather limited set of firms (Bloom, Eifert, Mahajan, McKenzie, & Roberts, 2013). In micro research, high dimensionality typically arises from the encoding of categorical variables (i.e., dummy variables). An additional source of high dimensionality is the incorporation of interaction terms (Aguinis & Gottfredson, 2010). When using a moderation model, the number of predictors in the multiple regression rises exponentially. For example, a regression model with four independent variables has six interaction terms whereas a model with five independent variables has 10.

Penalizing, or regularizing, model complexity[6] is a common approach in ML algorithms to prevent a model operating in a high-dimensional setting from capturing noise rather than signal. One common algorithm for regularization is LASSO. This regression method minimizes prediction error by penalizing the sum of the absolute values of the regression coefficients, encouraging sparsity by shrinking some coefficients to zero and effectively performing variable selection (Tibshiranit, 1996). In the context of statistical modeling and ML, sparsity implies that only a small subset of the covariates has non-zero coefficients, indicating that they are the most relevant or influential variables, while the rest are deemed irrelevant and effectively excluded from the model. These techniques are known to work well in predictive settings where the coefficient values of each independent variable are not of interest, but the overall predictive power of the model is (He et al., 2020). When performing regression analysis in causal estimation, however, the regularization in methods like LASSO might severely bias causal estimates because of a "shrinkage of coefficients," where a variable important for controlling confounding but weakly correlated with the outcome is excluded from the model. By these means, regularization might lead to inconsistent estimation of the treatment effect (Ban, El Karoui, & Lim, 2018).

DML has been designed to deal with high-dimensional settings and variable selection in causal inference (Chernozhukov et al., 2018). It is designed to counter regularization bias and provide estimation errors that are asymptotically normally distributed. First, DML employs orthogonalization to ensure that regularization does not introduce bias. Orthogonalization is well known to organization scholars in the context of dealing with collinearity (Barreto, 2012; Liang, Marquis, Renneboog, & Sun, 2019), but in our context, it involves isolating the causal

[6] For a more extensive explanation of the bias-variance tradeoff (i.e., the U-shaped relationship between model complexity and prediction error), we refer to Shrestha et al. (2021).

effect of the treatment on the outcome by removing the influence of confounding covariates (akin to regressing out irrelevant influences).

Second, DML uses high-quality ML models to estimate the relationships between the outcome (dependent variable), the treatment (independent variable of interest), and the covariates. In the context of DML, quality is measured in terms of "convergence rates," which relate to the speed at which an estimator approaches the true parameter value as the sample size increases. Higher convergence rates enable the model to retain asymptotic normality, i.e., the error converges to a normal distribution, which in turn provides statistically valid *p* values and confidence intervals. In general, ML algorithms such as LASSO have high convergence rates, because the penalty shrinks many parameter estimates to zero, effectively enforcing a sparse model[7]. The DML framework is flexible in terms of the ML algorithm. It can be applied with LASSO, decision trees, neural networks, and others, as long as the convergence rate is sufficiently high.

Third, DML incorporates "cross-fitting," which divides the dataset into multiple folds, training the ML models on one subset of the data while using another subset to estimate residuals. Cross-fitting prevents the overlap of training and evaluation data, thereby reducing overfitting and enhancing the robustness of the causal estimates. By combining these steps, DML addresses common pitfalls of ML methods for advanced pattern discovery in exploratory analysis (Choudhury et al., 2021; Shrestha et al., 2021).

To the best of our knowledge, there are no existing publications within the field of organizational science that explicitly introduce or address the concept of regularization bias.

[7] Note that the regularization previously recognized as a potential source of bias becomes advantageous in this context. This shift arises due to the applied orthogonalization procedure, which ensures that any bias introduced in estimating control variables does not impact the final estimation of the treatment effect.

We attribute this to the fact that ML methods for causal effect estimation are not yet broadly adopted this literature. Leveraging the power of ML for causal effect estimation, however, requires a careful consideration of regularization bias, which can fundamentally alter parameter estimates when regularization techniques are employed. However, regularization bias is not the only form of bias that researchers need to be aware of when applying causal ML. In the next section, we outline the relationship between covariate selection and causal estimation bias.

Combining DML with Graphical Causal Models

While DML offers a strong basis for causal analysis with many covariates, it assumes that all included covariates are exogenous, meaning they do not introduce bias when conditioning on them. This is problematic if the researchers using DML include so called "bad controls" (Hünermund et al. 2025). A bad control is a variable that is affected by the treatment—lying on the causal pathway between the cause and the outcome—such that including it in the analysis inadvertently removes part of the cause's effect, thereby biasing the estimation of the direct causal relationship. Additionally, a bad control can distort the estimation process by inducing spurious associations through unmeasured confounders that influence the bad control (i.e., mediator), creating new, artificial paths of correlation between the independent variable and the outcome. As previous articles have demonstrated, DML is particularly vulnerable to the inclusions of bad controls (Hünermund, Louw, & Caspi, 2023), thus a principal task for researchers applying DML is to avoid the inclusion of endogenous variables. One efficient and scalable approach for this task is to use graphical models such as DAGs (Hünermund et al., 2024). Below, we refer to the combination of DML and DAGs for robust estimation of significant interaction terms as "graph-based double machine learning."

APPLYING GRAPH-BASED DOUBLE MACHINE LEARNING

To enable the robust estimation of many interaction terms using graph-based DML, we propose an algorithm-supported abduction process as shown in Figure 2. This workflow is not designed to automate theorizing but instead complements the process of theory elaboration through abductive reasoning by allowing researchers to explore additional sources of heterogeneity in their models. Figure 2 captures the basic principle that abductive reasoning aims to generate alternative explanations for puzzling aspects of phenomena and assess their plausibility (Bamberger, 2019). Graph-based DML algorithmically supports the generation of such plausible explanations by providing robust significance values for all possible interaction terms of independent variables considered in the model.

The graph-based DML workflow is iterative. For example, if a researcher fails to collect sufficient data on the covariates included in the DAG, it might be necessary to go back to re-formulate the research question to obtain a valid result.

Insert Figure 2 about here

The proposed workflow consists of seven key steps designed to systematically guide the application of graph-based DML. Before starting with Step 1, it is crucial to narrow the scope of inquiry to a specific organizational phenomenon and establish a related theoretical framework to be elaborated. In Step 1, researchers encode the existing state of the literature with respect to causal relationships into a DAG. In Step 2, relevant data is collected and pre-processed, including, for example, the construction of dummy variables for categorical measures. Step 3 involves selecting appropriate ML estimators—such as LASSO regression—to separate the effect of the variable of interest from those of the control variables. In Step 4, interaction terms are specified based on their relevance in the context of the investigated phenomenon and target theory. In Step 5, the treatment and outcome models are trained, which

involves parameter tuning and cross validation to ensure robust and accurate estimation of treatment effects. Step 6 uses sensitivity analysis to evaluate the robustness of the causal estimates by examining the potential impact of unobserved confounders, ensuring confidence in the derived conclusions. Step 7 involves selecting the interaction terms that robustly appear significant. To demonstrate the value of the proposed workflow model, below, we provide a step-by-step walk through based on a real-world dataset, studying the gender-wage gap in the US.

Identify Phenomenon and Target Theory

The workflow begins by narrowing the scope of inquiry to a specific organizational phenomenon and identifying a related theory for elaboration. For demonstration purposes, we examine the phenomenon that, despite similar levels of experience and education, women earn less than men, a disparity known as the "gender wage gap." The unadjusted gender wage gap is often measured as the ratio of median earnings of women to men, expressed as a percentage value. The wage gap example is relevant for two reasons. First, it represents a significant social issue that compels managers in most organizations to seek solutions. Second, it engages a broad scholarly audience, as HCT, the dominant framework for examining the wage gap, is well established and informs many debates in organization science.

Despite global efforts to achieve gender parity, significant gaps remain. For instance, the global gender wage gap stood at around 68% in 2023, indicating that women earn approximately 68 cents for every dollar earned by men (World Economic Forum, 2023). The gender wage gap varies by country, industry, and individual career stage, reflecting complex socio-economic dynamics. It highlights systemic inequalities in the labor market, influenced by factors such as occupational segregation, differences in work experience, educational attainment, and discrimination (Blau & Kahn, 2017).

Whereas there are multiple theories that may explain the gender wage gap, HCT has been adopted most widely (Blau & Kahn, 2017; Goldin, 2014). HCT posits that individuals increase their productivity and earnings potential through investment in education, training, and experience (Becker, 1964; Schultz, 1961). This investment reflects the expectation of future returns in the form of higher wages. On the individual level, investing in higher education is assumed to increase lifetime earnings, improve access to higher paid jobs, and reduce time in unemployment (Becker, 1993). A large body of empirical research has shown that, overall, higher education represents an attractive investment opportunity yielding between 7% and 19% in private returns (Blöndal, Field, & Girouard, 2003).

Previous studies have attributed differences in wages between female and male employees to variations in their human capital (Blau & Kahn, 2017). A paramount concern in these studies has been disentangling the effects of discrimination from those attributed to human capital investments. Decomposition methods, based on OLS regression, allow researchers to quantify the extent to which differences in wages (such as those between genders) could be attributed to observable (i.e., "explained") characteristics, such as education and experience, and to unobservable (i.e., "unexplained") factors, potentially indicative of discrimination (Blau & Kahn, 2017). Applying the decomposition methods to the gender wage gap, researchers have found that the portion of the gap explained by human capital attainment shrunk from about 28% in 1980 to below 15% in 2010 (Blau & Kahn, 2017). Yet, a significant part of the gap remains unexplained, suggesting the influence of other factors, including potential labor market discrimination against women.

Several studies have reported a substantial increase in women's human capital levels between 1980 and 2010—a phenomenon described as the "grand gender convergence" (Goldin, 2014). To illustrate this situation, we provide summary statistics of a common real-world dataset for analyzing the gender wage gap in the US, the Panel Study of Income

Dynamics (PSID, Table 2). Leveraging the most recent data available from this source (2021), we obtain a dataset composed of 4,658 survey responses[8], providing a representative sample of the US population. We observe that there are hardly any disparities with respect to human capital characteristics[9] between women and men in our dataset—the share of women with a Bachelor's degree is comparable to that of men, and in terms of professional experience, the average levels do not differ significantly between genders.

Insert Table 2 about here

Examining the gender wage gap through the framework of HCT in contexts where men and women exhibit parity in human capital attainment necessitates a focus on gender-based disparities in the returns on human capital investments. In terms of higher education, previous literature is inconclusive whether the college wage premium (i.e., the additional earnings that college graduates receive compared to non-college graduates) differs between men and women (Blau & Kahn, 2017). Earlier work by Dougherty (2005) suggested that returns were higher for women, but Hubbard (2011) presents evidence that this is not the case when preventing outliers from skewing the analysis via top coding. In addition to the gender dimension, several studies have investigated various sources of heterogeneity with respect to returns on higher education and found significant effects for study areas and degree types (Blöndal et al., 2003; Psacharopoulos & Patrinos, 2004; Wahrenburg Mark & Weldi Martin, 2007). Card (1999) identifies heterogeneity in educational returns based on factors such as school quality, family background, individual ability, and declining marginal returns (i.e., as individuals attain higher degrees, the additional benefits gained from each successive degree tend to decrease). While family background, especially parental education, influences returns, evidence for its

[8] PSID Family-level 2021 (incl. spouse data).
[9] Note that human capital attainment is encoded in the columns.

magnitude remains mixed. Individual ability, often measured by IQ, correlates positively with returns, though estimates vary. Moreover, a concave relationship between schooling and earnings suggests that individuals with lower education levels experience higher marginal returns, aligning with evidence that policies targeting disadvantaged populations yield significant benefits. These findings illustrate the role of individual, institutional, and socio-economic contexts in shaping educational returns (Card, 1999). Additionally, evidence indicates that individuals from wealthier backgrounds benefit from enhanced returns on educational investments due to greater access to superior resources and opportunities (Diemer, Marchand, & Mistry, 2020). Furthermore, the importance of accounting for effect heterogeneity in analyzing the gender wage gap has been demonstrated also with a focus on the methodological perspective by quantifying the extent of heterogeneity using DML (Bach, Chernozhukov, & Spindler, 2024b).

The previous literature aimed at elaborating on HCT by examining moderation effects has not explored how the investigated factors (e.g., parental education, family wealth, etc.) interact with gender, limiting its explanatory power in the context of the gender wage gap. A conclusive understanding of whether returns to education differ between women and men could provide clearer insights into gender disparities, inform equitable education policies, and guide managerial interventions to close earnings gaps. Furthermore, if disparities in returns to education between genders would be quantified alongside certain socio-economic factors[10], researchers could gain an even more elaborate understanding of human capital dynamics in relation to the gender wage gap. Hence, to demonstrate the value of algorithm-supported abduction via graph-based DML and avoid an *ex-ante* restriction to a limited set of interaction

[10] Note that this constitutes a three-way interaction term. We provide more information on three-way interaction terms in Step 4: Specify Interactions of Interest.

terms, we formulate the following research question: **To what extent do individual and socio-economic characteristics cause disparate returns on investments into human capital for women vs. men?**

In the following sections, we demonstrate that graph-based DML can handle high numbers of interaction terms and therefore provides a more granular picture of wage penalties compared to the commonly used average estimates. It is important to stress that the target coefficients, which we test with the graph-based DML process, refer to the interactions of certain covariates (e.g., college degree, family wealth) with gender. We interpret a significant interaction coefficient, validated through sensitivity analysis, as evidence of heterogeneity and therefore worthy of consideration in the theory elaboration process.

Step 1: Encode Existing Knowledge into a Directed Acyclic Graph

Based on the specified research question, the next step is to encode the assumptions regarding the data generating process into a DAG, which requires in-depth review of existing theory and domain knowledge. Within the scope of our analysis, human capital attainment is measured via a binary variable representing whether an individual has obtained a Bachelor's degree or not, which is our treatment variable. Individuals who obtained higher degrees (e.g., Master's degree) are not distinguished from the individuals who obtained only a Bachelor's degree. The outcome variable is represented by the dollar amount earned per hour worked and therefore implicitly controls for various levels of employment[11].

[11] Note that we do not consider the costs of obtaining a college degree (incl. opportunity costs), as done in other studies (e.g., Blöndal et al., 2003) but instead define the returns on human capital investment simply as the percentage increase in hourly salary after obtaining a college degree. Therefore, we anticipate obtaining higher percentage returns compared to other studies.

Next, we consider potential confounding factors between the treatment and the outcome based on the available literature. This reflection is crucial for addressing endogeneity without being constrained by available datasets. Based on findings from previous studies, we formulate the DAG to incorporate the family background of an individual confounding the treatment and the outcome variable (Angrist & Pischke, 2009; Kunze, 2008). The individual's educational attainment is significantly influenced by the educational level of their parents, a relationship that can be understood through the lens of social capital theory (Coleman, 1988). Parents with higher educational backgrounds typically have greater access to professional networks and resources, which they can leverage to support their children's educational and career aspirations. Such access facilitates opportunities for mentorship, internships, and job placements, perpetuating a cycle of educational and professional advantage. This also affects the opportunity to learn, which captures the conditions, resources, and settings that enable or constrain educational experiences. Parental education levels influence children's educational outcomes by providing enriched learning environments, support, and expectations that foster academic success (Perry et al., 2024). We represent these insights by including four confounders into our DAG: the educational attainment of the mother, the educational attainment of the father, the cumulated family wealth (Gould et al., 2020), and a binary variable representing whether the individual was born in the United States or not, since there are persistent earnings differences and human capital disparities between migrants and non-migrants (Chiswick, 1978). The model is provided in Figure 3.

In addition to the family background, we account for regional differences in higher education (e.g., rural vs. metropolitan areas) and the fact that individuals in economically stronger regions receive higher salaries. We introduce the geographical region as an additional confounder. Also, as common in the analysis of the gender wage gap, we include ethnicity as a confounder potentially affecting both the treatment and the outcome variable (Blau & Kahn,

2017). Finally, we include gender as a confounder between obtaining a college degree and the hourly salary, which enables us to represent gender-specific career and education choices as well as the discriminatory effect of paying lower salaries to women.

Additionally, we include two variables that are influenced by the abovementioned ones: childcare responsibilities and the area of study. Childcare responsibilities may affect an individual's salary, because the demands of childcare could make it challenging to advance one's career, thus impacting earning potential and ultimately leading to lower salaries (Stone & Hernandez, 2013). Our causal model, shown in Figure 3, encodes the assumption that gender influences the amount of childcare an individual performs, since women tend to take on more childcare responsibilities[12] (Ly & Jena, 2018). In a similar fashion, the area of study (e.g., engineering, liberal arts, etc.), an important determinant of salary levels, may be affected by gender, since some study areas are male dominated, while others are female dominated (Ammermüller & Weber, 2005).

Insert Figure 3 about here

Step 2: Collect and Pre-Process Data

This important step involves gathering relevant datasets and organizing them in a clean, usable format by removing errors, filling in gaps and ensuring consistency (e.g., with respect to units). For our analysis, we leverage an open-source longitudinal household survey called the PSID[13], which provides high-quality, comprehensive data on income, employment, education, and other variables over time, covering all the necessary variables in our causal model. We exclusively use data from the most recent 2021 survey.

[12] Note that this pattern might not hold in countries outside the US.
[13] www.psidonline.isr.umich.edu

We follow standard statistical modeling practices to construct dummy variables from categorical data (e.g., degree), providing a structured dataset of 24 columns and 4,658 observations. An overview of all variables is provided in Table 3. Finally, we construct all possible two- and three-way interaction terms given the set of covariates, resulting in a data structure with 2,051 columns. For more details, please refer to the provided code repository[14].

Insert Table 3 about here

Step 3: Specify Model and Learners

As discussed above, within the DML framework, the researcher is free to choose what type of model and estimator[15] to use. To provide an accessible introduction to the algorithm-supported abduction process, we choose a partially linear model in combination with a LASSO regression algorithm for both the outcome and the treatment models. Partially linear models are a class of statistical models that blend the interpretability of parametric components with the flexibility of nonparametric methods. The parametric portion provides straightforward interpretability, akin to OLS coefficients, while the nonparametric component does not assume any specific functional form and therefore allows for capturing nonlinear relationships. Partially linear models are designed to separate the causal effect of interest from those of the control variables, which are estimated flexibly using a combination of regularization and orthogonalization. This reduces model misspecification risk while ensuring robustness to complex confounding structures.

[14] *Link to online code repository incl. the full dataset will be added.*

[15] For a more detailed introduction of various learning algorithms, we refer to Choudhury et al. (2021).

The partially linear models take the following form:

$$\text{Outcome Model: } ln(s_i) = \theta * d_i + g_0(x_i) + \zeta \quad \textit{(Equation 1)}$$

$$\text{Treatment Model: } d_i = m_0(x_i) + V \quad \textit{(Equation 2)}$$

where s_i is the hourly salary of an individual i and d_i represents the treatment variable. x_i is a vector consisting of all confounding factors, i.e., gender, education of the parents, whether an individual was born in the US, and ethnicity. In our case, $g_0(X)$ and $m_0(X)$ are linear functions that include all two-way and three-way interactions, but no polynomial terms. ζ and V are stochastic errors with $\mathbb{E}(\zeta \mid Degree,X) = 0$ and $\mathbb{E}(V \mid X) = 0$.

Since we are interested in the coefficients of many interaction terms, we treat each variable of interest in turn as d_i and apply the described estimation and inference for that specific variable. This is a standard procedure in simultaneous inference on multiple variables (Chernozhukov, Hansen, & Spindler, 2016). We include more information on multiple treatments and joint confidence interval in the online appendix.

Step 4: Specify Interactions of Interest

Given our research question and theoretical context, we specify all two-way and three-way interactions that include Gender and Degree as interactions of interest. This means that, while the model includes all possible interaction terms to allow for a high level of complexity, we focus our exploration efforts on a subset of 25 interaction terms (i.e., one two-way interaction and 24 three-way interactions). Including further interaction terms would constitute either a departure from HCT or the context of the gender wage gap since it is, for example, not of interest for us whether regional factors alone moderate the return on human capital but rather the combination of regional factors with gender. It is important to note that estimating an interaction effect is akin to conditioning on that variable. Therefore, to avoid the introduction

of endogenous variables as controls, we can only consider interactions with root nodes (see more detailed explanation in the section "Combining DML with Graphical Causal Models" above). For example, we consider the number of hours spent on childcare to be dependent on an individual's gender (Blau & Kahn, 2017). Therefore, we consider any estimate of an interaction term which includes the number of hours spent for childcare potentially biased, i.e., post-treatment bias (Hünermund et al., 2024)).

Scholars often find interpreting three-way interaction terms challenging, because such interactions require understanding how the relationship between two variables changes depending on the level of a third variable. This is especially true in models with continuous variables, where the interaction effect varies across the range of the interacting variables. When the variables are binary, interpretation becomes more manageable, because the interaction can be evaluated at specific combinations of the variables (e.g., all possible 0-1 configurations). Scholars can examine the coefficients of the three-way interaction to determine how the effect of one variable on the outcome changes depending on the combination of the other two variables. For this reason, all three-way interactions are constructed from binary variables. For a more comprehensive discussion on three-way interactions, including a typology providing a structure approach to understanding and interpreting these interactions in management research, we refer to Lam, Chuang, Wong, & Zhu (2019).

Step 5: Fit Treatment and Outcome Model

Training our ML models requires the specification of parameter values, known as "hyperparameters," on two levels: Learner-specific hyperparameters and DML parameters. First, depending on the choice of learning algorithms for the treatment and outcome models, we must specify learner-specific parameters. In our example working with LASSO, the key parameter is the magnitude of the penalty, which we obtain via hyperparameter tuning

(Tibshiranit, 1996). The penalty parameter in LASSO regression controls the strength of the regularization, balancing the trade-off between minimizing the residual sum of squares and shrinking coefficients towards zero, thus influencing variable selection and model complexity.

Second, DML-specific parameters include the number of folds for cross-fitting and the number of repetitions. These parameters govern the cross-fitting procedure and are critical to reduce overfitting. The number of folds specifies the number of data splits for cross-fitting, where each fold is used alternately for testing while the remaining folds train ML models; higher values improve generalization but increase computational cost. The number of repetitions defines the number of times the cross-fitting procedure is repeated with different random splits, averaging the causal estimates to enhance robustness. Together, these parameters balance computational efficiency and the precision of causal effect estimates, with typical values being 5-10 for the number of folds and 1-10 for the number of repetitions, depending on the dataset size and available resources. For more details, please refer to the code repository[16] and associated commentary.

The model outcome after training (i.e., fitting) is presented in Table 4. We report the estimate for each interaction term of interest, as well as corresponding *p* values and joint confidence intervals at the 99.9% level. We include a detailed description of joint confidence intervals in the online appendix. The estimated coefficient of discrete regressors can be understood as changes compared to the baseline group. The baseline group for this regression is white male without a college degree (see Table 3). This, in combination with the fact that we use log salaries, means we can interpret the estimates as percentage values. The estimates in Table 4 show that obtaining a Bachelor's degree leads to an approximate 45% increase in salary for men. Table 4 also provides an estimation of the gender wage gap for individuals without

[16] *Link to online code repository incl. the full dataset will be added.*

Bachelor's degrees: Being female leads to an approximate salary penalty of 24% (i.e., women earn about 76% of what men earn). The estimation results in Table 4 are obtained by controlling for all root nodes in Figure 3.

Insert Table 4 about here

The heterogeneity of the gender wage gap can be quantified based on the results presented in Table 4. For instance, Table 4 shows the coefficient for the interaction term between obtaining a degree and being female. This result indicates that women experience an estimated 18% higher wage when obtaining a Bachelor's degree compared to men. In other words, women obtaining a Bachelor's degree experience a wage gap that is *ceteris paribus* about 18% smaller than that of women who have not obtained a degree[17]. We provide a detailed explanation in the online appendix.

The results demonstrate that women earn less than men but obtaining a Bachelor's degree helps them to close the gap. To understand what other factors contribute to closing the gap, Table 4. provides the parameter values for three-way interaction terms between various socio-economic factors and *Degree*Female*. Women in the highest quartile of family wealth experience a wage gap that is *ceteris paribus* about 35% smaller than women with a Bachelor's degree from the third family wealth quartile[18]. Women living in north-eastern US states obtaining a Bachelor's degree seem to enjoy a reduced wage gap of about 33% compared to women with a Bachelor's degree living in Northwestern states[18]. Finally, women with more educated parents who obtain a Bachelor's degree appear to experience lower wage gaps—between approximately 20% and 24%—compared to women with less educated parents[18] who obtain a Bachelor's degree. Note that Table 4 only provides the estimation values for the

[17] The intercept of the regression corresponds to the gender wage gap for the baseline group.
[18] See Table 3 for a definition of the baseline group.

interaction effect with p values of 0.001. A full list of all estimation results is provided in the online appendix.

Step 6: Perform Sensitivity Analysis

The obtained *p* values of our estimates do not provide means for confirmation or rejection of the assumptions made in the causal model (Bliese, Certo, Smith, Wang, & Gruber, 2024). An unobserved confounder between the treatment and the outcome could still lead to a significant effect and potentially a wrong conclusion. Since we are not able to perform a randomized control trial and the ground truth is missing, the accuracy of our estimates relies on the assumptions reflected in the causal model (see Figure 3). Despite the lack of a ground truth, we can use sensitivity analysis to evaluate the robustness of the estimated treatment effects against deliberate violations of different assumptions (Bach et al., 2024a). A comprehensive overview of the available methods for sensitivity analysis lies beyond the scope of this paper[19], and here we restrict our descriptions to a recent approach that deals with the assumption of unconfoundedness. As mentioned before, our estimates assume that we control for all confounding factors between *Degree* and *Salary*. If this unconfoundedness assumption is not met, see Figure 4, sensitivity analysis can be used to quantify the impact of such unobserved confounding. This can be achieved by means of sensitivity parameters measuring the strength of confounding in both the treatment assignment mechanism and the outcome equation (Bach et al., 2024a). The OVB discussed above can thus be represented in a mathematical form, which allows us to assess to how much bias any confounding scenario (as given by specific values of sensitivity parameters) would correspond. If a researcher does not have a particular confounding scenario in mind, we recommend reporting so called robustness

[19] More information can be found in Hünermund et al. (2024) and Lonati & Wulff (2024)

values (RV), which indicate the minimal strength that (a collection of) confounders would need to have to revert the causal effect (i.e., change the sign of the effect) (Bach et al., 2024a). In other words, the robustness value answers the following question: "How strong would an unobserved confounder have to be—relative to the observed controls—to explain away the observed treatment effect?" The estimation of robustness values can be performed using DML (Chernozhukov, Cinelli, Newey, Sharma, & Syrgkanis, 2021).

Insert Figure 4 about here

The obtained RVs are reported alongside the estimates, in Table 4. For example, Table 4 shows that a possible confounder would need to explain about 31% of the residual variation of both the treatment and the outcome to be capable of reversing our conclusion. The residual variation of the outcome refers to the remaining variability in an individual's salary that cannot be explained by the covariates (control variables) included in the model, whereas the residual variation of the treatment refers to the percentage of the chances of obtaining a Bachelor's degree that cannot be explained by the covariates. After the main causal relationships have been accounted for, this residual variation captures what is left unexplained and serves as a critical factor in evaluating model robustness. The higher the RVs, the more robust the estimate.

Step 7: Select Robust Interactions

The final step of the workflow is comprised of selecting interaction terms based on the results obtained by the previous step. To ensure rigor, we adopt a two-stage process to identify robust interaction terms within our theoretical framework. First, we discard all interaction effects that yield a p value larger than 0.001. The choice of this comparatively high confidence level is justified by the multiple testing approach (see online appendix). In this way, we put relatively more weight on avoiding false positives, which is appropriate in an exploratory study

design. Second, we use the robustness values obtained from sensitivity analysis and use these to rank our results (see Table 4). Estimates with a robustness value of around 5% and above can be viewed as robust, since it is unlikely that an omitted variable exists which can explain over 5% of the residual variation of the outcome (Bach et al., 2024a). To put this threshold into perspective, we compute the partial *R* squared for the variable gender on salary, yielding 0.023. This indicates that, after controlling for all other covariates in our model, approximately 2.3% of the residual variance in salary is uniquely attributable to gender. Given that gender is a significant confounder, and its partial *R* squared value falls below half of our predefined cutoff threshold, the validity of our results would only be challenged if an omitted variable exerted a confounding effect more than twice as strong as that of gender.

Given the described selection logic, we discard the interaction estimate for ethnic Asians (robustness value of 0.82, see Table 4). It is important to recognize that while the robustness value is low, the coefficient value is comparatively high. In this way, we demonstrate that there is no one-to-one relationship between effect size and robustness to unobserved confounding. In total, our algorithm-supported abduction method has identified five out of 25 interaction terms that are significant and robust.

Note that at this stage, we do not yet make any claims regarding the novelty of the selected interaction terms. This task, covered in the next section, cannot be supported by the algorithmic system and hence relies on the researcher's expertise and knowledge of the literature.

IMPLICATIONS FOR THEORY AND PRACTICE

The interpretation of the estimation results obtained from graph-based DML and the formulation of potential implications for the target theory require (collective) human

sensemaking and a deep understanding of the existing literature. The purpose of this step in our algorithm-supported abduction process is to specify clear contributions to the existing body of knowledge based on the findings obtained. We draw from these findings to formulate plausible explanations for the phenomenon under investigation. However, the findings are not guaranteed to be novel.

Our findings from the previous section are:

- *Finding 1*: Women obtain on average an approx. 18 percentage points higher return on obtaining a Bachelor's degree compared to men[20] (see Table 4).
- *Finding 2*: Women with a high level of family wealth who obtain a Bachelor's degree experience a wage gap that is approx. 35% lower compared to that of women who obtain a Bachelor's degree with a lower level of family wealth[21] (see Table 4).
- *Finding 3*: Women living in Northeastern states who obtain a Bachelor's degree experience a wage gap that is approx. 32% lower compared to women who obtain a Bachelor's degree living in the Northwestern states[21] (see Table 4).
- *Finding 4*: Women with highly educated parents who obtain a Bachelor's degree experience a wage gap that is between 20-24% lower than that of women who obtain a Bachelor's degree with lower-educated parents[21] (see Table 4).

Higher relative returns on human capital investments for women compared to men (*Finding 1*) enables women who obtain a bachelor's degree to reduce the gender wage gap. In other words, Women obtaining a Bachelor's degree experience a wage gap that is approx. 18% lower than that of women without a Bachelor's degree. This empirical finding contributes to elaborating HCT, specifically to the literature on college wage premiums which finds

[20] Note that this result does not translate to higher absolute returns for women
[21] The baseline categories are defined in Table 3.

inconclusive results in terms of gender-related disparities (Card, 1999; Dougherty, 2005; Hubbard, 2011). Previous studies have outlined that schooling influences female earnings through both a direct productivity effect and an indirect effect by mitigating the negative impact of discrimination. The literature suggests that the relationship between education and discrimination is inverse, as higher education levels increase the likelihood of obtaining standardized wage offers through formal qualifications and enhance the capacity to resist discriminatory practices. Additionally, obtaining higher education may increase women's propensity to seek employment beyond traditionally low-paying female-dominated occupations.

Expanding on our first finding, *Findings 2-4* offer an even more nuanced understanding of what individual characteristics affect returns obtained from education for women. First, our results signify that wealth provides a critical buffer against the constraints that limit career advancement and earnings potential for women. Individuals with higher wealth are less dependent on immediate income, enabling them to take on roles or work in sectors with higher long-term returns (Diemer et al., 2020). Wealth also grants access to premium educational opportunities, career coaching, and professional networks, further amplifying the impact of their human capital. Additionally, women with a higher level of family wealth may be more likely to afford childcare and household help, reducing career interruptions and enhancing their ability to focus on professional growth. While it is well known that wealth shapes return on investments in education (Pfeffer & Schoeni, 2016), our results demonstrate how this mechanism affects equitable pay outcomes specifically for women.

Second, regional labor market dynamics seem to counteract some of the disadvantages women face in terms of returns on education. Higher returns on education for women in the

Northeastern United States, compared with the Northwestern region[22], might stem from the region's industrial composition—featuring finance and professional services—which often rewards additional schooling more generously than the resource-based and manufacturing-oriented sectors of the Northwest (Moretti, 2012). The region's high-caliber universities and dense professional networks amplify these educational payoffs (Dale & Krueger, 2014). While this insight is not novel across genders, our analysis unveils the magnitude of this effect specifically for women.

Third, highly educated parents tend to pass on higher levels of social and cultural capital, which translate into better access to professional networks and mentorship opportunities. These advantages may help their offspring leverage their educational achievements to navigate structural barriers and secure higher-paying jobs. Moreover, educated parents often foster aspirations and confidence in their offspring, encouraging them to negotiate for better compensation. While this mechanism is, again, well documented across genders, our findings suggests that intergenerational transfers of privilege mitigate, but do not eliminate, the systemic disadvantages faced by women in the labor market (Breen & Jonsson, 2005).

While there may be alternative explanations for these findings based on constructs not included in our analysis, such as differences in negotiation behaviors (Kray, Kennedy, & Lee, 2024), abductive reasoning in theory elaboration compels us to prioritize the most plausible and empirically supported explanations. This approach allows for a focused analysis that builds on existing evidence while remaining open to further exploration of less apparent factors. However, it is important to reiterate that the workflow, see Figure 2, is designed to be iterative.

[22] As outlined in Table 3, the baseline region is northwest.

Should the selected interaction terms not seem plausible to researchers, the problem setting needs to be reframed.

DISCUSSION

The presented findings contribute to a more nuanced understanding of HCT in the context of the gender wage gap, showing that investments in higher education yield heterogenous returns across individuals. Specifically, the theory should account for the moderating role of family wealth, regional job market dynamics and parental education levels. This extends HCT by highlighting individual characteristics that enable or constrain women's ability to realize the full value of their human capital investments (Dale & Krueger, 2014; Moretti, 2012; Pfeffer & Schoeni, 2016). By unveiling how previously examined mechanisms are moderated by gender, the theory can better capture the intricate dynamics that constitute the phenomenon of the gender wage gap.

From a managerial perspective, closing the gender wage gap requires addressing the structural factors that impact the returns on human capital investment for women. Organizations should implement targeted policies to reduce biases in hiring, promotions, and pay-setting processes, ensuring that women's qualifications and experience are equitably rewarded. Additionally, managers can invest in mentorship programs and leadership development opportunities, particularly for women from less advantaged backgrounds, to bridge the gap in social capital. Providing resources such as childcare support and flexible work arrangements can further enable women, especially those with lower family wealth, to fully leverage their human capital. By adopting these practices, organizations not only promote equity but also enhance their ability to attract and retain diverse talent.

The present paper aims at providing organizational scholars with an accessible demonstration of how to support abductive theorizing by leveraging graph-based DML to perform robust estimation of interaction effects. Considering effect heterogeneity in an abductive manner marks a significant departure from the results obtained through traditional regression methods, which require a pre-selection of a subset of interaction terms. Our methodology contributes to the existing methodological literature in organization science focused on ML methods. While previous contributions focused on the theory building process through exploratory analysis based on advanced pattern detection (Choudhury et al., 2021; Shrestha et al., 2021), our approach leverages ML methods to unveil previously unknown interactions within well-studied phenomena. The study of new phenomena might pose challenges with respect to constructing a valid DAG which can in turn significantly reduce the validity of the obtained results.

Causality research typically encompasses two branches—causal discovery, which extracts causal relationships directly from observational data, and causal inference, which estimates the effect of changing a particular variable on an outcome—with this study explicitly focusing solely on causal inference and not addressing causal discovery (Nogueira, Pugnana, Ruggieri, Pedreschi, & Gama, 2022). We contribute to the methodological literature in organization science, first, by demonstrating how scholars can effectively mitigate "regularization bias," which is introduced when applying ML methods for causal inference, and second, by providing organization scholars with a framework based on DAGs to counter endogeneity concerns introduced by the selection of control variables. By rigorously following the DAG construction guidelines presented in this paper, we help scholars avoid problematic practices such as p-hacking, which involves strategically selecting or excluding control variables and repeatedly altering regression specifications to achieve statistical significance (Sturman, Sturman, & Sturman, 2022).

LIMITATIONS OF GRAPH-BASED DOUBLE MACHINE LEARNING

Algorithmic abduction supported by graph-based DML does have some limitations. Despite its robustness in handling high-dimensional settings, the validity of the findings relies heavily on the assumptions embedded in the causal models, represented by the DAG. These assumptions, while informed by theory and domain knowledge and scrutinized via sensitivity analysis, are ultimately unverifiable and could influence the conclusions drawn from the analysis. Importantly, our approach does not introduce any new assumptions but rather reflects the current state of the literature; within this framework, it enables the detection of effect heterogeneity, thereby contributing to a more nuanced understanding of existing theories and enhancing their applicability to the complexity of the social world. In terms of widespread adoption and dissemination of best practices, the introduction of advanced analysis methods, like graph-based DML, into organization scholarship poses challenges, due to a shortage of programming skills among researchers. Achieving scaled adoption requires upskilling researchers, fostering interdisciplinary collaboration, and establishing generalizable workflows like the one proposed in this paper. Finally, the proposed workflow is primarily suitable for the elaboration of variance theories, since these focus on explaining estimation changes in an outcome variable as a function of changes in independent variables. Process theories, in contrast, emphasize sequences of events and mechanisms over time and may not always be easily expressed by the estimation of interaction terms.

CONCLUSIONS

This paper offers an accessible demonstration of how organization scholars can use DML in combination with DAGs to explore and estimate interaction effects in a high-dimensional setting. With this approach, we aim to accelerate the adoption of causal ML

methods that enable researchers to make robust causal statements about effect heterogeneity to derive more nuanced theory. The method's potential to support the abductive reasoning process through robust estimation is a step forward in addressing the complexity of organizational phenomena. We illustrate our methodology with the gender wage gap, where it reveals the nuanced impact of socio-economic conditions on human capital investments of women and men. Based on the demonstration, we add to scholarly understanding of how individual attributes, like wealth, region of residence, and parental education, moderate the relationship between educational attainment and wage outcomes. By doing so, we advance the discourse around human capital and pay equity, previously constrained by studying the mechanisms across genders instead of at a gender-specific level. Our findings contribute to the field by demonstrating that nuanced policies and practices are essential for addressing systemic inequities in the workplace. Looking beyond HCT, the algorithm-supported abduction process presented in this paper offers a robust framework for exploring a variety of organizational phenomena in which interactions play an important role. Its generalizability and adaptability make it a valuable tool for abductive theorizing across multiple domains within organization science. Future research can leverage this methodology to unpack complex phenomena in organizational behavior, strategy, and decision making, where traditional approaches may fall short due to their inability to adequately capture effect heterogeneity. In conclusion, the presented workflow for causal ML enriches the methodological toolkit of quantitative organizational scholars, enabling a more detailed and robust examination of the causal mechanisms underlying complex managerial phenomena.

References

Abraham M (2017) Pay formalization revisited: Considering the effects of manager gender and discretion on closing the gender wage gap. *Academy of Management Journal* (Academy of Management), 29–54.

Aceves P, Evans JA (2024) Mobilizing Conceptual Spaces: How Word Embedding Models Can Inform Measurement and Theory Within Organization Science. *Organization Science* 35(3):788–814.

Aguinis H, Gottfredson RK (2010) Best-practice recommendations for estimating interaction effects using moderated multiple regression. *J Organ Behav* 31(6):776–786.

Ammermüller A, Weber AM (2005) *Educational Attainment and Returns to Education in Germany*

Angrist JD, Pischke JS (2009) *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press).

Antonakis J, Bendahan S, Jacquart P, Lalive R (2010) On making causal claims: A review and recommendations. *Leadership Quarterly* 21(6):1086–1120.

Athey S, Imbens G (2019) Machine Learning Methods That Economists Should Know About. *Annu Rev Econom* 11:685–725.

Bach P, Chernozhukov V, Cinelli C, Jia L, Klaassen S, Skotara N, Spindler M (2024) *Sensitivity Analysis for Causal ML: A Use Case at Booking.com*

Bach P, Chernozhukov V, Spindler M (2024) Heterogeneity in the US gender wage gap. *J R Stat Soc Ser A Stat Soc* 187(1):207–228.

Bamberger PA (2019) On The Replicability of Abductive Research in Management and Organizations: Internal Replication and Its Alternatives. *Academy of Management Discoveries* 5(2):103–108. (June 1).

Ban GY, El Karoui N, Lim AEB (2018) Machine learning and portfolio optimization. *Manage Sci* 64(3):1136–1154.

Barreto I (2012) A behavioral theory of market expansion based on the opportunity prospects rule. *Organization Science* 23(4):1008–1023.

Becker GS (1964) *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education* First Edition. (National Bureau of Economic Research).

Becker GS (1993) *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education* Third Edition. (University of Chicago Press, Chicago).

Behfar K, Okhuysen GA (2018) Discovery within validation logic: Deliberately surfacing, complementing, and substituting abductive reasoning in hypothetico-deductive inquiry. *Organization Science* 29(2):323–340.

Belliveau MA (2012) Engendering inequity? How social accounts create vs. merely explain unfavorable pay outcomes for women. *Organization Science* 23(4):1154–1174.

Blau FD, Kahn LM (2007) The Gender Pay Gap: Have Women Gone as Far as They Can? Executive Overview. *Academy of Management Perspectives* 21(1):7–23.

Blau FD, Kahn LM (2017) The gender wage gap: Extent, trends, & explanations. *J Econ Lit* 55(3):789–865.

Blevins DP, Sauerwald S, Hoobler JM, Robertson CJ (2019) Gender differences in pay levels: An examination of the compensation of university presidents. *Organization Science* 30(3):600–616.

Bliese PD, Certo ST, Smith AD, Wang M, Gruber M (2024) Strengthening Theory-Methods-Data Links. *Academy of Management Journal* 67(4):893–902.

Blöndal S, Field S, Girouard N (2003) Investment in human capital through upper-secondary and tertiary education. *OECD Economic Studies* 2002(1):41–89.

Bloom N, Eifert B, Mahajan A, McKenzie D, Roberts J (2013) Does management matter? Evidence from India. *Quarterly Journal of Economics* 128(1):1–51.

Breen R, Jonsson JO (2005) Inequality of opportunity in comparative perspective: Recent research on educational attainment and social mobility. *Annu Rev Sociol* 31:223–243.

Busenbark JR, Yoon H, Gamache DL, Withers MC (2022) Omitted Variable Bias: Examining Management Research With the Impact Threshold of a Confounding Variable (ITCV). *J Manage* 48(1):17–48.

Card D (1999) Chapter 30 - The Causal Effect of Education on Earnings. Ashenfelter OC, Card D, eds. *Handbook of Labor Economics* (Elsevier), 1801–1863.

Castilla EJ (2015) Accounting for the gap: A firm study manipulating organizational accountability and transparency in pay decisions. *Organization Science* 26(2):311–333.

Chang J, Boyd-Graber J, Gerrish S, Wang C, Blei DM Reading Tea Leaves: How Humans Interpret Topic Models.

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(1):C1–C68.

Chernozhukov V, Cinelli C, Newey W, Sharma A, Syrgkanis V (2022) *Long Story Short: Omitted Variable Bias in Causal Machine Learning*

Chernozhukov V, Hansen C, Kallus N, Spindler M, Syrgkanis V (2025) *Applied Causal Inference Powered by ML and AI*

Chernozhukov V, Hansen C, Spindler M (2016) High-Dimensional Metrics in R.

Chiswick BR (1978) The Effect of Americanization on the Earnings of Foreign-born Men. *Journal of Political Economy* 86(5):897–921.

Choi J, Menon A, Tabakovic H (2021) Using machine learning to revisit the diversification–performance relationship. *Strategic Management Journal* 42(9):1632–1661.

Choudhury P, Allen RT, Endres MG (2021) Machine learning for pattern discovery in management research. *Strategic Management Journal* 42(1):30–57.

Choudhury P, Kim DY (2019) The ethnic migrant inventor effect: Codification and recombination of knowledge across borders. *Strategic Management Journal* 40(2):203–229.

Choudhury P, Starr E, Agarwal R (2020) Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management Journal* 41(8):1381–1411.

Choudhury P, Wang D, Carlson NA, Khanna T (2019) Machine learning approaches to facial and text analysis: Discovering CEO oral communication styles. *Strategic Management Journal* 40(11):1705–1732.

Cinelli C, Hazlett C (2020) Making sense of sensitivity: extending omitted variable bias. *J R Stat Soc Series B Stat Methodol* 82(1):39–67.

Coleman JS (1988) Social Capital in the Creation of Human Capital. *American Journal of Sociology* 94:S95–S120.

Cortina JM (1993) *Interaction, Nonlinearity, and Multicollinearity: implications for Mu/tip/e Regression*

Cowgill B, Agan A, Gee LK (2024) The Gender Disclosure Gap: Salary History Bans Unravel When Men Volunteer Their Income. *Organization Science* 35(5):1571–1588.

Dale SB, Krueger AB (2014) Estimating the Effects of College Characteristics over the Career Using Administrative Earnings Data. *J Hum Resour* 49(2):323–358.

Diemer MA, Marchand AD, Mistry RS (2020) Charting How Wealth Shapes Educational Pathways from Childhood to Early Adulthood: A Developmental Process Model. *J Youth Adolesc* 49(5):1073–1091.

Dougherty C (2005) Why Are the Returns to Schooling Higher for Women than for Men? *Journal of Human Resources* 40(4):969–988.

Durand R, Vaara E (2009) Causation, counterfactuals, and competitive advantage. *Strategic Management Journal* 30(12):1245–1264.

Elkjaer B, Simpson B (2011) Pragmatism: A lived and living philosophy. What can it offer to contemporary organization theory? *Research in the Sociology of Organizations* 32:55–84.

Ellsaesser F, Tsang EWK, Runde J (2014) Models of causal inference: Imperfect but applicable is better than perfect but inapplicable. *Strategic Management Journal* 35(10):1541–1551.

Feuerriegel S, Frauen D, Melnychuk V, Schweisthal J, Hess K, Curth A, Bauer S, Kilbertus N, Kohane IS, van der Schaar M (2024) Causal machine learning for predicting treatment outcomes. *Nat Med* 30(4):958–968.

Goldin C (2014) A grand gender convergence: Its last chapter. *American Economic Review* 104(4):1091–1119.

Gould ED, Simhon A, Weinberg BA, Lach S, Hamermesh D, Angrist J, Lavy V, et al. (2020) *Does Parental Quality Matter? Evidence on the Transmission of Human Capital Using Variation in Parental Influence from Death, Divorce, and Family Size*

Grimes M, Von Krogh G, Feuerriegel S, Rink F, Gruber M (2023) From Scarcity to Abundance: Scholars and Scholarship in an Age of Generative Artificial Intelligence. *Academy of Management Journal* 66(6):1617–1624.

Grimmer J, Stewart BM (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3):267–297.

Hannigan TR, Haan RFJ, Vakili K, Tchalian H, Glaser VL, Wang MS, Kaplan S, Jennings PD (2019) Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals* 13(2):586–632.

He VF, Puranam P, Shrestha YR, von Krogh G (2020) Resolving governance disputes in communities: A study of software license decisions. *Strategic Management Journal* 41(10):1837–1868.

Hubbard WHJ (2011) The Phantom Gender Difference in the College Wage Premium. *Journal of Human Resources* 46(3):568.

Hünermund P, Bareinboim E (2025) Causal inference and data fusion in econometrics. *Econom J* 28(1):41–82.

Hünermund P, Louw B, Caspi I (2023) Double machine learning and automated confounder selection: A cautionary tale. *J Causal Inference* 11(1).

Hünermund P, Louw B, Rönkkö M (2025) The choice of control variables in empirical management research: How causal diagrams can inform the decision. *Leadership Quarterly* 36(2).

Kistruck GM, Slade Shantz A (2022) Research on Grand Challenges: Adopting an Abductive Experimentation Methodology. *Organization Studies* 43(9):1479–1505.

Kray LJ, Kennedy JA, Lee M (2024) Now, Women Do Ask: A Call To Update Beliefs About The Gender Pay Gap. *Academy of Management Discoveries* 10(1):7–33.

von Krogh G, Roberson Q, Gruber M (2023) Recognizing and Utilizing Novel Research Opportunities with Artificial Intelligence. *Academy of Management Journal* 66(2):367–373.

Kunze A (2008) Gender wage gap studies: Consistency and decomposition. *Empir Econ* 35(1):63–76.

Lam LW, Chuang A, Wong CS, Zhu JNY (2019) A typology of three-way interaction models: Applications and suggestions for Asian management research. *Asia Pacific Journal of Management* 36(1):1–16.

Liang H, Marquis C, Renneboog L, Sun SL (2019) Future-time framing: The effect of language on corporate future orientation. *Organization Science* 29(6):1093–1111.

Lonati S, Wulff JN (2024) Hic Sunt Dracones: On the Risks of Comparing the ITCV With Control Variable Correlations. *J Manage*.

Luo X, Jia N, Ouyang E, Fang Z (2024) Introducing machine-learning-based data fusion methods for analyzing multimodal data: An application of measuring trustworthiness of microenterprises. *Strategic Management Journal* 45(8):1597–1629.

Ly DP, Jena AB (2018) Sex Differences in Time Spent on Household Activities and Care of Children Among US Physicians, 2003-2016. *Mayo Clin Proc* 93(10):1484–1487.

Miric M, Jia N, Huang KG (2023) Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. *Strategic Management Journal* 44(2):491–519.

Moretti E (2012) *The New Geography of Jobs* (Houghton Mifflin Harcourt, Boston).

Nogueira AR, Pugnana A, Ruggieri S, Pedreschi D, Gama J (2022) Methods and tools for causal discovery and causal inference. *Wiley Interdiscip Rev Data Min Knowl Discov* 12(2). (March 1).

Overgoor G, Chica M, Rand W, Weishampel A (2019) Letting the computers take over: Using AI to solve marketing problems. *Calif Manage Rev* 61(4):156–185.

Pearl J (2009) *Causality* 2nd ed. (Cambridge University Press, Cambridge).

Peirce CS (1878) Deduction, Induction, and Hypothesis. *Popular Science Monthly* 13:470–482.

Perry LB, Thier M, Beach P, Anderson RC, Thoennessen NM, Roberts P (2024) Opportunities and conditions to learn (OCL): A conceptual framework. *Prospects (Paris)* 54(1):55–72.

Pfeffer FT, Schoeni RF (2016) How wealth inequality shapes our future. *RSF: The Russell Sage Foundation Journal of the Social Sciences* 2(6):2–22.

Psacharopoulos G, Patrinos HA (2004) Returns to investment in education: A further update. *Educ Econ* 12(2):111–134.

Rathje J, Katila R, Reineke P (2024) Making the most of AI and machine learning in organizations and strategy research: Supervised machine learning, causal inference, and matching models. *Strategic Management Journal* 45(10):1926–1953.

Rathje JM, Katila R (2021) Enabling technologies and the role of private firms: A machine learning matching analysis. *Strategy Science* 6(1):5–21.

Ren Z, Zhou Z (2024) Dynamic Batch Learning in High-Dimensional Sparse Linear Contextual Bandits. *Manage Sci* 70(2):1315–1342.

Sætre AS, Van De Ven A (2021) Generating theory by abduction. *Academy of Management Review* 46(4):684–701.

Schultz TW (1961) Investment in Human Capital. *Am Econ Rev* 51(1):1–17.

Shani AB, Coghlan D, Alexander BN (2020) Rediscovering Abductive Reasoning in Organization Development and Change Research. *Journal of Applied Behavioral Science* 56(1):60–72.

Shrestha YR, He VF, Puranam P, von Krogh G (2021) Algorithm supported induction for building theory: How can we use prediction models to theorize? *Organization Science* 32(3):856–880.

Sitzmann T, Campbell EM (2021) The hidden cost of prayer: Religiosity and the gender wage gap. *Academy of Management Journal* 64(4):1016–1048.

Stone P, Hernandez LA (2013) The all-or-nothing workplace: Flexibility stigma and "opting out" among professional-managerial women. *Journal of Social Issues* 69(2):235–256.

Sturman MC, Sturman AJ, Sturman CJ (2022) Uncontrolled Control Variables: The Extent That a Researcher's Degrees of Freedom With Control Variables Increases Various Types of Statistical Errors. *Journal of Applied Psychology* 107(1):9–22.

Tibshiranit R (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society* 58(1):267–288.

Tidhar R, Eisenhardt KM (2020) Get rich or die trying… finding revenue model fit using machine learning and multiple cases. *Strategic Management Journal* 41(7):1245–1273.

Tsang EWK (2022) Explaining Management Phenomena: A Philosophical Treatise. (Cambridge University Press, Cambridge), 32–74.

Vanneste BS, Gulati R (2022) Generalized Trust, External Sourcing, and Firm Performance in Economic Downturns. *Organization Science* 33(4):1599–1619.

Wager S, Athey S (2018) Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *J Am Stat Assoc* 113(523):1228–1242.

Wahrenburg Mark, Weldi Martin (2007) *Return on Investment in Higher Education – Evidence for Different Subjects, Degrees and Gender in Germany*

World Economic Forum (2023) *Global Gender Gap Report*

Table 1

Overview of ML Methods in Organization Science

High-Level Goal	Objective	ML Approach	Example References
Operationalization of Constructs	Identify meaningful groups in (large) datasets	Unsupervised ML (e.g., latent Dirichlet allocation, principal component analysis)	(Choi, Menon, & Tabakovic, 2021; Hannigan et al., 2019)
	Construct categorical variables in a (large) dataset where manual coding is infeasible	Supervised ML (e.g., random forest, neural network) Sentiment analysis	(Choudhury & Kim, 2019; Miric, Jia, & Huang, 2023)
	Construct numerical representation of textual, image, video or audio data	Supervised ML (e.g., random forest, neural Networks.) Unsupervised ML (e.g., word embeddings)	(Aceves & Evans, 2024; Choudhury, Starr, & Agarwal, 2020; Luo, Jia, Ouyang, & Fang, 2024)
Exploratory Data Analysis	Identify which independent variables are associated with the dependent variable	Supervised ML (e.g., LASSO, random forest)	(Choudhury et al., 2021; Shrestha et al., 2021; Tidhar & Eisenhardt, 2020)
Causal Effect Estimation	Robustly handle high-dimensional settings when performing conditioning procedures	Causal ML (e.g., double ML, causal forests)	(Durand and Vaara 2009, Wager and Athey 2018, Rathje and Katila 2021, Feuerriegel et al. 2024, Rathje et al. 2024)

Table 1

Table 2

Summary Statistics PSID

	No Degree		*Bachelor's degree*	
	Female	*Male*	*Female*	*Male*
# of Individuals	1,505	1,616	701	836
Individuals [% of Total]	68.2	65.9	31.8	34.1
Mean Hourly Salary [$]	22.38	26.80	32.16	45.84
Median Hourly Salary [$]	17.50	21.55	27.00	35.29
Std. Dev. Hourly Salary [$]	33.76	26.50	28.83	50.73
Mean Prof. Experience [Years]	9.9	9.7	8.9	8.6
Std. Dev. Prof. Experience [Years]	8.4	7.9	7.1	7.0
Region West [%]	13.3	16.2	14.8	19.1
Region South [%]	47.8	45.5	42.1	39.0
Region Northeast [%]	10.6	11.1	15.6	17.0
Region North Central [%]	28.0	26.8	27.4	24.5
Ethnic Black [%]	46.3	32.7	28.1	16.3
Ethnic White [%]	51.6	65.2	70.2	82.1
Ethnic Asian [%]	0.1	0.2	0.1	1.0
Ethnic Pacific Islander [%]	0.1	0.1	0.0	0.1
Ethnic Other [%]	1.0	1.0	1.1	0.5

Table 2

Table 3

List of Variables

Variable	*Description*	*Type*	*Baseline category*
Dependent Variable			
Salary	Log hourly salary from regular employment	Continuous	-
Independent Variables			
Degree	Bachelor's degree	Binary	No degree
Gender		Binary	Male
Education of the father		8 categories	0-5 grades of schooling
Education of the mother		8 categories	0-5 grades of schooling
Born outside of the US		Binary	Born in the US
Family wealth	Incl. various asset types (e.g., real estate, financial assets)	4 categories	3rd quartile
Region		4 categories	Northwest
Ethnicity		5 categories	White

Table 3

Table 4

Estimation Results

Variable	Estimate	*p* Value	0.05% CI	99.95% CI	RV [%]
Human Capital					
Bachelor's degree	0.4519	0.001	0.3761	0.5284	31.24
Gender					
Female	-0.2397	0.001	-0.3121	-0.1680	18.30
Two-Way Interaction with Gender					
Bachelor's degree * Female	0.1829	0.001	0.0879	0.278	10.33
Three-Way Interaction with Gender and Bachelor's degree					
Family wealth					
4th quartile * Bachelor's degree * Female	0.3524	0.001	0.2092	0.4949	11.44
Region					
Northeast * Bachelor's degree * Female	0.3253	0.001	0.1118	0.5388	7.86
Parental Education					
Father Bac. deg. * Bachelor's deg. * Female	0.2430	0.001	0.0854	0.4008	7.42
Mother Bac. deg. * Bachelor's deg. * Female	0.2047	0.001	0.0507	0.3551	6.04
Mother adv. deg. * Bachelor's deg. * Female	0.2284	0.001	0.0349	0.4219	5.74
Ethnicity					
Asian * Bachelor's degree * Female	0.3402	0.001	0.2899	0.3863	0.82

Table 4

Figure 1

Simple DAG with two confounders

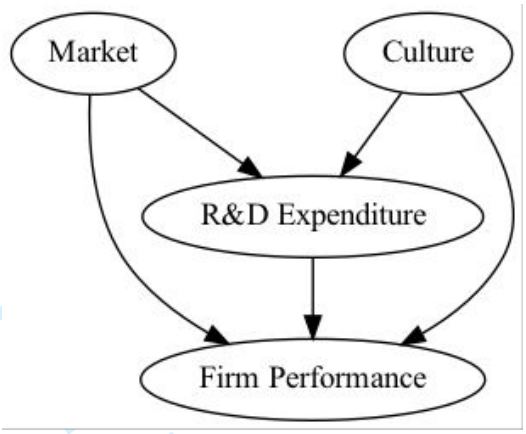

Figure 1

Figure 2

Workflow Diagram

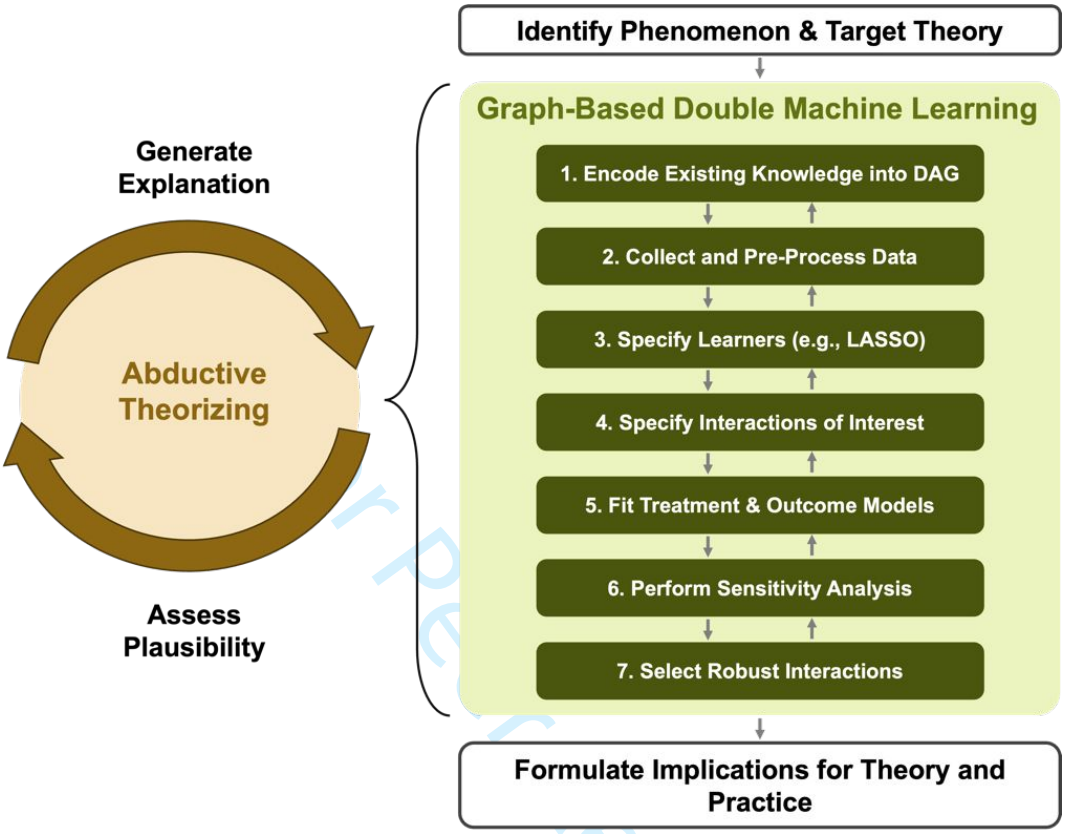

Figure 2

Figure 3

Directed Acyclic Graph (DAG) Representing the Relationship between Degree and Salary

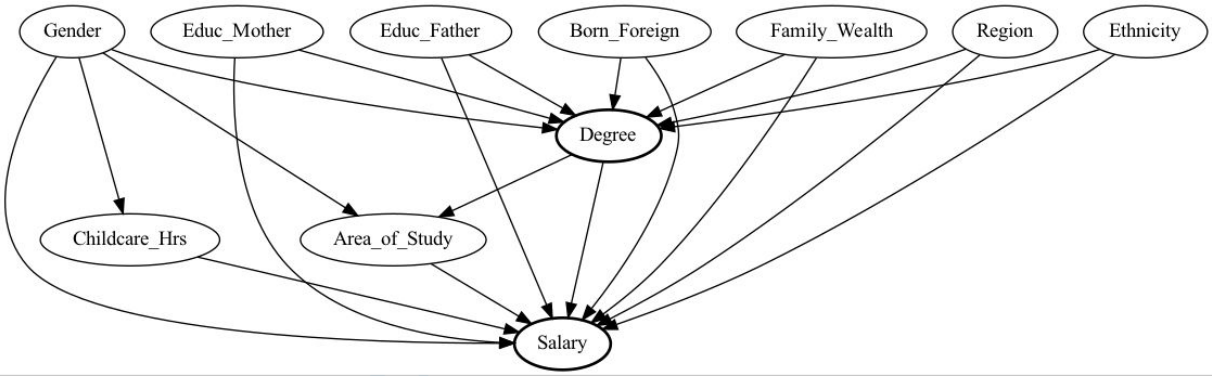

Figure 3

Figure 4

Treatment and Outcome Confounded by an Unobservable Variable

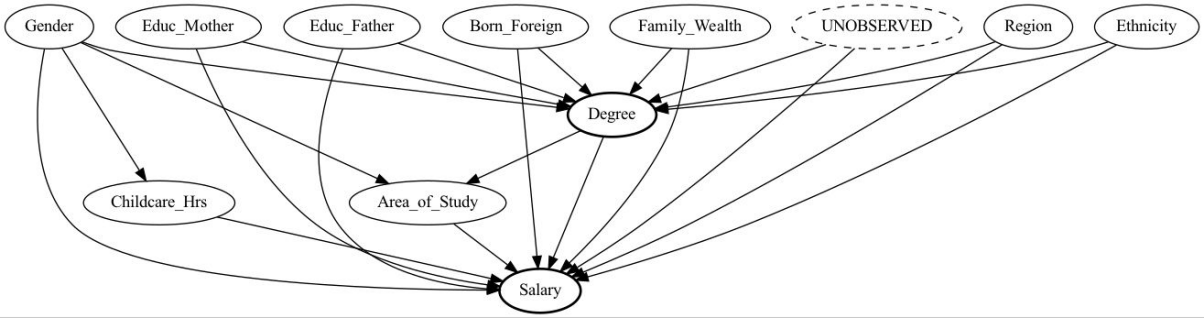

Figure 4