

PhD Project Description

CAREER-DML: Causal Embeddings for Labor Market Analysis

Rodolf Mikel Ghannam Neto

rodolf@cical.com.br · github.com/RodolfGhannam/CAREER-DML

Application to the PhD Programme in Strategy and Innovation

Copenhagen Business School

1. Research Purpose and Motivation

What is the causal impact of a strategic job move on an individual’s long-term earnings trajectory? This question is central to labor economics and corporate strategy, yet notoriously difficult to answer due to fundamental selection bias. My research addresses this gap by developing and validating a novel framework, CAREER-DML, designed for causal inference on sequential data. The framework stands as a bridge between the **two cultures of modern econometrics**—causal Machine Learning and structural modeling—by combining deep sequence embeddings with the semiparametric efficiency guarantees of Double Machine Learning (DML) (Chernozhukov et al., 2018).

This project is directly motivated by my 15+ years of executive experience, where I observed significant heterogeneity in employee development needs, highlighting the critical need for methods that can move beyond average effects. This proposal therefore aligns directly with the CBS Strategy Innovation department’s research agenda on “Digital Transformations” and “AI adoption and careers,” leveraging my practitioner-researcher background to apply advanced quantitative methods to large-scale administrative data.

2. Core Contributions and Positioning

This research makes three core contributions that form the basis of the proposed PhD project:

1. **The Sequential Embedding Ordering Phenomenon:** We document a robust empirical finding that causally-motivated embeddings (e.g., VIB) consistently yield *more* estimation bias than simpler predictive embeddings in this domain. This challenges the direct application of some causal representation learning methods to socio-economic trajectories.
2. **The Signal-to-Noise Frontier:** We characterize the sample size requirements for detecting realistic effect sizes, providing a rigorous, data-driven justification for the necessity of large-scale administrative data ($N > 1,034$) over typical survey data.
3. **A Bridge Between Two Cultures:** We build an explicit conceptual bridge between causal ML and the structural econometrics tradition, showing how our learned embeddings serve as non-parametric analogs of classical latent variables (e.g., human capital à la Ben-Porath (1967), latent types à la Keane and Wolpin (1997)).

3. Methodology and Data

3.1 The CAREER-DML Framework

The pipeline proceeds in two stages: (1) A Gated Recurrent Unit (GRU) learns a fixed-length embedding vector $z_i \in \mathbb{R}^{64}$ from an individual’s career history; (2) The CausalForestDML estima-

tor (Microsoft Research, 2019) uses these embeddings as high-dimensional controls to estimate the causal effect θ in the partially linear model $Y = \theta T + g(z) + \epsilon$.

3.2 Data Source Note

This project has been validated on a semi-synthetic Data Generating Process (DGP) calibrated with real-world US labor market data (NLSY79) and AI exposure scores (Felten et al., 2021). The primary data source for the PhD will be the Danish Integrated Database for Labour Market Research (IDA), which provides longitudinal employer-employee matched records for the entire Danish population.

Note on the Generative AI Shock: The reliance on the 2021 AIOE scores is a deliberate methodological choice. It provides a clean, pre-treatment baseline right before the major structural shift caused by the generative AI boom of late 2022. This temporal gap is an analytical advantage, allowing this project to set the stage to longitudinally disentangle the baseline effects of "traditional" AI adoption from the subsequent, compounding wage trajectories triggered by the Generative AI boom within the Danish registers.

4. Work Plan and Fit with CBS

This research is a perfect fit for the '**Digital Transformations**' research area at CBS. The proposed work plan is structured to produce three high-quality, independent but cohesive papers suitable for top-tier journals.

Year 1: Apply the validated CAREER-DML framework to the Danish registers. Write **Paper 1** (methodological), focusing on the Sequential Embedding Ordering Phenomenon. Target: AISTATS or a top econometrics journal.

Year 2: Write **Paper 2** (substantive), analyzing treatment effect heterogeneity of AI adoption across different demographic groups and industries in Denmark. Target: *Management Science* or *Strategic Management Journal*.

Year 3: Write **Paper 3** (policy-focused), simulating the long-term impacts of different AI adoption scenarios on wage inequality and career mobility. Consolidate papers into PhD thesis.

Key References: Chernozhukov, V. et al. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*. • Ben-Porath, Y. (1967). The Production of Human Capital and the Life Cycle of Earnings. *Journal of Political Economy*. • Keane, M. P., Wolpin, K. I. (1997). The Career Decisions of Young Men. *Journal of Political Economy*. • Felten, E., et al. (2021). Occupational hierarchy and the labor market impacts of automation. *AEA Papers and Proceedings*.

References

- Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *Journal of Political Economy*, 75(4):352–365.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The*

Econometrics Journal, 21(1):C1–C68.

Felten, E., Raj, M., and Seamans, R. (2021). Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, 42(12):2195–2217.

Keane, M. P. and Wolpin, K. I. (1997). The career decisions of young men. *Journal of Political Economy*, 105(3):473–522.

Microsoft Research (2019). EconML: A python package for econometric machine learning. Technical report, Microsoft.