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 AI-driven career transitions on individual earnings trajectories? This question is central to labor economics and corporate strategy, yet notoriously difficult to answer due to fundamental selection bias: workers who change jobs differ systematically from those who do not, and these differences are entangled with the very outcomes we wish to measure. Classical approaches—from Heckman selection models to fixed-effects panel regressions—have made progress, but they rely on restrictive functional form assumptions or require instruments that are rarely available in practice.

My research addresses this gap by developing and validating a novel framework, CAREER-DML, designed for causal inference on sequential career 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). Rather than imposing a parametric structure on career histories, the framework *learns* a rich, low-dimensional representation of an individual's entire occupational trajectory and uses it to flexibly control for selection.

This project is directly motivated by my 15+ years of executive experience, where I observed significant heterogeneity in employee development needs—a pattern that standard aggregate analyses consistently failed to capture. At Grupo CICAL, I designed a data-driven customer service methodology whose analytics revealed that employees with ostensibly similar backgrounds diverged sharply in their development trajectories, requiring a shift from group-level training to personalized coaching. This lived experience is the direct inspiration for a research agenda focused on understanding *why* career outcomes differ, not merely *that* they differ. The heterogeneity I observed in the field—employees with identical backgrounds diverging sharply—is precisely the variation that the nuisance function $g(z)$ in the DML framework is designed to capture and control for. The 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.

The urgency of this question has intensified: AI adoption in the Nordic countries has surpassed 35%, with 34.7% of the Danish workforce already exposed to generative AI in tasks representing 20% or more of their work hours (OECD, 2024; 2026). Understanding the causal—not merely correlational—impact of this exposure on career trajectories is a first-order policy question for Denmark and for the broader European labor market.

2. Core Contributions and Positioning

This research makes three core contributions that form the basis of the proposed PhD project and position it at the intersection of causal inference, machine learning, and labor economics:

1. **The Sequential Embedding Ordering Phenomenon:** We document a robust empirical finding that causally-motivated embeddings (e.g., the Variational Information Bottleneck approach of Veitch et al. 2020) consistently yield *more* estimation bias than simpler predictive embeddings in the DML pipeline for career data. This result holds even after controlling for model dimensionality and training dynamics, and it challenges the direct application of some causal representation learning methods to socio-economic trajectories. The finding suggests that, in domains where the confounding structure is diffuse and high-dimensional, the information bottleneck may discard precisely the variation needed for debiasing.

2. **The Signal-to-Noise Frontier:** We characterize the sample size requirements for detecting realistic effect sizes in observational career data. Our calibration experiments show that for plausible treatment effects ($\theta \approx 0.05$–0.10 standard deviations), reliable detection requires sample sizes on the order of $N > 1{,}000$, far exceeding what is available in typical survey datasets like the NLSY79. This provides a rigorous, data-driven justification for the necessity of large-scale administrative data—and a direct motivation for applying the framework to the Danish registers.

3. **A Bridge Between Two Cultures:** We build an explicit conceptual bridge between causal ML and the structural econometrics tradition. We show that the learned GRU embeddings serve as non-parametric analogs of classical latent variables—human capital accumulation à la Ben-Porath (1967) and latent occupational types à la Keane and Wolpin (1997). Linear probing experiments confirm that the embeddings encode interpretable economic structure (occupation, industry, tenure) without being explicitly trained on these labels. This bridges the gap between the "prediction-first" culture of ML and the "structure-first" culture of economics, offering a principled way to combine the strengths of both.

This positioning is distinctive. While recent work by Chernozhukov et al. (2018) and the EconML project (Microsoft Research, 2019) has advanced the toolkit for causal ML, applications to sequential career data remain scarce. Similarly, while the structural tradition (Keane and Wolpin, 1997) provides deep theoretical insight, it requires strong parametric assumptions that limit scalability. CAREER-DML occupies the space between these two traditions: it does not implement structural models directly, but it captures empirically what they theorize. The learned embeddings serve as data-driven proxies for the latent constructs that structural economists postulate but rarely observe—human capital stocks, latent occupational types, and task-specific skill bundles. This dialogue with the structural tradition is not merely rhetorical; linear probing experiments confirm that the embeddings encode interpretable economic structure (occupation, industry, tenure) without being explicitly trained on these labels, providing empirical evidence that the representations are economically meaningful rather than statistical artifacts.

3. Methodology and Data

3.1 The CAREER-DML Framework

The pipeline proceeds in two stages. In the **first stage**, a Gated Recurrent Unit (GRU) network processes an individual's sequential career history—a time-ordered sequence of occupation codes, industry codes, and tenure durations—and produces a fixed-length embedding vector $z_i \in \mathbb{R}^{64}$. The GRU architecture is chosen for its ability to capture long-range dependencies in sequential data while remaining computationally tractable. We implement and compare four embedding variants:

- **Predictive GRU:** Trained to predict next-period outcomes (standard sequence prediction).
- **VIB (Variational Information Bottleneck):** A faithful implementation of the two-stage procedure in Veitch et al. (2020), which explicitly compresses the representation to retain only treatment-relevant information.
- **Autoencoder GRU:** Trained for sequence reconstruction, capturing the full informational content of the career history.
- **Hybrid GRU:** A multi-task variant combining predictive and reconstruction objectives.

In the **second stage**, the `CausalForestDML` estimator (Microsoft Research, 2019) uses these embeddings as high-dimensional controls to estimate the causal effect θ in the partially linear model:

$$Y_i = \theta T_i + g(z_i) + \epsilon_i$$

where Y_i is the outcome (log wages), T_i is the treatment indicator (e.g., job transition to an AI-exposed occupation), $g(\cdot)$ is a flexible nuisance function estimated via gradient boosting, and z_i is the learned embedding. The DML framework provides $\sqrt{n}$-consistent and asymptotically normal estimates of θ under weak regularity conditions, even when $g(\cdot)$ is estimated non-parametrically.

3.2 Validation: The Semi-Synthetic Data Laboratory

A critical methodological contribution is the construction of a semi-synthetic Data Generating Process (DGP) calibrated with real-world parameters. The DGP uses occupational transition matrices from the NLSY79 and AI exposure scores from Felten et al. (2021) to generate realistic career sequences with a *known* true treatment effect ($\theta_0 = 0.50$). This allows us to benchmark each embedding variant against ground truth—a luxury unavailable in purely observational studies.

Preliminary results from this laboratory are encouraging. The Predictive GRU achieves an ATE estimate of $\hat{\theta} = 0.5378$ (SE: 0.0520, 95% CI: [0.4358, 0.6397]), representing a bias of only 7.6%, compared to 945% for naïve OLS—a reduction of over 90%—and a substantial improvement over the Heckman two-step model. A formal GATES heterogeneity test confirms statistically significant skill-biased treatment effects, consistent with the theoretical prediction that AI adoption disproportionately affects workers in routine-intensive occupations.

3.3 Data for the PhD: The Danish Registers

The primary data source for the PhD will be the **Danish Integrated Database for Labour Market Research (IDA)**, administered by Statistics Denmark. The IDA provides longitudinal employer-employee matched records for the *entire* Danish population, including detailed

information on occupation, industry, firm, wages, education, and demographic characteristics. This dataset offers three decisive advantages over the survey data used in the proof-of-concept:

- **Scale:** $N > 1{,}000{,}000$ individuals, far exceeding the Signal-to-Noise Frontier identified in our calibration.
- **Temporal depth:** $T > 30$ years of annual observations, enabling the study of long-run career dynamics and life-cycle effects.
- **Population coverage:** No sampling bias or attrition, eliminating a major source of concern in survey-based studies.

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 longitudinally disentangle the baseline effects of "traditional" AI adoption from the subsequent, compounding wage trajectories triggered by the Generative AI shock within the Danish registers.

4. Work Plan and Fit with CBS

This research is a natural 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, under the guidance of Professor Tom Grad, whose research on competitive dynamics and digital transformation provides a natural intellectual anchor for this project. Data access will be facilitated through the CBS research infrastructure and Statistics Denmark's researcher access program.

Year 1 — Methodological Foundation: Apply the validated CAREER-DML framework to the Danish IDA registers. Adapt the GRU architecture to the Danish occupational classification (DISCO-08) and train embedding models on the full population. Write **Paper 1** (methodological), providing a rigorous analysis of the Sequential Embedding Ordering Phenomenon at population scale. Investigate whether the phenomenon persists, attenuates, or reverses when the Signal-to-Noise Frontier is crossed. My working hypothesis is that the phenomenon attenuates at population scale, where the richer variation in the Danish registers provides the information that the bottleneck currently discards. Target: *The Econometrics Journal*, *AISTATS*, or *Journal of Machine Learning Research*. The modular architecture of CAREER-DML is designed for this transition: the GRU embedding stage and the DML estimation stage scale independently, allowing the pipeline to process millions of career sequences without architectural modification. Coursework in advanced econometrics and causal inference.

Year 2 — Substantive Analysis: Write **Paper 2** (substantive), estimating heterogeneous treatment effects of AI adoption across demographic groups (age, gender, education), industries, and regions in Denmark. Leverage the GATES and CLAN frameworks to identify which subpopulations are most affected by AI-driven career transitions. Explore the bridge to structural models by comparing embedding-derived latent types with those from parametric mixture models. Target: *Management Science* or *Strategic Management Journal*. Research stay at a partner institution (e.g., MIT Economics, Stanford SIEPR).

Year 3 — Policy Implications and Synthesis: Write **Paper 3** (policy-focused), simulating the long-term impacts of different AI adoption scenarios on wage inequality and career mobility in Denmark. Use the pre-2022 baseline to construct counterfactual trajectories under varying Generative AI diffusion rates. Consolidate the three papers into a coherent PhD thesis. Target: *American Economic Review: Insights* or *Journal of Labor Economics*. Dissemination at major conferences (NBER Summer Institute, European Economic Association).

5. Expected Contributions

This project begins where the proof-of-concept ends—at the precise boundary where survey data fails and administrative data becomes essential. It is expected to make contributions along three dimensions. **Methodologically**, it will provide the first large-scale empirical test of causal representation learning methods on administrative career data, advancing our understanding of when and why these methods succeed or fail. **Substantively**, it will produce among the most comprehensive causal estimates to date of AI adoption's impact on individual career trajectories in a Nordic labor market, with direct relevance to Danish and European policy. **Theoretically**, the bridge between causal ML and structural econometrics opens a new research agenda: using learned representations to test and extend classical models of human capital formation, occupational choice, and technological change—without the parametric restrictions that have historically limited their empirical reach. For the causal representation learning community specifically, the Sequential Embedding Ordering Phenomenon provides a rigorous empirical benchmark—documenting when and why information bottleneck approaches fail in high-dimensional socio-economic settings, and pointing toward design principles for the next generation of causal embedding methods.

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.* • Veitch, V., et al. (2020). Adapting Text Embeddings for Causal Inference. *UAI.*

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.

Veitch, V., Sridhar, D., and Blei, D. M. (2020). Adapting text embeddings for causal inference. In *Uncertainty in Artificial Intelligence*.