# Taking Out the Human Factor: Comparing Algorithmic Profiling Against Human Baselines

Christoph Kern[1,2][0000−0001−7363−4299], Max Kunaschk[3], Julia Lang[3][0000−0002−6832−8665], Ruben Bach[4][0000−0001−5690−2829], and Frauke Kreuter[1,2][0000−0002−7339−2645]

[1] Department of Statistics, LMU Munich, Ludwigstraße 33, 80539 Munich, Germany
[2] Munich Center for Machine Learning (MCML), Geschwister-Scholl-Platz 1, 80539 Munich, Germany
{christoph.kern,frauke.kreuter}@stat.uni-muenchen.de
https://www.stat.lmu.de/soda/en/
[3] Institute for Employment Research (IAB), Regensburger Straße 104, 90478 Nuremberg, Germany
{max.kunaschk,julia.lang}@iab.de
[4] Mannheim Centre for European Social Research, University of Mannheim, A5, 6, 68159 Mannheim, Germany
r.bach@uni-mannheim.de

**Keywords:** Algorithmic Decision-Making · Human-Algorithm Comparison · Public Employment Services · Algorithmic Fairness.

## Extended Abstract

Public administrations, facing growing demands and limited resources, increasingly turn towards algorithmic decision-making (ADM) systems to enhance efficiency and effectiveness across a range of public services. From criminal justice and child protection to immigration and employment services, algorithmic tools promise consistency and improved accuracy over human judgment in tasks such as triaging cases, allocating interventions, and forecasting risk. At the same time, research in algorithmic fairness pushes back against the use of algorithms in public settings, following reports of systematic prediction errors and false accusations that disproportionally affect minority groups. These risks can be attributed to social biases embedded in the models' training data: when previous judgments of humans were affected by stereotyping, discrimination, and prejudice, algorithms can learn and potentially amplify such patterns.

Both phenomena – hopes of increased accuracy and concerns of algorithmic bias – give rise to a fundamental question: *what would have been the outcome if a case had been assessed by an algorithm rather than by a human and vice versa*? The shift from human to algorithmic decision-making raises fundamental questions regarding the role of human discretion in decision-making and the implications of replacing human judgment with predictions in sensitive domains. Human decision-makers bring contextual understanding and social sensitivity to their work, but also have biases and inconsistencies. ADM systems offer the allure

of neutrality and scale, but often lack the nuanced judgment required in complex social contexts and risk amplifying societal biases embedded in training data. Directly comparing human and algorithmic assessments on the same grounds is critical to understanding how such trade-offs between the two types of decision-making unfold, where they disagree, and how they might complement each other. Yet, research in ADM often operates under implicit assumptions regarding the accuracy and fairness of the non-algorithmic baseline: the decisions made by human caseworkers in real-world operational processes. Until now, algorithms have rarely been evaluated against a meaningful human baseline, likely due to a lack of data.

This paper presents a *comparative analysis* of human versus algorithmic predictions in the context of public employment services in Germany. Specifically, we study the task of predicting long-term unemployment risk using a unique dataset that combines rich administrative labor market records with contemporaneous caseworker assessments collected at initial intake. The caseworker assessments include a prognosis about each job seeker's labour market re-integration chances, providing a real-world comparison point for model-based predictions that is often missing in ADM research. Risk predictions are used to identify job seekers in need of special support and to allocate interventions accordingly. This setting enables us to directly compare three approaches to decision-making: (1) human caseworkers at local job agencies, (2) a machine learning model trained on historical administrative data and observed outcomes, and (3) a hybrid model trained to emulate the assessments of caseworkers. By evaluating all three approaches on the same grounds, we show how humans and algorithms systematically diverge in their assessment of social subgroups, providing new empirical evidence on the hidden costs of automation in public domains.