

과업의 반복성과 AI 노출이 고용에 미치는 인지 수준별 이질적 영향: 생성형 AI 확산기를 중심으로

정덕재¹

한국관세무역개발원

초 록

본 연구는 생성형 AI 확산기에 고용이 인지(cognitive) 수준에 따라 어떻게 달라졌는지를 반복성 강도(routine)와 AI 노출 강도를 사용하여 분석한다. 먼저 O*NET 과업 정보를 대규모 언어모델(LLM)로 채점해 SOC 6자리 직업분류에서 반복성 점수와 인지 점수를 세밀하게 구하였다. 그 후 반복성 강도를 연속형 처치로, 인지 점수를 표본 분할 기준으로 삼아, 2단계 이중차분으로 미국의 고용량을 분석하였다. 반복성 강도를 처치로 하면 사후 계수의 부호가 인지 수준을 경계로 갈렸다. 저인지의 음(-)의 사후계수는 AI에 의한 새로운 현상이 아니라 반복편향 기술변화(Routine-Biased Technological Change)의 지속된 패턴으로 설명되는 반면, 고인지의 양(+)은 반복편향 기술변화의 예측과 다르게 고인지 그룹에서 오히려 반복적인 직업이 떠오르고 있음을 시사한다. 한편 같은 설계에 AI 노출지수를 처치로 넣으면, 저인지 직업군에서 사후 계수가 유의하지 않았다. 반면 고인지 직업군에서는 사후 계수가 음(-)으로 추정되어 반복성을 처치군으로 썼던 경우와 부호가 반대이다. 반복성과 AI 노출지수라는 두 측정치는 상관관계가 없으므로, 두 지표는 노동시장의 서로 다른 메커니즘을 비춘다고 해석된다. 이는 반복편향 기술변화 문헌이 의지해 온 반복성 지표가 생성형 AI의 고용 영향을 설명하는 메커니즘을 대변하지 않을 수도 있음을 시사한다.

JEL 코드: C23, C81, J21, J23, J24, O33

키워드: 인공지능, 반복과업강도, 인지과업강도, AI 노출강도, 직업별 고용, 연속 처치 이중차분

¹ 이메일: ubuzuz@gmail.com, 홈페이지: www.jayjeo.com

본 연구에서 사용된 모든 데이터와 STATA 코드는 아래 링크에서 공개한다.

<https://github.com/jayjeo/Effects-of-Routine-Intensity-and-AI-Exposure>

Cognitive-Level Heterogeneity in the Employment Effects of Routine Intensity and AI Exposure: Evidence from the Diffusion of AI

Deokjae Jeong

Korea Customs and Trade Development Institute

Abstract

Using routine intensity and AI exposure, this paper analyzes how employment shifted across cognitive levels during the diffusion of AI. This study first constructs the data by using a large language model to score O*NET task descriptions, obtaining granular routine and cognitive scores at the 6-digit SOC occupation level. The paper then analyzes U.S. employment with a two-stage Difference-in-Differences design that treats routine intensity as a continuous treatment and uses the cognitive score to split the sample. The sign of the post-treatment coefficient splits at the cognitive threshold. The negative coefficient in the low-cognitive group is not a new, AI-driven phenomenon but is explained by the continued pattern of Routine-Biased Technological Change, whereas the positive coefficient in the high-cognitive group suggests that, contrary to what Routine-Biased Technological Change predicts, routine occupations are gaining ground within the high-cognitive group. When the same design instead uses AI exposure as the treatment, the post-treatment coefficient is not statistically significant for the low-cognitive group, whereas for the high-cognitive group it is estimated to be negative –opposite in sign to the case using routine intensity as the treatment. Since these two measures, routine intensity and AI exposure, are uncorrelated, they are interpreted as offering different perspectives on the labor market. This suggests that the routine measure on which the Routine-Biased Technological Change literature has relied does not represent the mechanism through which AI affects employment.

JEL Code: C23, C81, J21, J23, J24, O33

Keywords: Artificial Intelligence, Routine Task Intensity, Cognitive Task Intensity, AI exposure, Occupational Employment, Continuous Difference-in-Differences

제1장. 서론

생성형 인공지능이 2022년 말 대화형 모델의 공개와 함께 빠르게 보급되면서, 어떤 일자리가 이 기술에 영향을 받을 것인가라는 물음이 노동경제학의 과제로 떠올랐다. 생성형 AI는 규칙을 따르기만 하면 되는 반복적 일을 대신하는 기술이라기보다 사람처럼 사고하는 능력에 가깝고, 그렇다면 고인지 직군에서 반복성이 약한 비정형 과업을 더 깊이 건드릴 수 있다. 반면 손과 신체를 쓰는 저인지 직업은 생성형 AI가 대신하기 어려우므로 고용에 별다른 영향이 없으리라는 데에는 이견이 적다. 본 연구는 이러한 예상들을 미국의 인구현황조사(Current Population Survey)를 사용해 실증적으로 검토하되, 질문을 “생성형 AI가 고용을 얼마나 줄였는가”에 두지 않고 더 넓은 시야에서 묻는다. 본 연구가 답하고자 하는 질문은, 생성형 AI가 본격적으로 확산하기 시작한 2023년 전후를 지나며 직업의 반복성(routine)의 강도와 고용량의 관계가 인지(cognitive) 수준별 직업에 따라 어떻게 달라졌으며, 그 변화가 무엇으로 설명되는가이다. 또한 같은 질문을 반복성 대신 AI 노출지수를 사용했을 경우는 어떻게 결과가 바뀌는가이다.² 반복성의 강도란 1 (routine)에 가까울수록 정형적·자동화 가능한 과업을, 0 (non-routine)에 가까울수록 예상 밖의 대처가 요구되는 비정형 과업을 가리킨다. 인지 강도란 1 (cognitive)에 가까울수록 정신적 처리가 중심인 인지적 과업을, 0 (manual)에 가까울수록 손과 신체를 쓰는 육체적 과업을 가리킨다.

본 연구가 던진 질문에 답하기 위해 세 가지 단계로 분석한다. 첫째, 직업의 반복성과 인지성을 새로운 방식으로 측정한다. 반복편향 기술변화(Routine-Biased Technological Change, 이하 RBTC)를 다루는 선행 연구는 직업의 반복과업강도(Routine Task Intensity, 이하 RTI)를 어떻게 측정하느냐에 따라 고용영향분석의 결론이 달라진다는 측정의 문제를 인지하고 있었다(Walo 2023; Haslberger 2022). 이에 본 연구는 O*NET³의 직업별 과업 기술(task description)을 대규모 언어모델(LLM)로 직접 해석하게 하여 0에서 1 사이의 반복성 점수와 인지 점수로 채점한다. 이는 저자가 아는 한 반복과업강도 측정에 LLM을 적용한 첫 시도다. 이 방식으로 반복성과 인지 점수를 측정하면 기존 선행연구의 측정보다 훨씬 세밀한 직업분류별 측정이 가능하다. 둘째, 이렇게 얻은 반복성 강도를 시간 불변의 연속강도 변수인 처치군으로 삼아 2단계 이중차분(Difference-in-Differences)을 추정한다. 구체적으로, 직업군마다 취업자수의 장기 추세가 다르므로, 1단계에서 직업군별 장기 추세를 제거한 취업자수를 얻는다. 그 다음, 이 종속변수를 사용해서 2단계

² 생성형 AI와 노동을 잇는 최근 연구는 어떤 직업이 AI에 노출되는지를 측정하는 데 집중한다(Eloundou et al. 2024; Webb 2019; Felten, Raj, and Seamans 2021). 노출지수는 그 자체로는 어떤 직업이 영향을 받을 잠재적 여지가 크다는 사실을 가리킬 뿐, 그것이 고용의 증가로 같지 감소로 같지까지 지표값에 담고 있지는 않다. 본 연구는 이 노출지수를 처치 강도로 삼아 이중차분으로 분석하면, 적어도 인지 수준으로 표본을 나눌 때 고인지 직군에서 AI 노출도가 큰 직업일수록 고용이 감소함을 보인다. 다만 생성형 AI의 채택이 빠르게 변지고 있음에도(Bick, Blandin, and Deming 2025; Humlum and Vestergaard 2024) 확산 초기의 노동시장 효과는 대체로 작거나 그 식별이 제한적이라는 평가가 이어지고(del Rio-Chanona et al. 2025), 같은 시기에 AI와 관련없는 충격이 겹친 상황에서 AI라는 단일 원인으로의 귀속은 설득력을 얻기 어려울 수 있다.

³ O*NET(Occupational Information Network)은 미국 노동부(U.S. Department of Labor) 산하 고용훈련청(Employment and Training Administration, ETA)이 후원하는 직업정보 데이터베이스로, 900여 개 직업에 대한 지식·기술·능력, 과업, 작업 환경 등의 정보를 표준화된 형태로 제공한다.

에서 시점별 용량-반응 계수를 이벤트 스터디로 추정한다. 이때 사후 시점의 계수가 0보다 작게 추정되면 본 논문에서는 이를 음(-)의 계수, 0보다 크게 추정되면 양(+의 계수로 서술한다. 한편, 전 과정의 정확한 표준오차를 구하기 위해 직업군 군집 부트스트랩(cluster bootstrap)을 사용한다. 셋째, 전체 표본을 인지 점수 0.5를 기준으로 두 집단으로 나누어, 저인지(육체 과업 중심)와 고인지(두뇌 과업 중심) 직업군에서 반복성 강도에 대한 사후 계수가 어떻게 갈리는지를 비교한다. 또한 반복성 강도와 대등하게 AI 노출지수를 같은 2단계 이중차분의 처치 강도로 삼아 추정하여, 두 측정 축이 고용 변화를 어떻게 다르게 비추는지 비교한다.

분석의 핵심 발견부터 밝힌다. 반복성 강도에 대한 사후 계수의 부호가 인지 수준을 경계로 갈라졌다. 두뇌 과업이 적은 저인지 직업군에서는 처치 이후 반복성이 높을수록 장기추세를 제거한 고용량이 상대적으로 더 떨어졌고, 그 결과 사후 계수가 음(-)으로 추정되었다. 반면 두뇌 과업이 많은 고인지 직업군에서는 반복성이 높을수록 도리어 고용량이 상대적으로 개선되어, 사후 계수가 양(+)으로 추정되었다. 더구나 두 집단의 사후 계수는 방향만 반대인 것이 아니라 크기도 한쪽으로 쏠려, 저인지 쪽 음(-)의 계수가 고인지 쪽 양(+의 계수의 두세 배에 달했다. 다만 여기서 음과 양은 일자리의 절대적 증감이 아니라 직업군별 장기 고용량 추세를 걷어낸 뒤의 상대적인 고용량 크기임을 분명히 해 둘 필요가 있다. 즉, 사후 계수가 양(+)으로 추정되더라도 고용이 늘었다는 신호가 아니라 이전 추세만큼 줄지는 않았다는 신호일 수 있다.

본 연구는 이 부호 반전을 생성형 AI의 단일 효과로 단정하지 않고, 부호가 갈린 두 집단의 사후 계수가 각각 어디에서 비롯하는지를 갈래로 나누어 분석한다. 먼저 저인지 집단에서 관측된 음(-)의 사후 계수는 새삼스러운 현상이 아니라, 반복적·육체적 과업이 기계와 소프트웨어에 밀려난 반복편향 기술변화(Routine-Biased Technological Change, 이하 RBTC)의 오랜 패턴 위에서 비교적 자연스럽게 설명된다. RBTC의 관점에서 노동을 본 초기 연구(Autor, Levy, and Murnane 2003)⁴ 이래로 규칙성이 있는 반복적 과업은 기계로 옮겨지기 쉽다고 이해되어 왔고, 그러한 대체효과가 중간 숙련자의 반복성 직업을 깎아내며 노동시장을 양극으로 벌려 온 사정은 널리 보고되었다(Autor and Dorn 2013). 로봇과 물리적 자동화를 직접 들여다본 연구들도 그 축출효과(displacement effect)⁵가 반복적·육체적 직무에 집중됨을 확인한다(Acemoglu and Restrepo 2020; Dauth et al. 2017).⁶ 저인지·고반복 직업의 상대적 위축은 이처럼 생성형 AI 이전부터 작동하던 힘이 반복성(routine)이라는 원인을 따라 또다시 나타난 것으로 보는 편이 자연스럽다.

반대로 고인지 집단에서 관측된 양(+의 사후 계수는 AI의 영향 때문인지 불확실하다. 반복성은 과업이

4 보다 엄밀하게, 그들은 컴퓨터 자동화가 고인지·반복적 과업과 저인지·반복적 과업을 “대체”하고, 고인지·비반복적 과업과 저인지·비반복적 과업을 “보완”한다고 표현하였다.

5 Acemoglu and Restrepo (2020)이 사용한 용어로, 대체효과(substitution effect) 및 복원효과(reinstatement effect)와 구별되는 개념이다.

6 보다 엄밀하게, Acemoglu and Restrepo, Dauth et al. (2020)은 자동화의 부정적 고용효과가 주로 제조업, 저인지 반복 직업군, 블루 컬러 직종, 조립 작업을 하는 직종에 집중된다고 표현하였다.

규칙적이냐 비규칙적이냐의 정도를 짚 뿐, 생성형 AI가 과업을 대체할 수 있는지와 곧바로 1:1로 대응하지 않는다. AI의 발전에 대한 직업의 노출을 측정할 지표들은 비반복적 두뇌 과업에서 AI 노출값이 높다고 보고된다(Eloundou et al. 2024; Webb 2019; Felten, Raj, and Seamans 2021). 본 연구에서 분석한 결과에 따르면, 반복성 강도와 AI 노출강도는 전체 집단에서는 음의 상관관계를 지니나, 고인지 집단에만 한정해서 보았을 때는 두 강도는 아무런 상관관계가 없다. 따라서 왜 고인지 집단에서, AI 노출이 많이 되는 직업일수록 고용이 감소한 반면, 반복성이 강한 직업일수록 고용이 증가했는지는 미지수이다. 애초에 고인지 집단에서 두 강도의 상관관계가 없기 때문에, 반복성의 결과가 오직 AI의 영향 때문이라는 단정은 위험하다.

고인지 집단에서 반복적인 직업일수록 고용이 증가했다는 결과는 기존의 반복편향 기술변화(RBTC)가 설명해온 메커니즘과 대치되는 현상이 AI가 도입되기 시작한 2023년 이후로 관찰되고 있음을 시사한다. 반복편향 기술변화(RBTC)에 따르면 근로자의 인지 수준과 무관하게 반복성이 강한 직업일수록 자동화 기술에 의해 대체되어 고용비중을 겪는 것이 일반적이다. 그런데 오히려 2023년 이후로는 고인지 집단에서 반복성이 강한 직업이 고용의 입지를 넓히는, 기존의 이론적 예측을 벗어난 이질적 변화가 확인된다. 이러한 고용 궤적의 반전이 전적으로 생성형 AI의 도입에 기인한 것인지 인과적으로 단언하기는 어려우나, 적어도 구조적 변화의 발생 시점이 AI의 확산기와 정확히 일치한다는 점은 AI 기술이 고인지 반복 과업에 대해 기존의 컴퓨터 자본과는 다른 강력한 보완적 생산성 충격으로 작용하고 있을 가능성을 강하게 뒷받침한다.

같은 이중차분 설계에 반복성 강도 대신 AI 노출지수를 처치변수로 넣으면 사뭇 다른 그림이 나타난다. 고인지 직업군에서 사후 계수가 음(-)으로 추정되어, AI의 노출을 많이 받는 직업일수록 고용이 부정적인 영향을 받았다. 이는 반복성을 처치변수로 놓았을 때 양(+),을 가리킨 것과는 정반대 부호이다. 하지만 두 추정치는 고인지 조건하에서 사실상 독립이어서, 음(-)과 양(+),의 반대부호 결과를 반복성 지표가 AI노출 지표의 거울반사이기 때문이라고 말할 수는 없다. 오히려 반복성 지표와 AI 노출지수는 노동시장의 서로 다른 관점을 비춘다. 한편 저인지 직업군에서는 사후 계수가 0에서 유의하지 않아, 아무런 효과가 없었다. 이는 AI가 육체를 많이 쓰는 직업군에는 영향이 없을 것이라는 상식적인 예상과 합치한다.

이 글의 구성은 다음과 같다. 이어지는 ‘데이터 구축’에서는 종속변수인 고용량과 대규모 언어모델로 채점한 반복성 점수 및 인지 점수, 그리고 또 하나의 처치 측인 AI 노출지수를 설명한다. ‘이중차분 분석’에서는 직업군별로 상이한 고용량의 장기 추세를 제거해내는 2단계 연속 처치 이중차분 모형을 세운 뒤, 분석 결과를 세 방향으로 나누어 제시한다. 먼저 반복성 강도를 처치로 하여 인지 수준에 따라 사후 계수의 부호가 갈리는 양상(부호 반전)을 보이고, 이어 같은 설계에 AI 노출지수를 처치로 넣어 고인지 직업군에서 음(-)의 사후 계수가, 저인지 직업군에서는 유의하지 않음을 보인다. ‘강건성 검사’에서는 wild cluster bootstrap으로 반복성 강도와 AI 노출지수 두 축의 핵심 결과가 추론 방법을 바꾸어도 강건하게 유지됨을 보인다. 끝으로 ‘결론’에서 발견과 그 함의를 정리한다.

제III장. 데이터 구축

본 연구에서 사용된 모든 데이터와 STATA 코드는 아래 각주의 링크에서 공개한다.⁷

제1절. 종속변수

본 연구의 종속변수 후보는 취업자수(고용량), 시간당임금, 신규 진입자수 세 가지이며, 모두 미국 인구현황조사(Current Population Survey, 이하 CPS)의 IPUMS 추출 자료에서 구축했다(Flood et al. 2025). 한국 통계청의 경제활동인구조사를 쓰지 않고 미국 CPS를 택한 데에는 직업 분류의 정밀도라는 분명한 이유가 있다. 경제활동인구조사는 한국표준직업분류(KSCO)의 1자리 대분류 수준만 공표하여 직업군을 충분히 세분화하지 못한다. 반면 미국 CPS는 센서스 직업코드(OCC)를 4자리까지 제공하므로, 직업군 단위의 장기 추세와 처치 강도를 훨씬 정밀하게 식별할 수 있다. 직업군별 추세 이질성을 명시적으로 다루는 본 연구의 설계에서 이 차이는 결정적이다. 본 연구의 분석 단위는 이 4자리 수준의 OCC 직업군이며, 분석 표본은 2003년 3월부터 2026년 4월까지의 월별 직업군 패널이다.

세 후보 가운데 본문에서 실제로 사용하는 종속변수는 취업자수 뿐이다. 시간당임금과 신규 진입자수를 종속변수로 둔 이중차분에서는 유의한 변화가 전혀 관측되지 않고 경로가 평탄하게 나타났다. 두 결과 역시 그 자체로 보고할 가치가 있으나, 논문이 지나치게 길어지는 것을 막기 위해 본문 서술에서는 제외하고 취업자수에 집중한다.

제2절. 설명변수

가. 설명변수 개요

이중차분의 설명변수는 반복과업점수(routine score), 인지과업점수(cognitive score), 그리고 AI 노출지수 세 가지다. 본 연구의 설계에서 반복과업점수와 AI 노출지수는 각각 별개의 연속형 처치 강도(treatment intensity)로 투입되어 서로 다른 과업 구조의 축을 비추며, 인지과업점수는 두 분석 모두에서 표본을 인지 수준에 따라 두 그룹으로 나누는 기준으로 쓰인다.

세 변수는 모두 미국 표준직업분류(Standard Occupational Classification, 이하 SOC)의 6자리 수준에서 구축되었다. 종속변수인 취업자수는 센서스 직업코드(OCC) 체계를 따르므로, OCC와 SOC의 두 체계를 잇는 미국 센서스국의 OCC2010 - SOC2010 연계표를 이용해 설명변수를 OCC 체계로 변환했다. 하나의 OCC 직업에 여러 SOC 직업이 대응하는 경우에는 그 점수들을 단순평균하여 OCC 수준의 값으로 집계했다.

반복과업점수는 1에 가까울수록 정형적·자동화 가능한 과업(routine)을, 0에 가까울수록 예상 밖의 대처가 요구되는 비정형 과업(non-routine)을 가리킨다. 인지과업점수는 1에 가까울수록 정신적 처리가 중심인 인

⁷ <https://github.com/jayjeo/Effects-of-Routine-Intensity-and-AI-Exposure>

지적 과업(cognitive)을, 0에 가까울수록 손과 신체를 쓰는 육체적 과업(manual)을 가리킨다. 두 점수의 정확한 정의와 구축 방법은 이어지는 ‘인지과업점수와 반복과업점수’에서 다룬다. AI 노출지수는 Eloundou et al. (2024)이 대규모 언어모델(LLM)의 노동시장 영향 잠재력을 직업별로 측정된 노출 지표로, 값이 클수록 해당 직업이 LLM에 더 크게 노출됨을 뜻한다. 다만 이 지표는 반복과업점수 및 인지과업점수와 측정 정의와 대상이 다르므로, 정확히 무엇을 대리(proxy)하며 두 점수와 어떻게 구별되는지는 뒤의 ‘OpenAI의 AI-노출지수’에서 따로 다룬다.

나. 인지과업점수와 반복과업점수

중간 숙련 직업의 고용 감소를 설명해 온 반복편향 기술변화(RBTC) 문헌에서 핵심 변수는 직업의 반복 과업강도(Routine Task Intensity, 이하 RTI)인데, 이를 어떻게 측정하느냐가 분석 결과를 크게 좌우한다 (Walo 2023). Walo(2023)는 기존 RTI 측정법을 약 여섯 가지로 정리하며, 가장 널리 쓰이는 Autor, Levy, and Murnane (2003)의 직업명 사전(DOT)⁸ 기반 측정을 비롯해 대부분이 소수의 변수를 골라 결합하는 방식임을 보인다. 그런데 이런 방식은 반복성(routine)의 핵심인 반복 그 자체(repetitiveness)를 충분히 담지 못한다는 비판을 받아 왔다(Haslberger 2022). 본 연구는 이 한계를 극복하기 위해 ONET(National Center for ONET Development 2023)이 제공하는 직업별 과업 기술(task description)을 대규모 언어모델(LLM)로 직접 해석해 반복성 점수를 매기는 새로운 측정을 사용한다. 이는 저자가 아는 한 RTI 측정에 LLM을 적용한 첫 시도이며, 같은 방식으로 인지성(cognitive) 측정치까지 함께 구축한다. 이러한 방법의 큰 장점은 극도로 세분화된 (SOC 6자리) 직업군별로 점수를 구축했다는 점이다. 기존 문헌들에서 사용한 방식은 중분류나 소분류 수준에서의 점수인 경우가 많다.

구축 절차는 다음과 같다. 각 직업(SOC 6자리)에 속한 다수의 과업 기술들에 대해서, LLM이 과업 하나 하나의 반복성을 0에서 1 사이 연속값으로 평가하게 한다. 1은 인간의 재량이나 예측 불가능한 판단 없이 완전히 자동화할 수 있는 정형 과업에, 0은 인간의 창의성·직관이나 예외 상황 대응이 필요해 현재 기술로 자동화하기 어려운 비정형 과업에 부여되며, 0.5 부근은 양쪽 성격이 섞여 한쪽으로 분류하기 모호한 과업에 둔다. 인지성도 같은 방식으로 1(인지적)에서 0(육체적)까지 평가한다. 이렇게 얻은 과업 점수를 직업 단위로 평균하여 직업별 반복·인지 과업점수를 각각 얻되, O*NET이 핵심 과업(core task)으로 분류한 항목에는 보조 과업(supplementary task)으로 분류한 항목의 다섯 배에 해당하는 가중치를 부여한다. LLM에 부여한 정확한 채점 지시문과 한 과업에 대한 채점 예시는 부록 B에 수록한다.

LLM 출력의 신뢰성을 보완하기 위해 본 논문에서 이 작업을 수행할 당시에 성능이 가장 앞섰던 두 모델, OpenAI의 GPT-4와 Anthropic의 OPUS-3을 각각 병렬로 쓰고 두 점수의 평균을 최종 값으로 삼는다. 두 모델의 평가를 비교하면 상당수 과업에서 점수가 비슷해 점들이 강한 양의 상관관계를 이루지만, 적지 않은 과업에서는 두 모델의 판단이 갈리기도 한다. 두 모델 결과의 산포도는 부록 A에 있다. 이렇게 갈리는

⁸ DOT(Dictionary of Occupational Titles, 직업명 사전)는 미국 노동부가 1939년부터 발간한 직업정보 자료집으로, 각 직업의 수행 직무와 요구 기능·작업 조건 등을 표준화하여 기술하였다. 1991년 개정판을 마지막으로, 1990년대 후반 O*NET으로 대체되었다.

과업은 대개 심지어 사람이 봐도 관점에 따라 반복·비반복 판단이 달라지는 경계 사례인데, 본 연구는 이 불일치를 인위적으로 제거하지 않고 두 점수를 그대로 평균한다. 평균을 취하면 그 값이 0.5 부근으로 모여 과업 기술 자체의 모호성을 자연스럽게 반영하기 때문이다.

이렇게 구축한 두 점수로 직업을 펼쳐 보면, 직업들은 반복성과 인지성이라는 두 축을 따라 그림 1과 같이 반복-인지, 반복-육체, 비반복-인지, 비반복-육체의 네 영역에 흩어진다. 그림 2와 같이 SOC 2자리 대분류 수준에서 보면 관리·법률 직군은 비반복이면서 인지적인 영역에, 생산·건설 직군은 반복이면서 육체적인 영역에 자리 잡아, 두 점수가 그려 내는 직업 지형이 직관과 부합한다.

그림 1 SOC 6자리 직업의 인지-반복 과업점수 분포

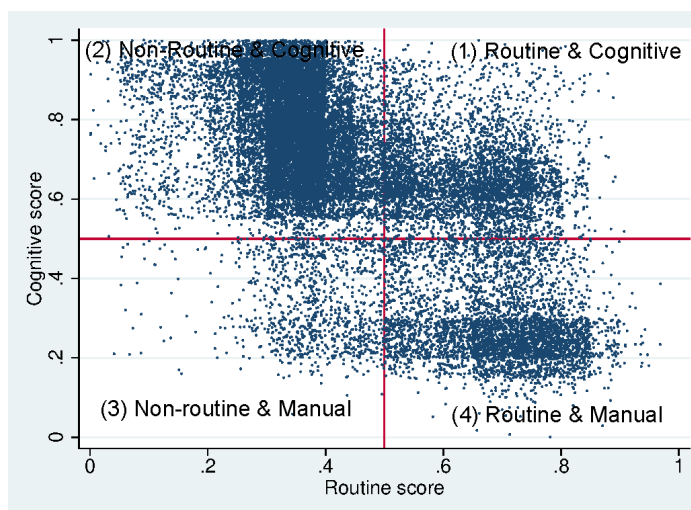


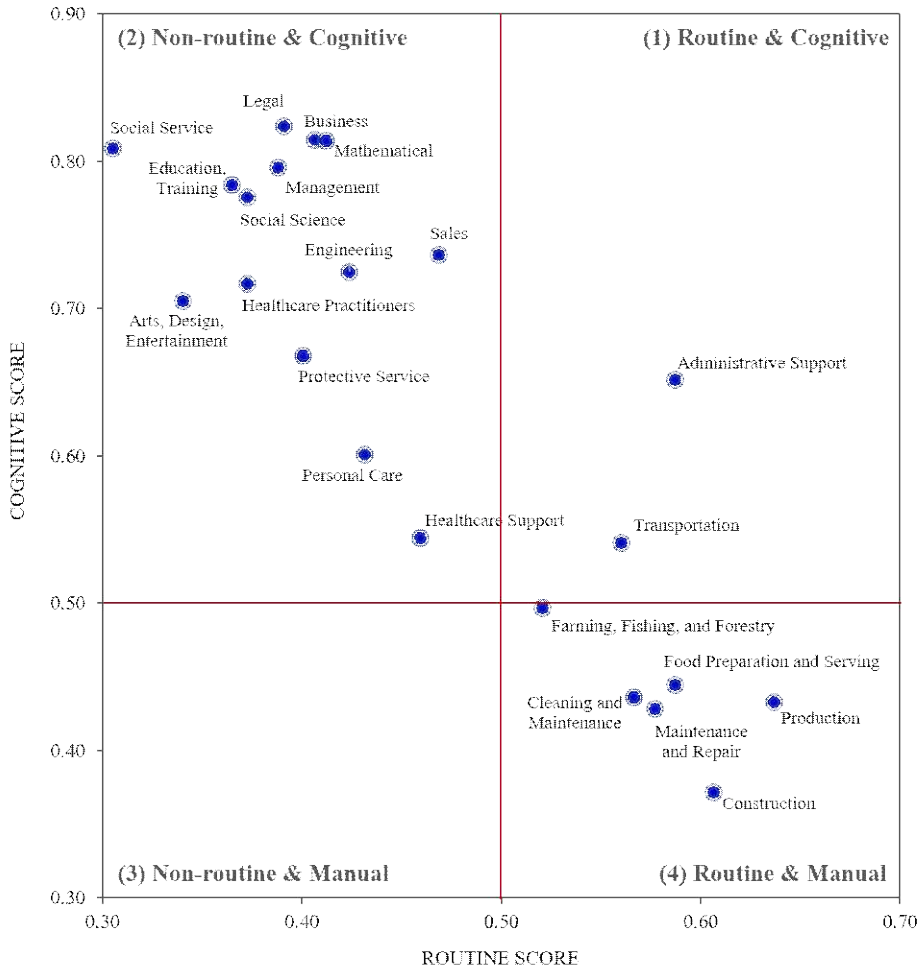
표 1 반복과업점수와 기존 RTI 측정치의 상관관계

| | 고용비중 | RTI Autor | RTI Acemoglu | RTI Goos | RTI Dengler | RTI Marcolin | RTI 본 논문 |
|--------------|-------|-----------|--------------|----------|-------------|--------------|----------|
| 고용비중 | 1.00 | | | | | | |
| RTI Autor | -0.35 | 1.00 | | | | | |
| RTI Acemoglu | -0.42 | 0.38*** | 1.00 | | | | |
| RTI Goos | -0.35 | -0.08 | 0.68*** | 1.00 | | | |
| RTI Dengler | -0.50 | 0.32*** | 0.58*** | 0.68*** | 1.00 | | |
| RTI Marcolin | -0.33 | 0.05 | 0.59*** | 0.73*** | 0.44*** | 1.00 | |
| RTI 본 논문 | -0.46 | 0.37*** | 0.76*** | 0.71*** | 0.64*** | 0.66*** | 1.00 |

본 논문에서 측정된 반복성 점수가 타당한지는 기존 RTI 측정치들과의 상관관계로 가늠할 수 있다. Walo(2023)가 정리한 상관 행렬에 본 연구의 LLM 기반 반복과업점수를 추가해 보면, 이 점수는 Autor 등(0.37), Acemoglu(0.76), Goos 등(0.71), Dengler(0.64), Marcolin(0.66) 등 서로 다른 방법으로 만든 기존

RTI 측정치들과 일관되게 유의한 양의 상관을 보이고, 1980~2005년의 고용 비중과는 음의 상관(-0.46)을 보인다. 이는 개념적으로 같은 대상을 겨냥한 측정치들이 서로 수렴한다는 수렴타당도(convergent validity)의 증거다(Walo 2023). 표 1의 상관 행렬의 자세한 해석은 부록 D에 둔다.

그림 2 SOC 2자리 대분류의 반복·인지 과업점수



LLM을 측정 도구로 삼는 접근은 경제학계에서 설명 불가능성(블랙박스)과 재현 불가능성(non-reproducibility)을 이유로 경계의 대상이 되기 쉽다. 본 연구는 다섯 가지 방법으로 이를 완화한다. 첫째, 본 연구에서 사용한 LLM은 단순한 웹 기반 챗봇 인터페이스가 아니라 API를 통해 호출하는 Python 프로그램이다. 이를 통해 모델의 동작을 결정하는 여러 파라미터를 직접 제어할 수 있는데, 그중 temperature를 0으로 설정하여 출력의 재현성을 극대화하였다. Temperature는 출력의 무작위성을 조절하는 파라미터로, 값이 높을수록(1) 답변이 다양하고 창의적인 대신 동일한 입력에도 실행할 때마다 결과가 달라지며, 0에 가까울수록 항상 확률이 가장 높은 응답을 선택하여 사실상 결정론적인 출력을 산출한다. 둘째, LLM은 각 과업에 점수만이 아니라 그 판단의 근거(reasoning)를 함께 출력하도록 Python을 프로그래밍하여, 모든 점수에 대해 사후에 검토할 수 있는 기록이 남아 판단 과정이 완전한 블랙박스에 머물지

않는다. 셋째, 채점에 사용한 지시문과 23,825개 과업의 점수·근거 전체를 공개하여 제3자가 점수와 그 근거를 직접 검토·재분석할 수 있게 한다. 넷째, 앞서 본 기존 RTI 측정치들과의 수렴타당도는 이 점수가 자의적 산물이 아님을 시사한다. 다섯째, 서로 독립된 두 모델의 평균은 단일 모델의 변동성을 줄인다. 그럼에도 한계는 분명하다. LLM을 사용한 채점 방식은 점수 판단을 전적으로 LLM에 위임한다는 점에서 신뢰성 우려가 남으며, 현재의 기술 수준을 기준으로 반복성을 판정하므로 2000년대 이전 시기의 반복성을 측정하는 데에는 적합하지 않다. 다만 본 연구의 관심이 생성형 AI 확산기의 최근 반복성에 있으므로, 이 한계는 본 연구의 목적과 충돌하지 않는다.

다. OpenAI의 AI-노출지수

AI 노출지수는 Eloundou et al. (2024)이 구축한 지표다. 이 지표에서 ‘노출(exposure)’의 정의는 대규모 언어모델(LLM) 또는 LLM 기반 소프트웨어가, 과업의 품질을 최소한 유지하거나 높이면서, 어떤 과업의 완수에 드는 시간을 최소 50% 줄일 수 있는 능력으로 정의된다. 연구진은 O*NET 데이터베이스의 직업별 과업을 사람과 GPT-4가 각각 수행하게 한 뒤, 한 직업에 속한 과업 가운데 ‘노출’된 과업의 비중을 직업 수준으로 집계하여 AI 노출지수를 얻는다.⁹

여기서 분명히 해 둘 점은, AI 노출지수 자체가 측정하는 것은 어떤 직업이 LLM에 의해 영향을 받을 잠재적 가능성이지, 실제 AI 도입의 정도나 그 영향이 고용 증가(보완)로 갈지 감소(대체)로 갈지의 방향까지 지표값에 담겨 있지는 않다는 것이다. 노출이 큰 직업의 고용이 실제로 어느 방향으로 움직였는지는 뒤의 이중차분 분석이 실증적으로 답할 질문이다. 또한 이 노출은 본 연구의 반복·인지 과업점수와 측정 대상이 다르다. 반복과업점수가 과업의 규칙화·자동화 가능성을, 인지과업점수가 과업의 정신적·육체적 성격을 재는 데 비해, AI 노출지수는 LLM이 시간을 절약해 줄 수 있는 잠재력을 잴다. 그 결과 노출은 텍스트·코드 생성이 많은 직업이나 고학력·고임금 직업에서 높게 나타나 반복성과는 결을 달리한다. 세 변수 사이의 실제 상관관계는 뒤의 AI 노출 측 분석에서 수치로 제시한다.

제III장. 이중차분 분석

본 장은 앞서 구축한 반복·인지 과업점수를 이용해 본 연구의 핵심 질문에 답한다. 그 질문은 생성형 AI가 확산하기 시작한 2023년 전후를 지나며 직업의 반복성 강도와 고용량의 관계가 직업의 인지 수준에 따라 어떻게 달라졌으며, 그 차이를 무엇으로 분해해 읽을 수 있는가이다. 이를 위해 반복성 강도를 연속적 처치로 삼고, 직업군별 고용량의 장기 추세를 제거한 뒤, 시점별 이중차분의 계수값을 추정하고, 인지 점수를 기준으로 표본을 나누어 그 계수값이 인지 수준에 따라 어떻게 갈리는지를 살핀다. 이어 같은 설계에 AI 노출지수를 처치로 넣어 고인지 직업군에서 음(-)의 사후 계수가 나타남을 보인 뒤, 두 축의 결과를 중

⁹ 노출 정도는 LLM에 직접 접근하는 경우(E1)부터 그 위에 추가 소프트웨어가 결합되어야 노출되는 경우(E2)까지 단 계적으로 정의된다. 그 후 $E1 + 0.5 \cdot E2$ 로 가중계산하여 최종 AI 노출지수를 구축한다.

합하여 측정 축에 따른 결과의 이질성을 논한다.

제1절. 인지과업점수와 반복과업점수

본 연구는 가장 중요한 변수로 인지과업점수(cognitive score)와 반복과업점수(routine score)를 사용한다. 두 점수가 개념적으로 구분되는 별개의 차원이라는 점이 중요하다. 반복성은 과업이 규칙화·자동화될 수 있는 정도를, 인지성은 과업이 정신적인지 육체적인지를 측정하므로, 한 직업은 반복(1)-인지(1), 반복(1)-육체(0), 비반복(0)-인지(1), 비반복(0)-육체(0) 라는 네 조합 중 어디에나 놓일 수 있다(Autor, Levy, and Murnane 2003). 예컨대 단순 사무직은 반복적이면서 인지적이고, 조립 공정은 반복적이면서 육체적이며, 관리·연구직은 비반복적이면서 인지적이다. 이하 본문에서는 두 점수를 맥락에 따라 인지 점수나 반복성 강도로 줄여 부르기도 한다.

제2절. 직업군별 고용 추세의 이질성

본 연구가 주목하는 생성형 AI는 어느 한 시점에 갑자기 출현한 것이 아니라, 2022년 말 대화형 모델의 공개를 기점으로 그 이후 여러 해에 걸쳐 점진적으로 보급되었다. 그 이전 시기(2003~2022)는 생성형 AI가 대중적으로 보급되기 전 기간이므로, 본 설계는 이 구간을 처치 이전(미처리) 구간으로 두고 1단계 고용량의 추세 제거에 사용한다. 모형에서 처치 개시 시점은 2023년 1월로 두어 모든 직업군에 공통으로 적용되, 이는 효과가 그 시점에 한꺼번에 발생하는 단절점이 아니라 생성형 AI 보급이 본격화한 기준점을 뜻한다. 보급이 처치 이후부터 점진적으로 진행되는 만큼, 처치 반응을 한 시점에 묶는 단일 사전·사후 더미 대신, 시점(k)마다 별도의 계수 β_k 를 두는 이벤트 스테디로 추정하여, 반복성과 고용의 관계가 처치 이후 어떻게 전개되는지를 동태적 경로로 관찰한다. 분석 자료는 2003년 3월부터 2026년 4월까지의 월별·개인별 패널이며, 인지 점수(cognitive score) 0.5를 기준으로 표본을 둘로 나누어, 0.5 미만인 저인지 직업군과 0.5 이상인 고인지 직업군에 대해 동일한 모형을 각각 적용한다.

처치는 처치군과 통제군으로 나뉘지 않고 연속 강도로 주어진다. 따라서 식별은 처치 유무에 의한 평균치 처치효과(Average Treatment Effect on the Treated, 이하 ATT)가 아니라, 각 시점 안에서 직업군 간 반복성 강도가 다르다는 점(횡단 변이)에 의존한다. 즉 같은 달에 반복성 강도가 높은 직업군과 낮은 직업군의 고용량을 공통의 시간 충격을 통제한 채 비교한다.

이중차분의 식별가정은 처치가 없었다면 처치군과 비교군이 평행하게 움직였으리라는 조건부 평행추세이다. 그런데 20여 년의 직업군 패널에서 고용량은 직업군마다 장기 추세가 뚜렷이 다르다. 어떤 직업군은 추세적으로 늘고 어떤 직업군은 준다. 따라서 고용량을 그대로 쓰면 두 가지 문제가 발생한다. 우선 이중차분은 처치군과 대조군이 시간고정효과와 패널고정효과로 걸러낸 나머지가 사전적으로 평행해야한다. 하지만 직업군 별로 고용량의 추세가 다른 것은 연속시간과 패널더미의 교차효과이므로 이중차분의 두 가지 고정효과로 걸러내지 못한다. 따라서 이질적인 추세가 존재하는 한 사전평행추세가 성립할 수 없다. 둘째, 이중차분은 처치 이전의 추세 차이를 처치효과로 귀속시켜 반복성 강도의 사후 효과를 편향(biased)되게 추정한다. 그러므로 효과를 제대로 식별하려면 직업군별 고용량 추세를 먼저 제거하고, 추세를 걷어낸 뒤

남는 고용량만 분석해야 한다. 이러한 방법을 회귀모형으로 명시적으로 쓰는 접근은 변위 노동자 연구 (Jacobson et al. 1993)¹⁰와 일방적 이혼법 연구(Friedberg, 1998)¹¹에서 확립되었고, 이중차분 이벤트 스타디(Two-way fixed effect event study)에서 인정되는 식별 전략의 하나로 정립되어 있다.¹²

가. 1단계: 직업군별 고용량 추세 제거

직업군별 추세는 1단계 회귀로 제거한다. 직업군을 i , 시간을 t ¹³, 고용량을 Y_{it} 라 하면 다음을 추정한다.

$$Y_{it} = \alpha_i + \delta_i t + \gamma_i t^2 + \epsilon_{it}, \quad \text{where } t < 2023\text{년}1\text{월}$$

여기서 α_i 는 직업군 고정효과, $\delta_i t$ 는 직업군별 선형 추세, $\gamma_i t^2$ 는 직업군별 2차 추세이다. 핵심은 이 회귀를 처치 개시 이전(2023년 1월) 구간만으로 적합한다는 점이다. 처치 이후 자료를 섞으면 추세계수가 처치 반응을 일부 빨아들이기 때문이다. 적합한 추세를 원본 고용량에서 빼면 추세를 제거한 고용량 \widetilde{Y}_{it} 를 얻는다. 이는 잔차(residual)를 그대로 사용하는 방법과 약간 차이가 있는데, 더 강건하다.

$$\widetilde{Y}_{it} = Y_{it} - (\widehat{\alpha}_i + \widehat{\delta}_i t + \widehat{\gamma}_i t^2)$$

직업군별 선형 및 2차 추세를 수준회귀에 직접 넣는 처리방법은 주별 추세를 쓴 Friedberg (1998)의 선례와 같은 계열이다.¹⁴ 한편 이 추세를 전체 표본이 아니라 미처리 구간만으로 적합하는 점은, 오차항(error term)을 미처리 관측치로만 추정한다는 현대 이중차분의 원리와 부합한다.¹⁵ \widetilde{Y}_{it} 는 직업군 고유의 장기 동학을 걸어낸 고용량으로, 공통의 시간 충격과 반복성 강도에 따른 차등 반응만 남긴다.

¹⁰ 단위별 선형추세를 식별 장치로 사용한 대표적 선례.

¹¹ 주별 선형·2차 추세를 수준회귀에 직접 포함하는 명세의 선례, Eq.(3) 및 fn.10 참조.

¹² Freyaldenhoven et al. 2021. 단위별 추세를 포함한 식별 전략 및 긴 패널에서 선형추세 가정의 강도에 관한 논의 (p.16) 참조.

¹³ 표본 시작월을 원점(1)으로 둔 연속변수이다. 더미변수가 아니다.

¹⁴ 주별 선형·2차 추세를 수준회귀에 직접 포함하는 명세의 선례, Eq.(3) 및 fn.10 참조.

¹⁵ Gardner et al. (2022): 미처리 관측치를 이용한 1단계 적합과 2단계 잔차 회귀의 구조, 생성된 종속변수에 따른 2단계 표준오차 조정과, 타당한 표준오차를 얻는 가장 단순한 방법으로서 두 단계를 매 복제에서 재추정하는 부트스트랩 및 그 대안인 GMM 동시추정을 제안.

Borusyak et al. (2024): 미처리 관측치만으로 단위·시간 고정효과를 적합하는 귀속 추정량, 비이분 처치의 이론적 수용, 모든 단위가 처치된 시점에서 시간 효과와 처치효과 분리 불가 및 never-treated 집단이 비식별을 방지한다는 논의(Simultaneous Treatment) 참조.

나. 생성된 종속변수 문제

2단계는 \widetilde{Y}_{it} 를 이벤트 스터디 형태의 처치 항에 회귀한다.

$$\widetilde{Y}_{it} = \zeta_i + \lambda_t + \sum_{k \neq k_0} [\beta_k \cdot D_i \cdot \mathbf{1}[t = k]] + \eta_{it}$$

여기서 D_i 는 시간 불변의 연속 처치강도, 곧 각 직업군의 반복성(routine) 강도이며, λ_t 는 시간 고정효과, ζ_i 는 직업군 고정효과(1단계의 α_i 와 구별된다), $\mathbf{1}[t = k]$ 는 시점 더미, $k_0 = 2022$ 년 12월은 기준시점으로 처치개시 시점인 2023년 1월의 직전월이다. 계수 β_k 는 반복성 강도가 1단위 높을 때 기준시점 대비 고용량의 경로가 얼마나 달라지는지를 나타내는 용량-반응 계수(dose-gradient)이다. 이하 본 논문에서 dose-gradient는 모두 이 시점별 용량-반응 계수를 가리키며, 결과를 서술할 때 사후 계수가 양(+) 또는 음(-)이라 함은 이 계수의 부호를 뜻한다.

여기서 한 가지를 짚어야 한다. \widetilde{Y}_{it} 는 1단계에서 추정된 계수로 만든 생성된 종속변수이다. 그 계수에는 표본 변동이 있다. 그런데 \widetilde{Y}_{it} 를 관측된 변수처럼 다뤄 2단계에 통상적 표준오차(군집-강건 표준오차(cluster robust standard error))를 적용하면, 1단계 추정오차가 반영되지 않는다. 그 결과 추론이 일치하지 않고, 대개 불확실성을 과소추정한다(Pagan 1984). Gardner (2022)도 2단계 이중차분에서 같은 문제를 지적한다. Gardner가 제시하는 가장 단순한 해법은 두 단계를 매 복제마다 모두 다시 추정하는 부트스트랩이고, GMM 동시추정은 그 대안이다. 1단계와 2단계 모형이 옳다면 β_k 의 점추정은 일치성(consistent)이 있지만, \widetilde{Y}_{it} 의 생성과정을 무시한 2단계 표준오차는 불일치(inconsistent) 하다.

다. 2단계 추론: 직업군 군집 부트스트랩

타당한 추론을 위해 전체 2단계 절차를 직업군 군집 수준에서 부트스트랩(bootstrap)한다. 각 복제에서 직업군을 복원추출로 다시 뽑고, 1단계 추세회귀와 \widetilde{Y}_{it} 재구성, 2단계 이벤트 스터디를 모두 다시 추정한다. 충분한 복제(본 연구는 999회)에 걸친 β_k 의 분포로 정확한 표준오차를 계산한다.

매 복제가 1단계를 다시 돌리므로 추세 추정의 불확실성이 2단계 추론에 그대로 반영된다. 요컨대 본 절차는 부트스트랩을 Gardner가 원하는 귀속 방식(imputation)이 아니라 추세 제거에 적용한 형태이다. 관측치가 아니라 직업군 단위로 다시 뽑으므로 직업군 내 시계열 상관도 해결한다. 추세가 있는 상황에서는 이러한 시계열 상관이 표준오차를 크게 키울 수 있어 군집 추론이 특히 중요하다(Lee et al. 2011).¹⁶ 복원추출에서는 같은 직업군이 중복 추출될 수 있다. 이때 재표본된 직업군마다 고유 식별자를 부여해, 고정효과 추정이 중복분을 별개 단위로 다루도록 한다. 신뢰구간은 군집 수가 많지 않은 점을 감안해 군집 수에

¹⁶ 동적 처치효과와 단위별 추세가 결합될 때 잔차 계열상관으로 군집 보정 표준오차가 크게 증가할 수 있음을 발견.

서 1을 뺀 자유도의 t-분포로 소표본 보정을 적용한다. 군집-강건 추론의 접근이 관측치 수가 아니라 군집 수에 의존하기 때문이다(Cameron et al. 2015).¹⁷

라. 해석상의 유의점

이 설계에는 세 가지 유의점이 있다. 첫째, 처치가 연속이고 시간불변이므로 추정된 계수는 단순한 평균처리효과(ATT)가 아니라 dose-gradient, 곧 반복성 강도에 대한 한계효과의 가중평균에 가깝다. 다만 Callaway, Goodman-Bacon, and Sant'Anna (2024)가 지적했듯, 일반적인 양방향 고정효과(Two-way Fixed Effect) 모형을 연속 처치에 적용하면 계수를 ATT로 해석할 때는 '음의 가중치' 문제가 발생하고, 한계효과로 해석할 때조차 집단 간 효과 이질성에 따른 '선택 편의(selection bias)'가 발생한다. 특히 미처치(never-treated) 집단이 없는 경우 이러한 편의는 더욱 복잡해진다. 따라서 이 효과를 단순한 '처치 대 미처치'의 효과로 읽기보다, 고반복·저반복 직업군에서 반복성 강도가 1단위 증가할 때 나타나는 평균적인 한계 반응으로 주의 깊게 해석해야 한다.

둘째, 직업군 군집 수가 모수 수에 비해 넉넉하지 않다. 군집이 적으면 군집-강건 분산이나 짝(pairs) 부트스트랩이 표준오차를 작게 잡을 수 있다(Cameron et al. 2015, Pustejovsky and Tipton 2018). 이를 보완하기 위해 Roodman et al. (2019)이 고안한 wild cluster bootstrap을 별도로 적용하며, 그 결과는 뒤의 강건성 검사에서 다룬다.

셋째, 1단계는 미처리 구간(2003~2022)에서 적합한 고용량 추세를 처치 이후로 외삽한다. 추세가 유연할수록 처치 이후 경로의 일부를 추세로 잘못 돌릴 위험이 커진다. 추세 명세가 처치 동태와 섞일 수 있다는 경고는 사후 자료까지 추세 적합에 넣은 설계에서 이미 제기되었고,¹⁸ 본 설계는 적합을 미처리 구간으로 제한해 이를 완화하나, 문제를 완전히 없애지는 못한다(Wolfers 2006, Meer and West 2016).

마. Gardner의 2단계 이중차분을 쓰지 않는 이유

본 논문에서 사용한 2단계 이중차분 구조는 길보기에 Gardner (2022)의 2단계 이중차분과 비슷하다. Gardner의 방법은 Borusyak·Jaravel·Spiess (2024)의 귀속 추정량처럼, 단위·시간 고정효과를 미처리 관측치로만 추정한 뒤 처치 관측치의 미처리 잠재결과를 귀속(impute)하고, 잔차를 처치변수에 회귀한다.¹⁹ 그러나 본 연구는 이 추정량을 쓰지 않는다. 이유는 연속 처치 때문은 아니다. 연속적인 강도의 처치는 Gardner의 방법에서도 이론적으로 구현 가능하기 때문이다.

¹⁷ 군집 수가 적을 때 군집-강건 추론의 하방 편의 및 군집 수 기반 접근 방법.

¹⁸ Wolfers 2006: 단위별 2차 추세가 처치 동학을 흡수하여 편의를 악화 또는 완화할 수 있음.

Meer and West 2016: 과가 수준이 아니라 성장경로에 작용할 때 추세 통제가 효과를 감쇠시킬 수 있음.

¹⁹ Borusyak·Jaravel·Spiess (2024): 미처리 관측치만으로 단위·시간 고정효과를 적합하는 귀속 추정량, 비이분 처치의 이론적 수용, 모든 단위가 처치된 시점에서 시간 효과와 처치효과의 분리 불가 및 never-treated 집단이 비식별을 방지한다는 논의(Simultaneous Treatment) 참조.

진짜 걸림돌은 공통 처치시작 시점의 사용과 미처치(never-treated) 집단의 부재가 겹친 데 있다. Gardner의 1단계는 시간 고정효과를 미처치 관측치로 추정해야 한다. 그런데 2023년 1월에 모든 직업군이 동시에 처치되므로, 처치 이후 시점에는 미처치 관측치가 하나도 없다. 따라서 처치 이후의 시간 고정효과를 1단계에서 식별할 수 없고, 처치 이후 잠재결과의 귀속도 불가능하다. Gardner의 식별은 미처치 표본에서 행렬이 가역²⁰일 것을 요구한다. 그런데 미처치 관측치가 모든 시점 더미를 포괄하지 못하면, 단위 고정효과와 시간 고정효과가 서로 분리 식별되지 않는다. 공통 처치시작 시점이라는 점은, 처치 이후의 시간더미에 대응하는 미처치 관측치가 전무하다는 의미이므로, 바로 이 가역성이 깨지게 된다. Borusyak·Jaravel·Spiess도 처치시점 이후에 모든 관측치가 전부 처치된 경우에는 시간 효과와 처치효과가 분리되지 않으며, 미처치 집단이 있어야 이 비식별을 막을 수 있다고 분명히 한다. 동적 처치효과와 시간 고정효과가 분리되지 않으면 기준더미 시점의 정규화만으로 부족하다는 지적도 같은 맥락이다 (Schmidheiny and Siegloch 2023).²¹

본 논문에서 사용한 모형은 시간 고정효과를 미처치 관측치로 귀속하지 않으므로 이 문제를 피한다. 1단계는 직업군별 추세와 수준만 제거하고, 시간 고정효과는 2단계로 넘겨 전 구간 자료로 dose-gradient와 함께 추정한다. β_k 는 각 시점 안의 반복성 강도의 횡단 변이 -즉, 시간 고정효과를 통제된 상태에서 고반복-저반복 직업군의 대비- 로 추정되므로, 처치 이후에 미처치 관측치가 없어도 추정이 가능하다. 이때 시간 고정효과가 모든 직업군에 공통인 사후 변화를 흡수하므로, 본 2단계 모형이 식별하는 것은 반복성 강도와 무관한 사후 평균처치효과(ATT)가 아니라 반복성 강도에 대한 용량-반응(dose-gradient) 계수다. 공통의 처치 개시시점에서 사후 평균처치효과 자체는 여전히 시간효과와 분리되지 않으나, 본 2단계 회귀 모형은 그 평균이 아니라 반복성 강도에 대한 가중평균된 한계효과를 겨냥한다. 요컨대 Gardner의 귀속 기반 2단계 이중차분(STATA 및 R의 패키지명 *did2s*)은 본 설계에 맞지 않다.

제3절. 분석 결과: 두 처치 축의 이중차분

가. 반복성 강도를 처치변수로 사용하는 경우

본 항의 결과를 읽기 전에 계수의 의미를 다시 분명히 해 둘 필요가 있다. 앞에서 서술했듯 2단계 추정의 종속변수는 직업군별 고용 추세를 제거한 고용량이며, 시점별 계수 β_k 는 원시 고용량의 레벨 차원의 변화

²⁰ 각 group과 period에 untreated and treated observations가 있어야 한다는 의미이다.

²¹ Schmidheiny and Siegloch (2023): 동적 처치효과와 추세적 시간 고정효과의 비분리 식별 및 단위별 추세 포함 시 추가 정규화 필요성 논의 참조.

Borusyak, Jaravel, and Spiess는 미처치(never-treated) 집단이 없을 때 동학적 처치효과 경로가 짐식별되지 않을 수 있음을 지적하며, Schmidheiny and Siegloch 역시 미처치 집단이 없는 경우 정규화만으로는 부족하고 추가 제약이 필요하다고 설명한다. 따라서 본 논문의 2단계 회귀는 사후 평균처치효과(ATT) 자체가 아니라, 전 표본의 시간 고정효과를 포함한 상태에서 반복성 강도에 대한 시점별 dose-gradient를 식별한다는 점에서 Gardner의 2단계 추정방법과 구별된다.

가 아니다. β_k 는 직업별 추세를 제거한 고용량에 대해서 반복성(routine) 강도 한 단위에 대한 상대적 용량-반응 계수이다. 따라서 β_k 가 음(-)이라는 것은 해당 시점에 반복성이 높은 직업군이 낮은 직업군보다, 그리고 사전 추세에서 외삽하여 추세를 제거한 고용량의 경로 대비, 상대적으로 더 부진했음을 뜻할 뿐 고용이 절대적으로 감소했음을 뜻하지 않는다. 마찬가지로 β_k 가 양(+)이라는 것도 절대적 증가가 아니라 추세 대비 덜 감소했거나 상대적으로 더 나아졌음으로 읽어야 한다.

그림 3 저인지 직업군의 반복성 강도 이벤트 스터디

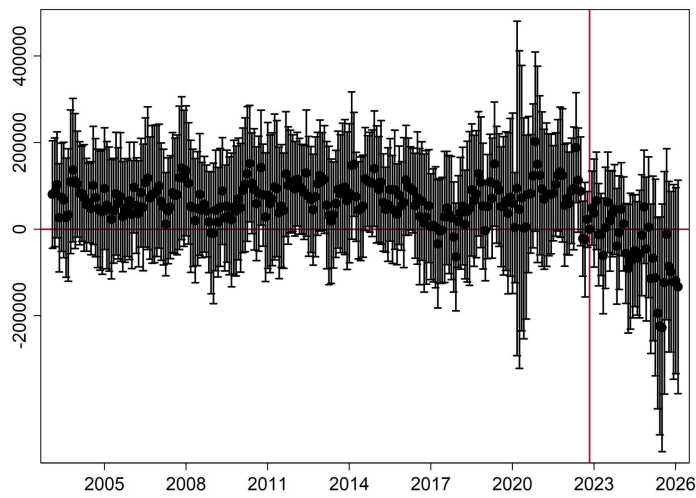
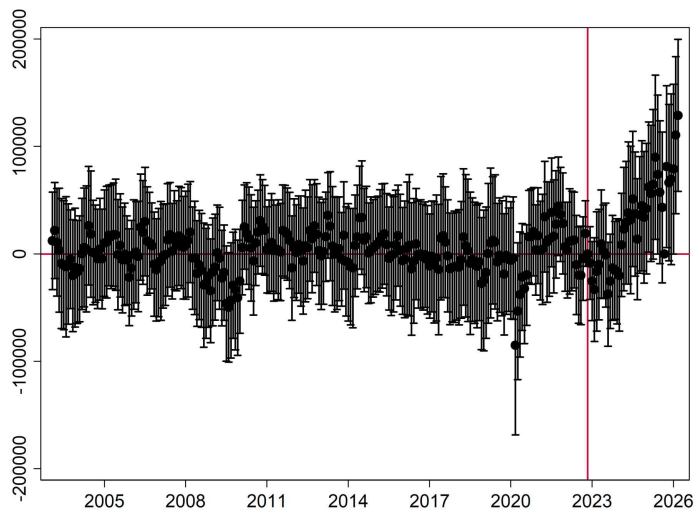


그림 4 고인지 직업군의 반복성 강도 이벤트 스터디



본 연구는 표본을 인지 점수(cognitive score)를 기준으로 둘로 나누어 분석한다. 인지 점수가 0.5 미만인 저인지 직업군(주로 육체 과업 중심)과 0.5 이상인 고인지 직업군(주로 두뇌 과업 중심)을 각각 별도로 추정하며, 두 집단의 반복성 강도의 용량-반응 계수 이벤트 스터디 경로와 95% 신뢰구간을 그림 3과 4에 제시한다. 두 그림 모두 수직선은 처치 개시(2023년 1월)를 의미한다.

먼저 두 집단 모두 사전 평행추세가 대체로 충족된다. 처치 이전(2003~2022) 구간에서 계수들은 0 근방에서 작게 등락하며 신뢰구간이 거의 모든 시점에서 0을 포함한다. 점추정의 등락폭은 저인지 집단에서 대략 $\pm 150,000$ 명 이내, 고인지 집단에서는 더 좁아 대략 $\pm 50,000$ 명 이내로 후자의 사전 경로가 특히 평탄하다. 다만 사전 시기의 평행추세만으로는 이중차분의 가장 중요한 전제조건인 “만약에 사건이 일어나지 않았더라면 사전과 사후를 포함한 모든 시기의 평행추세가 성립해야한다”를 완벽하게 입증하지 못함에 유의해야한다(Roth 2022).²²

처치 개시 이후 두 집단의 경로는 뚜렷이 갈린다. 저인지 집단에서는 2023년 이후 계수가 점차 하락하여 2025~2026년에는 약 $-250,000$ 에서 $-350,000$ 명 수준에 이르고, 사후 말기 여러 시점에서 신뢰구간이 0에서 유의하게 다르다. 곧 저인지 직업군 안에서 반복성이 높은 직업일수록 그 시기에 (추세 대비) 상대적으로 더 부진했고, 뒤집어 말하면 덜 반복적인 직업이 상대적으로 호전되었다. 반대로 고인지 집단에서는 같은 시점 이후 계수가 상승하여 2025~2026년에는 약 $+100,000$ 에서 $+150,000$ 명 수준에 이르고, 역시 사후 말기에 신뢰구간을 보면 점추정값이 0이라는 귀무가설을 0.05 유의수준에서 기각한다. 곧 고인지 직업군 안에서는 반복성이 높은 직업일수록 상대적으로 호전되었다. 두 효과는 부호가 반대일 뿐 아니라 크기도 비대칭적이어서, 저인지의 음(-)의 효과가 고인지의 양(+)의 효과보다 대략 2~3배 크다.

저인지 집단의 음(-)의 사후 계수, 곧 반복적 육체 과업 직업의 상대적 고용 부진은 사실 새로운 현상이 아니라 반복편향 기술변화(RBTC) 문헌이 오랫동안 기록해 온 패턴의 연장으로 읽힌다. Autor, Levy, and Murnane (2003)이 정식화한 과업 기반 틀에서 반복적 과업은 규칙으로 코드화되어 기계로 대체되기 쉬우며, Autor and Dorn (2013)은 이 대체(substitution)가 중간 숙련의 반복적 직업을 잠식하고 노동시장을 양극화(polarization)해 왔음을 보였다. 로봇과 물리적 자동화의 고용 효과를 직접 추정한 연구들은 이 대체가 특히 반복적·육체적 직업에 집중됨을 일관되게 확인한다. Acemoglu and Restrepo (2020)는 미국에서 산업용 로봇이 반복적·조립·육체 직업의 고용을 줄였음을 보고하며, Dauth et al. (2017)도 독일 제조업에서 로봇에 의한 마찬가지로 고용 감소를 확인한다.

따라서 저인지·고반복 직업의 사후 부진은 생성형 AI의 충격이라기보다, 로봇·소프트웨어·물리적 자동화라는 이전부터 작동해 온 기체가 반복성 메커니즘을 따라 다시 표면화한 결과로 보는 편이 자연스럽다. 실제로 생성형 AI 노출을 측정한 연구들은 육체적 반복 직업의 AI 노출이 대체로 낮다고 보고하므로²³, 저인지 집단에서 반복적 직업일수록 고용이 부진한 이유를 AI 노출이 높아서이기 때문이라고 말할 수 없다. 여기에 2021~2022년 인력난 시기에 팽창했던 육체-반복 부문의 고용이 코로나 이후 다시 줄어든 현

²² Angrist and Pischke (2008)이 명쾌하게 설명하듯이, 이중차분이 인과적 효과를 입증하기 위한 핵심 전제조건은 처치 비발생시의 잠재결과(처치군 및 대조군 모두의 잠재적 비처치 결과)는 오직 시간고정효과, 패널고정효과, 통제변수들만으로 완벽하게 설명될 수 있어야 한다는 것이다. 완벽하게 설명된다는 뜻은, 잠재결과(potential outcome)를 시간고정효과, 패널고정효과, 통제변수들만으로 회귀분석하면 사전추세든 사후추세든 그 잔차가 반드시 0에서 유의하게 다르지 않아야 함을 의미한다.

²³ Eloundou et al. 2024; Webb 2019; Felten, Raj, and Seamans 2021.

상도 영향이 있을 수 있다. 요컨대 저인지 집단 결과의 결과는 전통적 RBTC와 비(非)AI 충격에 비추어 비교적 자연스럽게 해석되는 부분이다.

한편, 고인지 집단의 양(+)의 사후 계수는 그 원인을 AI로 돌릴 수 있을지가 분명하지 않다. 반복성이라는 척도는 과업의 규칙성 정도를 측정할 뿐이며, 생성형 AI가 해당 과업을 대신할 수 있는지와 일대일로 맞물리지는 않기 때문이다. 실제로 직업이 AI 발전에 얼마나 노출되는지를 계량한 지표들은 비반복적 두뇌 과업에서 도리어 높은 값을 보인다고 알려져 있다(Eloundou et al. 2024; Webb 2019; Felten, Raj, and Seamans 2021). 본 연구의 분석에서도 반복성 강도와 AI 노출강도는 전체 표본에서는 음의 상관을 보이지만, 표본을 고인지 집단으로 좁히면 둘 사이의 상관은 사실상 사라진다. 그렇기에 고인지 집단 안에서 AI 노출이 큰 직업은 고용이 줄고 반복성이 큰 직업은 고용이 늘는, 이 상반된 양상이 왜 나타나는지는 여전히 설명되지 않은 채 남는다. 두 강도가 고인지 집단 하에서 애초에 무상관인 이상, 반복성에서 나온 결과를 전적으로 AI 탓으로 못 박는 해석은 무리가 따르기 때문이다.

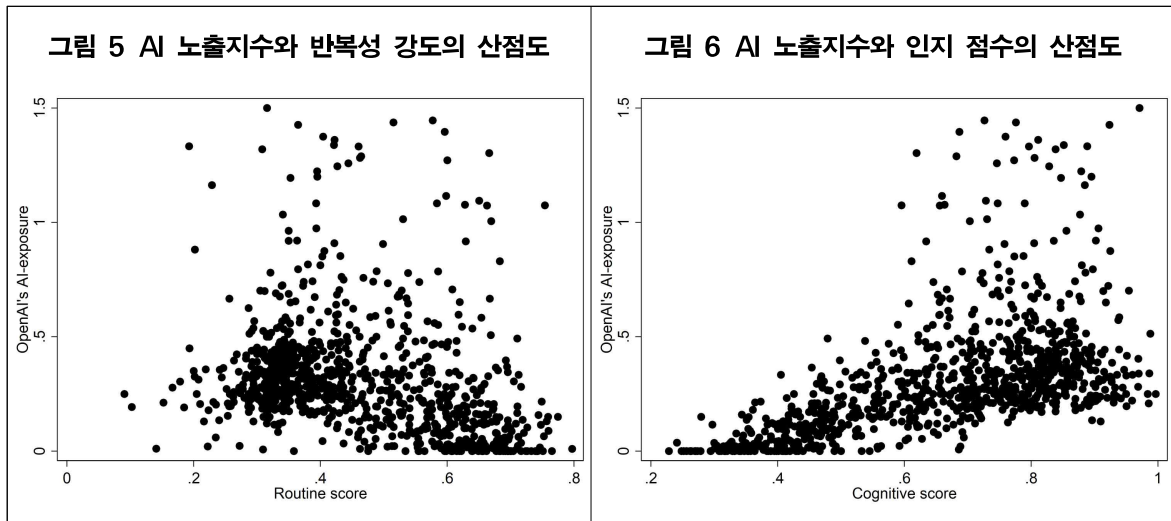
고인지 집단에서 반복성이 높은 직업의 고용이 증가했다는 사실은, 기존 RBTC가 제시해 온 작동 원리와는 정반대되는 현상이 AI가 본격 도입된 2023년을 기점으로 나타나기 시작했음을 보여준다. 이 변화가 AI 도입에서 비롯되었는지는 인과적으로 단언하기 어렵지만, 발생 시점만큼은 서로 들어맞는다. RBTC의 논리대로라면 집단이 고인지든 저인지든 반복성이 강한 직업은 자동화 기술에 밀려 고용이 감소해야 한다. 그런데 2023년 이후의 고인지 집단에서는 외려 반복성이 강한 직업이 입지를 넓히는, 예상과 어긋난 모습이 관찰된다.

이어지는 AI 노출지수 측 분석에서는 설명변수를 반복성 강도에서 AI 노출지수로 바꾸어 동일한 2단계 이중차분을 추정한다.

나. AI 노출지수를 처치변수로 사용하는 경우

본 항에서는 반복성 강도 대신 Eloundou et al. (2024)이 구축한 AI 노출지수를 동일한 2단계 이중차분의 처치 강도로 투입한 결과를 보고한다. AI 노출지수의 구축 방식은 앞의 데이터 구축 장에서 이미 정의했으므로 여기서는 분석에 필요한 핵심만 환기한다. 이 지표는 대규모 언어모델 또는 그 기반 소프트웨어가 어떤 직업의 과업 수행 시간을 품질을 유지한 채 절반 이상 줄일 수 있는 잠재적 능력을, O*NET 과업에 대한 평가를 직업 수준으로 집계해 측정한 것이다. 다만 이것은 특정 직업이 LLM에 노출될 잠재적 가능성의 예측치이지 실제 도입의 정도나 고용 증감의 방향을 재는 값이 아니며, 반복성 강도가 포착하는 과업의 규칙화·자동화 가능성과는 측정 대상 자체가 다르다.

두 측정치가 실제로 얼마나 다른지는 상관관계에서 드러난다. AI 노출지수와 반복성 강도의 상관계수는 -0.27 로, 음의 방향이기는 하나 절댓값이 크지 않은 약한 상관에 그친다. 그림 5의 산점도가 보이듯 두 변수의 관계는 상당히 분산적이어서, 반복성이 높은 직업이 일률적으로 AI 노출이 낮은 것도 그 역도 아니다.



반면 AI 노출지수와 인지 점수의 상관계수는 +0.54로 상당히 높은 양의 상관을 보인다. 그림 6의 산점도에서도 인지 점수가 높아질수록 AI 노출이 뚜렷하게 함께 높아지는 우상향 관계가 확인된다. 이는 AI 노출지수가 반복성 축보다 인지 축에 훨씬 가깝게 정렬되어 있음을 뜻한다. Eloundou et al. (2024) 자신도 이 노출이 텍스트·코드 생성이 많은 고학력·고임금의 비반복 두뇌 직업에서 높게 나타난다고 보고하며, 본 자료의 +0.54 상관은 이 특성을 직접 확인해 준다.

다만 위의 상관계수를 고인지와 저인지 그룹으로 제한해서 다시 측정하면 결과가 다르다. 고인지 하위표본 안에서 반복성 강도와 AI 노출지수의 상관은 약 +0.12, 저인지 하위표본 안에서는 약 -0.08로, 두 경우 모두 0에 가깝다.

AI 노출지수를 처치 강도로 삼으면 이중차분 결과의 모습이 달라진다. 추정치는 반복성 강도의 경우와 동일하게 직업군 추세를 제거한 고용량을 종속변수로 하고 999회 복제의 군집 부트스트랩으로 추론하되, 표본을 인지 점수로 두 그룹으로 나누어 본다. 고인지 직업군에서는 AI 노출지수의 사후 계수가 처치 이후 음(-)으로 내려가고, 사후 말기 여러 시점에서 신뢰구간이 0에서 유의하게 다르다. 곧 고인지 직업군 안에서 AI 노출이 큰 직업일수록 그 시기에 추세를 제거한 고용이 상대적으로 더 부진했다. 반면 저인지 직업군에서는 사후에도 계수가 0 근방에 머물고 신뢰구간이 0을 포함하여, 유의한 변화가 관측되지 않는다. 두 하위표본의 이벤트 스테디 경로와 95% 신뢰구간을 다음의 그림 7과 8에 제시한다.

그림 7 저인지 직업군의 AI 노출지수 이벤트 스터디

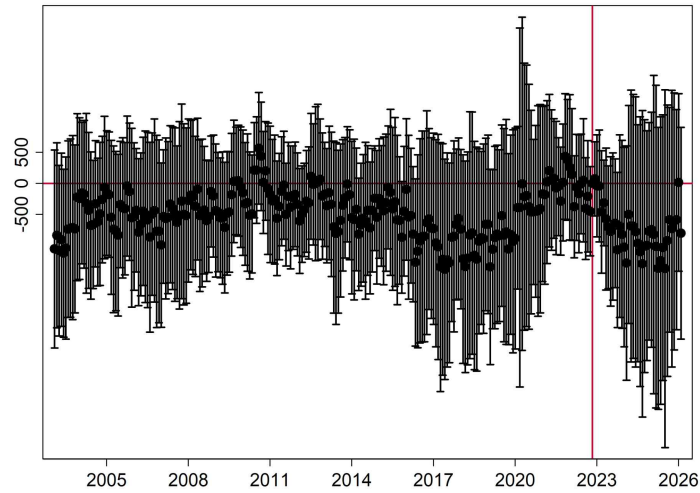
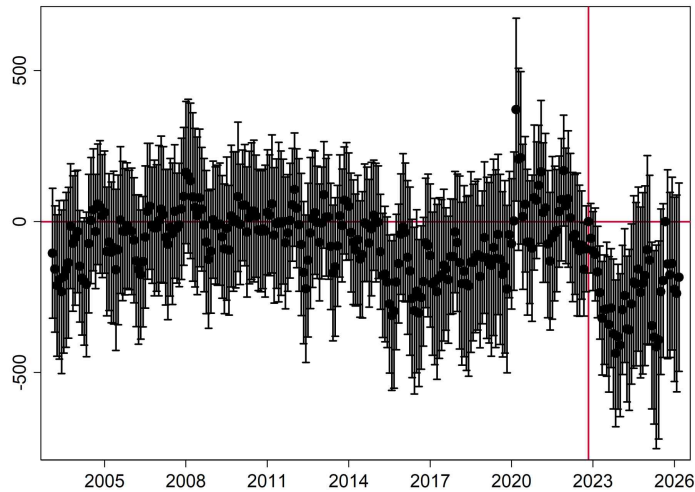


그림 8 고인지 직업군의 AI 노출지수 이벤트 스터디



다. 두 축의 종합: 측정 축의 갈라짐

반복성 강도 축과 AI 노출지수 축의 결과를 나란히 놓으면 한 가지 사실이 분명해진다. 고인지 직업군에서 두 축의 사후 계수는 부호가 반대다. 반복성 강도로 재면 양(+)이고, AI 노출지수로 재면 음(-)이다. 이상반된 부호는 모순이 아니라, 두 지수가 같은 직업군을 서로 다른 기준으로 줄 세우기 때문에 생긴다. 실제로 고인지 하위표본 안에서 반복성 강도와 AI 노출지수의 상관은 약 +0.12, 저인지 하위표본 안에서는 약 -0.08로, 두 경우 모두 0에 가깝다. 곧 인지 수준을 고정하고 보면 두 측정치는 사실상 독립이어서, 한 직업이 얼마나 반복적냐는 것과 얼마나 AI에 노출되어 있냐는 것은 별개의 정보다. 따라서 상반된 두 계수는 동전의 한 면을 뒤집은 반대면이 아니라, 서로 다른 축에서 따로 식별된 별개의 추정치다.

다만 측정치가 거의 독립이라는 사실이 두 계수의 배후 원인까지 서로 무관함을 뜻하지는 않는다. 실제로 2023년 전후 비반복 두뇌 직업의 상대적 위축이라는 사건은 AI 노출이라는 동전의 반대편일 수도 있다. 하지만 엄밀하게 이 위축의 원인은 생성형 AI의 대체효과일 수도, 같은 시기 비반복 두뇌 직업에 집중된 빅테크 해고·금리 인상이라는 비AI 충격일 수도 있어, 관측만으로 인과적 원인을 판정하기 어렵다.

한편 저인지 직업군에서 반복성 강도로 재면 음(-)이고, AI 노출지수로 재면 유의하지 않다. 반복성 강도의 음(-)의 결과는 전형적인 RBTC로 이해될 수 있고, AI 노출지수의 유의하지 않은 결과는 육체노동을 AI가 대신해주지 못한다는 직관과 일치한다.

제IV장. 강건성 검사 (Robustness Check)

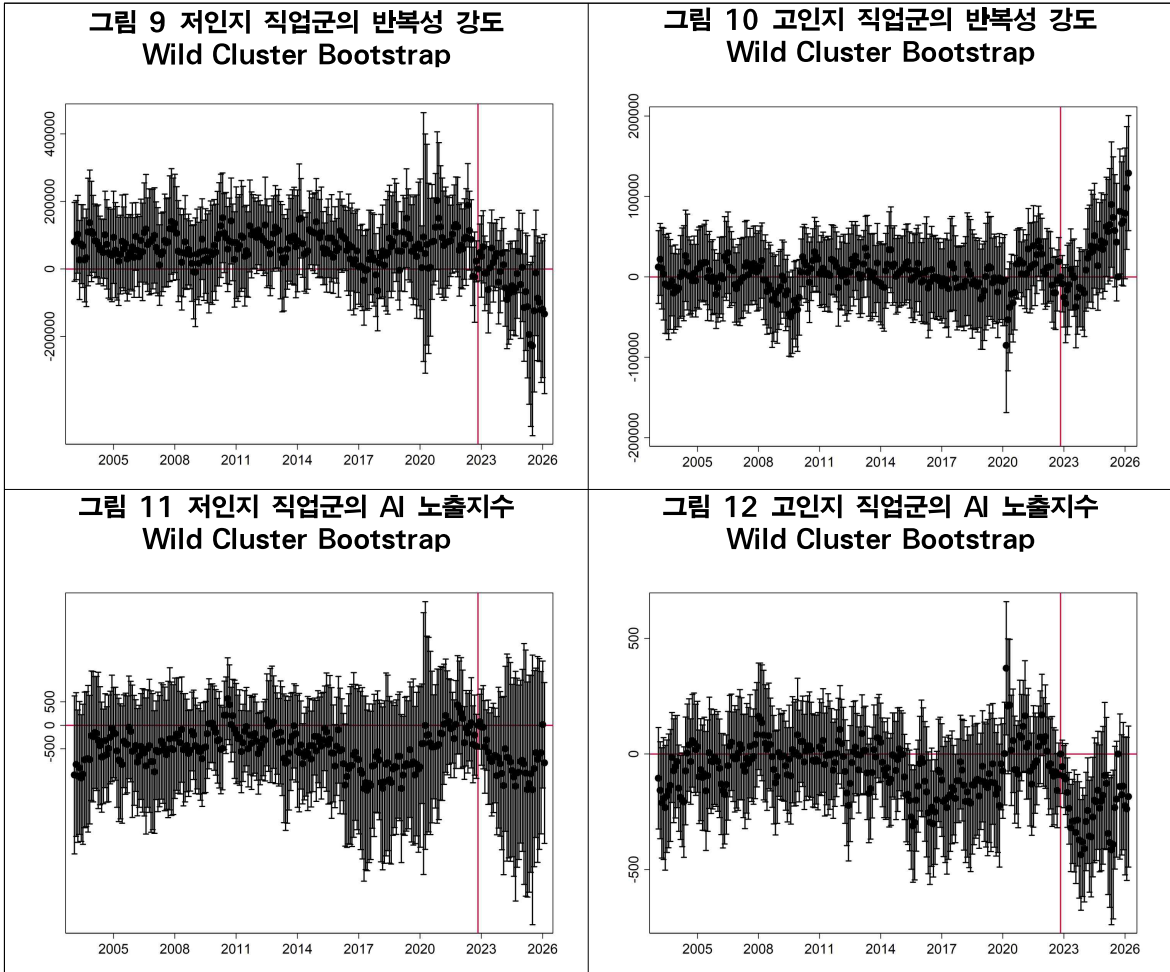
제1절. Wild Cluster Bootstrap

본 분석의 baseline (제III장)에서 사용한 모형은 직업군 군집 짝(pairs) 부트스트랩이다. 다만 이벤트 스터디에는 시점 고정효과와 반복성 강도×시점 항이 많아 추정 모수가 직업군 군집 수에 비해 크다. 전체 표본의 직업군 군집은 약 123개이며, 앞의 분석에서처럼 표본을 인지 점수로 양분하면 각 하위표본의 군집 수는 그보다 더 줄어들어 모수 대비 군집이 넉넉하지 않다는 우려가 오히려 커진다. 군집이 모수에 비해 부족하면 군집-강건 분산이나 짝 부트스트랩이 표준오차를 작게 잡을 수 있으므로, Roodman et al. (2019)은 더 나은 소표본 성질을 갖는 wild cluster bootstrap을 제안하였다. 본 연구는 이것을 반복성 강도 추과 AI 노출지수 추 모두에서 저인지·고인지 각 하위표본에 강건성 검사로 적용한다. 이 wild 검사는 1단계 추세제거 결과를 고정된 채 2단계 잔차만 직업군 가중으로 교란하므로, 2단계 군집 추론의 강건성만 본다. 1단계의 불확실성까지 담는 짝 부트스트랩이 여전히 baseline 추론이며, baseline 방법과 wild 방법은 서로 대체하지 않고 보완한다.

결과는 그림 9, 10에 제시되었으며 baseline과 사실상 같다. wild 방식은 동일한 2단계 점추정을 그대로 두고 군집 수준의 추론만 다시 수행하므로, 점추정은 짝 부트스트랩과 같고 달라지는 것은 신뢰구간과 그로부터의 유의성 판정뿐이다. 두 하위표본 모두 사전 구간(2003~2022)에서는 계수가 0 근방에 머물며 신뢰구간이 0을 포함한다. 처치 이후에는 baseline의 부호 반전이 그대로 유지된다. 저인지 집단에서는 계수가 점차 하락하여 2025~2026년에 약 -250,000에서 -350,000명 수준에 이르고, 후기 여러 시점에서 신뢰구간이 0을 배제한다. 반대로 고인지 집단에서는 계수가 상승하여 2025~2026년에 약 +100,000에서 +150,000명 수준에 이르고, 역시 후기에 신뢰구간이 0을 배제한다.

앞 장의 AI 노출지수 추 결과도 같은 wild cluster bootstrap으로 점검하며, 결과는 그림 11, 12에 제시했다. AI 노출지수의 사후 계수 역시 점추정은 짝 부트스트랩과 같고 달라지는 것은 신뢰구간뿐이다. 고인지 직업군에서는 처치 이후 음(-)의 계수가 wild 추론에서도 사후 말기 여러 시점에서 신뢰구간이 0을 배제

하여, 노출이 큰 비정형 두뇌 직업의 상대적 부진이라는 신호가 추론 방법을 바꾸어도 유지된다. 반면 저인지 직업군에서는 baseline과 마찬가지로 사후에도 신뢰구간이 0을 포함하여 유의한 변화가 관측되지 않는다.



요컨대 반복성 강도 측에서 인지 수준에 따라 사후 계수의 부호가 갈리는 핵심 결과도, AI 노출지수 측에서 고인지 직업군의 사후 계수가 음(-)으로 유의한 결과도, 추론 방법을 짝 부트스트랩에서 wild cluster bootstrap으로 바꾸어도 견고하게 유지된다.

제V장. 결론

본 연구는 생성형 AI가 본격적으로 보급되기 시작한 2023년 전후를 대상으로, 직업의 반복성(routine) 강도와 고용의 관계가 인지(cognitive) 수준에 따라 어떻게 달랐으며 그 분기가 어떤 요인들로 해석되는가를 물었다. 나아가 같은 설계를 AI 노출지수에도 대등하게 적용해, 반복성 측과 AI 노출 측이 노동시장을 어

땡게 다르게 비추는지를 함께 살폈다. 이를 위해 O*NET 과업 정보를 대규모 언어모델로 채점하는 방법으로 반복·인지 과업점수를 구축하는 작업부터 수행했다. 그 후 직업군별 고용 추세의 이질성을 1단계에서 제거한 뒤 2단계에서 연속 처치 이벤트 스터디를 추정하는 이중차분에 직업군 군집 부트스트랩 추론을 결합하였다. 분석 단위는 OCC 4자리 직업군이며, 표본은 2003년 3월부터 2026년 4월까지의 월별 패널이다.

첫 번째 핵심 발견은 반복성 강도에 대한 사후 계수의 부호가 인지 수준에 따라 갈린다는 것이다. 인지 점수 0.5 미만의 저인지 직업군에서는 처치 이후 반복성이 높은 직업일수록 추세 대비 상대적으로 고용이 더 부진하여 사후 계수가 음(-)으로 나타난 반면, 0.5 이상의 고인지 직업군에서는 반복성이 높은 직업일수록 상대적으로 고용상황이 더 나아져 사후 계수가 양(+)으로 나타났다. 저인지의 음(-)은 반복편향 기술 변화(RBTC)가 오래 예측해 온 바와 부합한다. 반대로 고인지의 양(+)은 명확한 인과적 원인을 단정하기 어렵다. 그러나 적어도 RBTC 이론에 따라서 ‘반복성이 높을수록 고용이 감소한다’는 예측은, 고인지 집단에서 2023년 이후 더 이상 성립하지 않는 것으로 보인다.

본 연구는 반복성 강도뿐 아니라 Eloundou et al. (2024)이 구축한 AI 노출지수를 같은 설계의 처치 강도로 삼아 추정했다. 그 결과 고인지 직업군에서 사후 계수가 유의한 음(-)으로 나타나, 두뇌 직업 안에서는 AI 노출이 클수록 고용이 상대적으로 부진해졌다. 이는 반복성을 처치 강도로 사용했을 때의 사후 계수가 양(+)이었던 것과 부호가 반대인데, 두 측정치가 고인지 그룹 하에서 사실상 무상관(상관관계 +0.12)임을 고려할 때, 이러한 반대 부호는 동전 앞면의 거울의 뒷면 결과가 아니라, 단지 서로 다른 관점을 바라보는 다른 지표의 결과다. 다만 이러한 해석이 무조건 AI의 영향 때문이 아니라는 단정 또한 아니다.

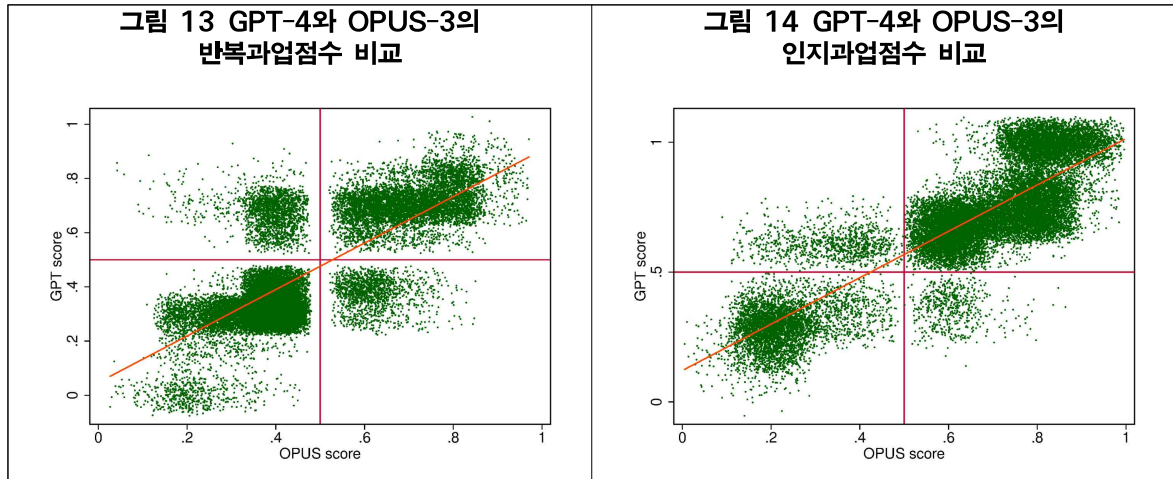
한편 저인지 직업군에서는 AI 노출지수를 처치 강도로 삼아 이중차분 했을 때, 사후 계수가 0에서 유의하게 다르지 않았다. 이는 육체를 주로 쓰는 직업에서는 AI가 고용을 대체하지도, 보완하지도 않았음을 의미한다. 이는 반복성을 처치 강도로 사용했을 때의 사후 계수가 음(-)이었던 것과 다른 결과이다. 이 또한 반복성 지수와 AI 노출지수가 단순히 거울처럼 반대인 하나의 지표가 아니라 서로 다른 메커니즘을 관찰할 수 있는 다른 정의의 지표임을 시사한다.

발견의 한계는 분명히 해 둘 필요가 있다. 본 연구는 생성형 AI가 고용을 특정 방향으로 바꾸었다는 깨끗한 인과적 증명을 하지 못했다. 또한 2023년 이후 사후 기간이 아직 길지 않아 후기 동학의 해석은 잠정적이다. 따라서 본 연구의 발견은 생성형 AI의 깨끗한 인과효과가 아니라 반복성 강도와 AI 노출지수라는 두 지표에서 관측된 이질적인 패턴으로 읽어야 한다.

이러한 유보에도 불구하고 본 연구의 기여는 분명하다. 첫째, 반복성과 고용의 관계가 생성형 AI 확산기에 단일한 방향으로 휩쓸린 것이 아니라 인지 수준에 따라 상반된 방향으로 갈렸음을 보였다. 둘째, 반복성 강도와 AI 노출지수를 같은 설계로 대등하게 추정해, 두 측정 축이 노동시장의 서로 다른 위험을 비추는 점을 드러냈다. 20여 년간 반복편향 기술변화(RBTC) 문헌이 의지해 온 반복성 지표가 생성형 AI의 고용 영향 축을 그대로 대신하지는 않으므로, 생성형 AI 확산기의 노동시장을 읽을 때 어떤 측정 축을 택하느냐가 결론을 가른다는 점을 본 연구가 알려준다.

부록

부록 A: 그림



부록 B: 채점 지시문

본문에서 밝혔듯 반복·인지 과업점수는 O*NET의 직업별 과업 기술(task description)을 대규모 언어모델(GPT-4와 OPUS-3)이 0에서 1 사이 연속값으로 평가하게 하여 구축했다. 과업의 반복성과 인지성에 대한 개념 정의는 본문에 이미 서술했으므로 여기서는 되풀이하지 않고, 모델에 실제로 입력한 채점 지시문을 영어 원문 그대로 실는다. 아래 지시문은 0과 1의 양극단, 0.3~0.7의 중간 구간, 불확실한 경우 0.5 부근으로 수렴시키는 규칙, 그리고 점수와 함께 판단 근거를 출력하게 하는 형식을 규정한다.

반복성(Routineness) 채점 지시문은 다음과 같다.

Extremely Non-Routine Task (Score: 0): Assign a value of 0 exclusively to tasks that necessitate human creativity, intuition, or involve complex, unpredictable problem solving that cannot be replicated by current automation technologies at all.

Extremely Routine Task (Score: 1): Assign a value of 1 only to tasks that are fully automatable with absolutely no need for human discretion or unpredictable judgment. This should be strictly limited to tasks where current technology can perform the task without any human oversight.

Moderately Routine or Non-Routine Tasks (Score range: 0.3 to 0.7): Assign values within this range to tasks that blend elements of both routine and non-routine characteristics, or when the classification into extreme categories is not clear. The middle range should be expanded slightly to encourage less extreme scoring, using 0.5 as a central point for truly ambiguous tasks.

Uncertainty Principle: If there is any uncertainty in classifying the task, default to a score closer to 0.5. Use the wider range of 0.3 to 0.7 to adjust the score slightly if there is a mild inclination towards routine or non-routine characteristics. This approach should ensure that only tasks with clear and definitive characteristics receive scores at the extremes.

Output Format: Begin your response with the score, followed by a colon and a detailed explanation of your reasoning. The explanation should thoroughly consider the task's characteristics, including the potential for automation and the level of human input or creativity required.

인지성(Cognitivity) 채점 지시문은 다음과 같다.

Extremely Manual Task (Score: 0): Assign a value of 0 only to tasks that are fully manual task.

Extremely Cognitive Task (Score: 1): Assign a value of 1 exclusively to tasks are full cognitive task.

Moderately Manual or Cognitive Tasks (Score range: 0.3 to 0.7): Assign values within this range to tasks that blend elements of both manual and cognitive characteristics, or when the classification into extreme categories is not clear. The middle range should be expanded slightly to encourage less extreme scoring, using 0.5 as a central point for truly ambiguous tasks.

Uncertainty Principle: If there is any uncertainty in classifying the task, default to a score closer to 0.5. Use the wider range of 0.3 to 0.7 to adjust the score slightly if there is a mild inclination towards manual or cognitive characteristics. This approach should ensure that only tasks with clear and definitive characteristics receive scores at the extremes.

Output Format: Begin your response with the score, followed by a colon and a detailed explanation of your reasoning. The explanation should thoroughly consider the task's characteristics.

전체 23,825개 과업에 대한 점수와 그 판단 근거는 별도로 공개하여, 제3자가 과업 코드로 임의의 O*NET 자료와 연결해 직접 검토·재분석할 수 있게 한다.

부록 C: 채점 예시

다음은 한 과업에 대해 두 모델이 산출한 점수와 근거(reasoning)의 예시다. 분량을 줄이기 위해 하나의 과업만 싣는다. 같은 과업에 대해 두 모델이 비슷하면서도 미묘하게 다른 점수를 부여하며, 각자 그 판단의 근거를 함께 출력함을 확인할 수 있다. 모델 출력은 영어 원문 그대로 옮긴다.

Task description: Resolve customer complaints regarding sales and service.

OPUS Routine score: 0.4

OPUS Routine reasoning: Resolving customer complaints regarding sales and service involves a blend

of routine and non-routine elements. While some aspects of the task, such as documenting complaints or following standardized resolution procedures, can be considered routine and potentially automatable, the core of the task requires human interaction, empathy, and problem-solving skills. Customer complaints often involve unique situations that demand active listening, understanding the customer's perspective, and finding creative solutions to address their concerns. Each complaint may present a different challenge that requires adapting to the specific context and using human judgment to determine the most appropriate course of action. However, the task is not entirely non-routine, as there are likely established guidelines.

GPT Routine score: 0.3

GPT Routine reasoning: Resolving customer complaints regarding sales and service is predominantly a non-routine task. While certain aspects of customer service can be automated, such as providing information about services or products, handling complaints often involves unpredictable elements and requires human intuition and creativity. Each customer complaint is unique and may require a different approach or solution, which is challenging to automate. The task involves understanding the customer's emotions, empathizing with them, and devising solutions to their problems, which are skills that are currently beyond the capabilities of automation technologies. However, since some aspects of the task, such as gathering initial information about the complaint, can be automated, it does not fully qualify as an extremely non-routine task.

OPUS Cognitive score: 0.8

OPUS Cognitive reasoning: Resolving customer complaints regarding sales and service is primarily a cognitive task. It involves several mental processes and skills, such as: 1. Communication and language comprehension: Understanding the customer's complaint and effectively communicating with them to gather information and provide solutions. 2. Problem-solving and decision-making: Analyzing the complaint, identifying the root cause of the issue, and determining the most appropriate course of action to resolve the problem. 3. Empathy and emotional intelligence: Being able to understand and relate to the customer's feelings and concerns, while maintaining a professional and composed demeanor. 4. Knowledge application: Utilizing knowledge of the company's products, services.

GPT Cognitive score: 0.8

GPT Cognitive reasoning: The task of resolving customer complaints regarding sales and service is primarily a cognitive task. It involves understanding the customer's problem (language comprehension), thinking of the best way to resolve the issue (problem-solving, decision making), and communicating that solution effectively (language comprehension again). It may also involve recalling relevant information about the company's policies or the customer's past transactions (memory). However, there might be some manual aspects involved such as typing responses or navigating through the customer service software, which is why the task doesn't score a full 1. The cognitive aspects of this

task clearly outweigh the manual ones, hence the score leans more towards the cognitive end of the scale.

부록 D: 상관 행렬의 해석

본문의 표 1은 본 연구가 측정한 LLM 기반 반복과업점수(이하 RTI-LLM)가 기존 측정치들과 얼마나 수렴하는지를 보여 준다. 비교 대상은 Walo(2023)가 한자리에 모아 정리한 다섯 가지 기존 RTI 측정치로, 직업사전(DOT)에 기반한 Autor 등의 측정을 비롯해 서로 다른 자료와 방법으로 구축된 Acemoglu, Goos 등, Dengler, Marcolin의 측정이다. 본 연구는 RTI-LLM을 이 상관 행렬에 같은 직업 단위로 추가해 비교했다.

해석의 핵심은 수렴타당도(convergent validity)다. 개념적으로 같은 대상, 곧 과업의 반복성을 겨냥한 측정치들이 서로 다른 방법으로 만들어졌음에도 일관되게 높은 양의 상관을 보인다면, 그 측정치들이 자의적 잡음이 아니라 공통의 실체를 포착하고 있다고 볼 근거가 된다. 표에서 RTI-LLM은 Acemoglu 측정과 0.76, Goos 등과 0.71, Marcolin과 0.66, Dengler와 0.64로 높은 상관을 보이고, 가장 오래된 DOT 기반의 Autor 등 측정과도 0.37의 양의 상관을 보인다(별표 셋은 1% 수준에서 유의함을 뜻한다). 과업 기반으로 정교하게 구축된 측정치들과의 상관이 특히 높다는 점은, RTI-LLM이 반복성의 핵심인 반복 그 자체를 비교적 잘 포착함을 시사한다.

표의 첫 행은 1980-2005년의 직업별 고용 비중과의 상관으로, RTI-LLM은 -0.46 의 음의 상관을 보인다. 이는 반복성이 높은 직업일수록 그 기간에 고용 비중이 줄어들었다는 반복편향 기술변화(RBTC) 문헌의 핵심 예측과 부합하며, 기존 측정치들이 보이는 음의 상관(-0.33 에서 -0.50)과도 같은 방향이다. 요컨대 상관 행렬은 RTI-LLM이 기존 측정치들과 수렴하는 동시에 반복편향 기술변화가 기록해 온 고용 패턴과도 정합함을 보여 주어, 새로운 측정 방식의 타당성을 뒷받침한다.

참고 문헌

- Acemoglu, D. and Restrepo, P., 2020. Robots and jobs: Evidence from US labor markets. *Journal of Political Economy*, 128(6), pp.2188 - 2244. <https://doi.org/10.1086/705716>
- Angrist, J.D. and Pischke, J.S., 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press. <https://doi.org/10.1515/9781400829828>
- Autor, D.H. and Dorn, D., 2013. The growth of low-skill service jobs and the polarization of the US labor market. *American Economic Review*, 103(5), pp.1553 - 1597. <https://doi.org/10.1257/aer.103.5.1553>
- Autor, D.H., Levy, F. and Murnane, R.J., 2003. The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, 118(4), pp.1279 - 1333. <https://doi.org/10.1162/003355303322552801>
- Bick, A., Blandin, A. and Deming, D.J., 2024. The rapid adoption of generative AI. NBER Working Paper No. 32966. Cambridge, MA: National Bureau of Economic Research.
- Borusyak, K., Jaravel, X. and Spiess, J., 2024. Revisiting event-study designs: Robust and efficient estimation. *The Review of Economic Studies*, 91(6), pp.3253 - 3285. <https://doi.org/10.1093/restud/rdae007>
- Callaway, B., Goodman-Bacon, A. and Sant'Anna, P.H.C., 2024. Difference-in-differences with a continuous treatment. NBER Working Paper No. 32117. Cambridge, MA: National Bureau of Economic Research.
- Cameron, A.C. and Miller, D.L., 2015. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2), pp.317 - 372. <https://doi.org/10.3368/jhr.50.2.317>
- Dauth, W., Findeisen, S., Südekum, J. and Woessner, N., 2017. German robots - the impact of industrial robots on workers. CEPR Discussion Paper No. 12306. London: Centre for Economic Policy Research.
- del Rio-Chanona, R.M., Ernst, E., Merola, R., Samaan, D. and Teutloff, O., 2025. AI and jobs. A review of theory, estimates, and evidence. arXiv preprint [arXiv:2509.15265](https://arxiv.org/abs/2509.15265).
- Eloundou, T., Manning, S., Mishkin, P. and Rock, D., 2024. GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702), pp.1306 - 1308. <https://doi.org/10.1126/science.adj0998>
- Felten, E.W., 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), pp.2195 - 2217. <https://doi.org/10.1002/smj.3286>
- Flood, S., King, M., Rodgers, R., Ruggles, S., Warren, J.R., Backman, D., Breton, E., Cooper, G., Rivera Drew, J.A., Richards, S., Van Riper, D. and Williams, K.C.W., 2025. IPUMS CPS: Version 13.0 [dataset]. Minneapolis, MN: IPUMS. <https://doi.org/10.18128/D030.V13.0>
- Freyaldenhoven, S., Hansen, C., Pérez Pérez, J. and Shapiro, J.M., 2021. Visualization, identification, and

- estimation in the linear panel event-study design. NBER Working Paper No. 29170. Cambridge, MA: National Bureau of Economic Research.
- Friedberg, L., 1998. Did unilateral divorce raise divorce rates? Evidence from panel data. *American Economic Review*, 88(3), pp.608 - 627.
- Gardner, J., 2022. Two-stage differences in differences. arXiv preprint arXiv:2207.05943.
- Haslberger, M., 2022. Rethinking the measurement of occupational task content. *The Economic and Labour Relations Review*, 33(1), pp.178 - 199. <https://doi.org/10.1177/10353046211037095>
- Humlum, A. and Vestergaard, E., 2024. The adoption of ChatGPT. Becker Friedman Institute Working Paper No. 2024-50. Chicago, IL: University of Chicago.
- Jacobson, L.S., LaLonde, R.J. and Sullivan, D.G., 1993. Earnings losses of displaced workers. *American Economic Review*, 83(4), pp.685 - 709.
- Lee, J.Y. and Solon, G., 2011. The fragility of estimated effects of unilateral divorce laws on divorce rates. *The B.E. Journal of Economic Analysis & Policy*, 11(1), pp.1 - 11. <https://doi.org/10.2202/1935-1682.2994>
- Meer, J. and West, J., 2016. Effects of the minimum wage on employment dynamics. *Journal of Human Resources*, 51(2), pp.500 - 522. <https://doi.org/10.3368/jhr.51.2.0414-6298R1>
- National Center for O*NET Development, 2023. O*NET Resource Center. U.S. Department of Labor. Available at: <https://www.onetcenter.org/>
- Pagan, A., 1984. Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, 25(1), pp.221 - 247. <https://doi.org/10.2307/2648877>
- Pustejovsky, J.E. and Tipton, E., 2018. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36(4), pp.672 - 683. <https://doi.org/10.1080/07350015.2016.1247004>
- Roodman, D., Nielsen, M.Ø., MacKinnon, J.G. and Webb, M.D., 2019. Fast and wild: Bootstrap inference in Stata using boottest. *The Stata Journal*, 19(1), pp.4 - 60. <https://doi.org/10.1177/1536867X19830877>
- Roth, J., 2022. Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3), pp.305 - 322. <https://doi.org/10.1257/aeri.20210236>
- Schmidheiny, K. and Siegloch, S., 2023. On event studies and distributed-lags in two-way fixed effects models: Identification, equivalence, and generalization. *Journal of Applied Econometrics*, 38(5), pp.695 - 713. <https://doi.org/10.1002/jae.2971>
- Walo, S., 2023. The link between routine tasks and job polarization: A task measurement problem? *LABOUR*, 37(3), pp.437 - 467. <https://doi.org/10.1111/labr.12251>
- Webb, M., 2019. The impact of artificial intelligence on the labor market. SSRN Working Paper No. 3482150.
- Wolfers, J., 2006. Did unilateral divorce laws raise divorce rates? A reconciliation and new results. *American Economic Review*, 96(5), pp.1802 - 1820. <https://doi.org/10.1257/aer.96.5.1802>