README

Disruption without Dividend? How the Digital Divide and Task Differences split GenAI's Global Impact

By Pawel Gmyrek, Mariana Viollaz and Hernan Winkler (Dec. 2025)

Overview

This document contains the necessary information to replicate the results in "Disruption without Dividend? How the Digital Divide and Task Differences split GenAI's Global Impact".

The main folders included in the package, which must be located in the same folder for a correct replication of the results, are as follows:

- Codes
  - 01_Country_level_GenAI_exposure
  - 02_GenAI_exposure_by_internet_access
  - 03_Task_content_job_measures
  - 04_GenAI_exposure_task_content
  - 05_Mlogit_model_annex
  - ado
- Data
  - Raw
  - Cleaned
- Outputs
  - Main
  - Annex

Data Availability

This paper uses multiple data sources based on household surveys from various economic contexts. The data come from EAPPOV (for East Asian economies) from the World Bank; GLD (for developing economies worldwide) from the World Bank; SEDLAC and LABLAC (for LAC economies) from The Center of Distributive, Labor and Social Studies (CEDLAS); PIAAC surveys (for OECD economies); WAEMU (for West African economies); and STEP (for economies included in the STEP program) from the World Bank. **No data can be made publicly available**. See the section "Folders and Contents" for further details on the list of datasets included in each raw data subfolder, the data source and direct links to the data sources when available. Some data are downloaded dynamically; in those cases, the date on which the data were accessed is reported. Lastly, the file name specifies the year to which the dataset corresponds for household surveys.

Statement about Rights

☐ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

☐ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package.

Folders and Contents

Data

- **Raw**

  The "Raw" folder contains the raw datasets used for this paper. It is organized by the following subfolders:

  - "EAPPOV": includes all microdata used in this paper for countries in the East Asia and the Pacific (EAP) from the World Bank. From this data source, we used only the countries with ISCO codes version 2008 (ISCO-08) and its latest year available; the corresponding year for each country is specified in the file name (e.g. KHM_2021 for Cambodia in 2021). Access date: May 1st, 2025. All the datasets in Stata format (.dta) included in the subfolder are:

    - KHM_2021_CSES_v01_M_v01_A_EAPPOV_I.dta
    - LAO_2018_LECS_v01_M_v06_A_EAPPOV_I.dta

    The subfolder "EAPPOV" contains all restricted-use EAPPOV databases included in this paper. The original data source is the World Bank Stata API *datalibweb*, for internal use only. The EAPPOV databases listed above can be used for reproducibility verification at the WB but cannot be shared publicly on the portal. Please note that not all these EAPPOV databases are publicly available, and hence cannot be shared to external researchers without specific prior written permission.

- "GLD": includes all microdata used in this paper from the Jobs Group's Global Labor Database (GLD) at the World Bank, which covers labor market data from household surveys worldwide, with a focus in developing economies. From this data source, we used only the countries with ISCO codes version 2008 (ISCO-08) and its latest year available; the corresponding year for each country is specified in the file name (e.g. ALB_2013 for Albania in 2013). Access date: February 21st, 2025. All the datasets in Stata format (.dta) included in the subfolder are:
    - ALB_2013_LFS_V01_M_V01_A_GLD_ALL.dta
    - BGD_2022_QLFS_V01_M_V02_A_GLD_ALL.dta
    - BOL_2021_ECE_V01_M_V02_A_GLD_ALL.dta
    - ETH_2021_LFS_V01_M_V03_A_GLD_ALL.dta
    - GEO_2022_LFS_V01_M_V01_A_GLD_ALL.dta
    - IND_2022_PLFS_V01_M_V02_A_GLD_ALL.dta
    - LKA_2021_LFS_V01_M_V03_A_GLD_ALL.dta
    - MEX_2023_ENOE_V01_M_V01_A_GLD_ALL.dta
    - MNG_2022_LFS_V01_M_V02_A_GLD_ALL.dta
    - PAK_2020_LFS_V01_M_V04_A_GLD_ALL.dta
    - RWA_2021_LFS_V01_M_V01_A_GLD_ALL.dta
    - THA_2021_harmonized_LFS.dta
    - TUR_2019_HLFS_V01_M_V03_A_GLD_ALL.dta
    - ZMB_2022_LFS_V01_M_V01_A_GLD_ALL.dta
    - ZWE_2022_QLFS_V01_M_V02_A_GLD_ALL.dta

  GLD data is stored on a server managed by the GLD team. The GLD databases listed above can be used for reproducibility verification at the WB but cannot be shared publicly on the portal. Datasets are available for World Bank staff, and they can be downloaded from the GLD server or the Stata API *datalibweb*. For researchers outside the World Bank, it is possible to reproduce the datasets using the published do-files in the GLD Github:

  https://worldbank.github.io/gld/README.html

- "SEDLAC": includes all microdata from the Socio-Economic Database for Latin America and the Caribbean (SEDLAC) used in this paper. From this data source, we used only the countries with ISCO codes version 2008 (ISCO-08) and its latest year available; the corresponding year for each country is

specified in the file name (e.g. ARG_2023 for Argentina in 2023). Access date: March 28th, 2025. All the datasets in R format (.rds) included in the subfolder are:

- ARG_2023_EPHC-S2_v01_M_v01_A_SEDLAC-03_all.rds
- BOL_2023_EH_v01_M_v01_A_SEDLAC-03_all.rds
- BRA_2023_PNADC-E1_v01_M_v01_A_SEDLAC-03_all.rds
- CHL_2022_CASEN_v01_M_v01_A_SEDLAC-03_all.rds
- DOM_2023_ECNFT-Q03_v01_M_v01_A_SEDLAC-03_all.rds
- ECU_2023_ENEMDU_v01_M_v01_A_SEDLAC-03_all.rds
- GTM_2023_ENCOVI_v01_M_v01_A_SEDLAC-03_all.rds
- HND_2023_EPHPM_v01_M_v01_A_SEDLAC-03_all.rds
- PER_2023_ENAHO_v01_M_v01_A_SEDLAC-03_all.rds
- SLV_2023_EHPM_v01_M_v01_A_SEDLAC-03_all.rds
- URY_2023_ECH_v01_M_v01_A_SEDLAC-03_all.rds

The subfolder "SEDLAC" contains all restricted-use SEDLAC databases included in this paper and sourced from SEDLAC harmonization project (World Bank and CEDLAS). The original data source is the World Bank Stata API *datalibweb*, for internal use only. The SEDLAC databases listed above can be used for reproducibility verification at the WB but cannot be shared publicly on the portal. Please note that not all these SEDLAC databases are publicly available, and hence cannot be shared to external researchers without specific prior written permission.

- "LABLAC": includes all microdata from the Labor Database for Latin America and the Caribbean (LABLAC) used in this paper. From this data source, we used only the countries with ISCO codes version 2008 (ISCO-08) and its latest year available; the corresponding year for each country is specified in the file name (e.g. bra_2024 for Brazil in 2024). Access date: March 6th, 2025. All the datasets in Stata format (.dta) included in the subfolder are:

  - LABLAC_bra_2024_q03_ALL.dta
  - LABLAC_col_2024_q03_ALL.dta
  - LABLAC_cri_2024_q03_ALL.dta
  - LABLAC_dom_2024_q03_ALL.dta
  - LABLAC_ecu_2024_q03_ALL.dta

- LABLAC_mex_2024_q02_ALL.dta
- LABLAC_per_2024_q03_ALL.dta
- LABLAC_slv_2023_q04_ALL.dta
- LABLAC_ury_2024_q02_ALL.dta

The subfolder "LABLAC" contains all restricted-use LABLAC databases included in this paper. LABLAC is a complementary source to the SEDLAC database, with a focus in labor statistics. The original data source is the World Bank Stata API *datalibweb*, for internal use only. The LABLAC databases listed above can be used for reproducibility verification at the WB but cannot be shared publicly on the portal. Please note that not all these LABLAC databases are publicly available, and hence cannot be shared to external researchers without specific prior written permission.

- "PIAAC": includes all microdata from the Survey of Adults Skills (PIAAC), from both 1st and 2nd cycles, used in this paper. All files are stored in R (.rds) and STATA (.dta) formats. The databases can be used for reproducibility verification at the WB but cannot be shared publicly on the portal. PIAAC survey databases were accessed in April 2025 and are publicly available at: https://www.oecd.org/en/data/datasets/piaac-2nd-cycle-database.html#data

- "WAEMU": includes all microdata used in this paper for countries in the West African Economic and Monetary Union (WAEMU). We used additional subfolders to store each country database separately. All data corresponds to 2021. The databases can be used for reproducibility verification at the WB but cannot be shared publicly on the portal. Access to the data can be requested by contacting the Africa Poverty Team (hwu4@worldbank.org, efoster1@worldbank.org). Access date: August 13th, 2025. The subfolders and the corresponding databases included are:

  - BEN/BEN_2021_GMD.dta
  - BFA/BFA_2021_GMD.dta
  - GNB/GNB_2021_GMD.dta
  - MLI/MLI_2021_GMD.dta
  - SEN/SEN_2021_GMD.dta
  - TGO/TGO_2021_GMD.dta

- "STEP": includes all microdata for some of the countries included in the STEP (Skills Toward Employment and Productivity) database. The STEP dataset includes surveys from a range of low- and middle-income countries across several regions. All files are stored in Stata format (.dta). Access date: March

1st, 2025. The databases can be used for reproducibility verification at the WB but cannot be shared publicly on the portal. Most of the microdata files can be found here:

https://microdata.worldbank.org/index.php/catalog/?page=1&collection%5B%5D=step&ps=15. The databases included are:

- STEP_PHL.dta
- STEP Kosovo_working_S11.dta
- STEP_SRB.dta
- el_salvador.dta
- STEP Armenia_working.dta
- STEP Bolivia_working.dta
- STEP Colombia_working.dta
- STEP Georgia_working.dta
- STEP Ghana_working.dta
- STEP Kenya_working.dta
- STEP Laos_working.dta
- STEP Macedonia_working.dta
- STEP Sri_Lanka_working.dta
- STEP Ukraine_working.dta
- STEP Vietnam_working.dta
- STEP Yunnan_working.dta

- "Other": includes all data from other data sources different to microdata based on household surveys. **The data in this folder can be shared publicly on the portal.**

  - Excel files "Country_emp_by_occupation_ILO.xlsx" contains the latest value of total employment by occupation (1-digit ISCO-08) and "Country_emp_by_ISCO2D_ILO.xlsx" contains the latest value of total employment by 2-digit ISCO-08 codes across countries. Both datasets are from the International Labour Organization (ILO), and can be dynamically downloaded at the ILOSTAT online portal: https://ilostat.ilo.org/topics/employment/ (the datasets used in the paper were downloaded on March 10th, 2025). On the other hand,

"mapping.csv" contains descriptions and labels for ISCO-08 occupation codes at all levels; this information is publicly available at:

https://ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation/

- "Final_Scores_ISCO08_Gmyrek_et_al_2025.xlsx": contains the 2025 GenAI scores at a 4-digit ISCO-08 level. This dataset was taken from Gmyrek et. al. (2025), and it can be downloaded at:

  https://github.com/pgmyrek/2025_GenAI_scores_ISCO08

- "Exposure_Gradients.xlsx" is a processed version of the original GenAI scores from "Final_Scores_ISCO08_Gmyrek_et_al_2025.xlsx" which keeps variables of interest for the paper and "AI_scores.dta" is just a version of this processed versions of the raw scores, but in Stata format (.dta).

- About datasets from World Bank World Development Indicators (WDI): These data were downloaded using the WB WDI database API. All data were downloaded on January 12th, 2026, except for "GDP_internet_from_WDI.rds," which was downloaded on January 9th, 2026. The WDI datasets include GDP per capita, total population, population with access to the internet, access to electricity, urban population, employed population aged 15+, and country region and income classification. See the "Specific instructions" section below for further details on the indicators used in each figure and their indicator IDs. We use 2024 as the reference year in all cases; however, if a country did not have data available for 2024, the latest available year was used. Table 1 below contains the list of countries used.

- **Cleaned**

  This subfolder of the "Data" folder contains all processed datasets used as an input for the results in the paper. All the datasets in this folder are the result of executing the codes (see the "Codes" section below for an explanation and details on where and how the "Cleaned" files were constructed). We used additional subfolders to store each dataset separately by exercise conducted in the paper.

Table 1: List of countries used for WDI data download

Afghanistan	Cyprus	Iran	Nauru	South Africa
Albania	Czechia	Iraq	Nepal	Spain
Angola	Denmark	Ireland	Netherlands	Sri Lanka
Argentina	Dominican Republic	Israel	Niger	Sudan
Australia	Ecuador	Italy	Nigeria	Suriname
Austria	Egypt	Jordan	North Macedonia	Sweden
Bahamas	El Salvador	Kenya	Norway	Switzerland
Bangladesh	Estonia	Kiribati	Pakistan	São Tomé & Príncipe
Barbados	Eswatini	Kosovo	Palau	Tajikistan
Belarus	Ethiopia	Kyrgyzstan	Palestinian Territories	Tanzania
Belgium	Fiji	Laos	Panama	Thailand
Belize	Finland	Latvia	Papua New Guinea	Timor-Leste
Benin	France	Lebanon	Peru	Togo
Bhutan	Gambia	Lesotho	Philippines	Tonga
Bolivia	Georgia	Liberia	Poland	Tunisia
Bosnia & Herzegovina	Germany	Lithuania	Portugal	Tuvalu
Botswana	Ghana	Luxembourg	Romania	Uganda
Brazil	Greece	Madagascar	Russia	United Arab Emirates
Brunei	Grenada	Maldives	Rwanda	United Kingdom
Bulgaria	Guatemala	Mali	Samoa	United States
Burkina Faso	Guinea	Marshall Islands	Senegal	Uruguay
Burundi	Guinea-Bissau	Mauritius	Serbia	Vanuatu
Cambodia	Guyana	Mexico	Seychelles	Viet Nam
Chile	Honduras	Micronesia (Federated States of)	Sierra Leone	Zambia
Colombia	Hungary	Mongolia	Singapore	Zimbabwe
Congo - Kinshasa	Iceland	Montenegro	Slovakia	
Costa Rica	India	Mozambique	Slovenia	
Croatia	Indonesia	Myanmar (Burma)	Somalia	

Codes

The "Codes" folder contains all the R/Stata codes needed for this replication. This folder contains separate folders with the codes for each exercise numerated with the order they should be executed. Inside each subfolder is a main script that allows to run all codes for a specific section after changing the top-level directory (e.g. main_for_01.R). The contents and the structure of the "Codes" folder is as follows:

- **01_Country_level_GenAI_exposure**
    - **00-preparing-GenAI-exposure.R:** this code takes raw Excel file "Data/Raw/Other/Final_Scores_ISCO08_Gmyrek_et_al_2025.xlsx", renames variables, and selects variables of interest for the paper. It then saved resulting database in "Data/Raw/ Other/Exposure_Gradients.xlsx".
    - **01-estimating_ISCO_averages.R:** this code merge GenAI exposure scores with raw microdata from GLD, LABLAC, PIAAC, and WAEMU for countries with 4-digit ISCO-08 codes available. It then estimates national average, and average exposure at 3, 2, and 1-digit ISCO-08 by country and saved results in "Data/Cleaned/ mean_exposure_isco_all"
    - **02-knn-closest_match(ILO-exercise).R:** this code takes file:
        - "Data/Raw/Other/Country_emp_by_occupation_ILO.xlsx"

      As an input, from ILOSTAT, which contains the latest value of total employment by occupation for all countries with data available, to estimate the share of employment at 1-digit ISCO-08. It then calls country-level data for total population, internet users and GDP per capita from the WDI API. The data are used to prepare and estimate a K-Nearest Neighbors algorithm (KNN). The result is saved in:
        - "Data/Cleaned/ knn_algorithm/Match_ILO_country_exercise.rds"
    - **03-graphs.R:** this code uses outputs from codes previous codes (01-02) to imputed mean exposure scores at a 2-digit level from the closest match for countries without microdata or 4-digit ISCO-08 codes available. The resulting dataset is then used to created Figures 1-2 and Tables A1 and A2. The outcomes of this code are saved in "Outputs" and "Main" or "Annex" depending on the figure.
- **02_GenAI_exposure_by_internet_access**
    - **01-processing-data-with-internet.R:** this code prepares raw microdata from SEDLAC, WAEMU and PIAAC for countries with internet access information. It harmonized variables of interest and combined all data sources into the dataset:
        - "Data/Cleaned/internet_use_prediction/ SEDLAC_WAEMU_PIAAC.rds".

- **02-processing-data-without-internet.R:** this code prepares raw microdata from GLD, EAPPOV and countries in SEDLAC that does not have internet access information. It harmonized other variables of interest and combined all data sources into the dataset:
    - "Data/Cleaned/internet_use_prediction/GLD_EAPCE_others.rds"
- **03-models-estimation.R:** this code uses outputs from previous codes (01-02). It estimates a logit-lasso model on countries with internet access data to predict internet access in countries without this information and export the results in files:
    - **"**Data/Cleaned/internet_use_prediction/ Pooled_logit_lasso_predictions.rds**"**
    - **"**Data/Cleaned/internet_use_prediction/ GLD_logit_lasso_predictions.rds**"**
- **04-estimating-avg-exposure-gradients.R:** this code uses outputs from codes 01 and 02 to estimate average exposure at 3, 2, and 1-digit ISCO-08 across countries with 4-digit ISCO-08 codes available. Results are saved in "Data/Cleaned/ internet_use_prediction/" with files are named as: "ISCO_(…)_gradient_avg.rds"
- **05-knn-closest_match(internet-exercise).R:** this code follows the structure of the KNN algorithm in the code "02-knn-closest_match(ILO-exercise).R" from "ILO_country_exercise" (see explanation above), but it estimates the algorithm just for the set of countries with microdata available. It divides the sample of countries with microdata for all data sources into two sets: the set of countries with 4-digit ISCO-08 codes (used to impute), and the set of countries without 4-digit ISCO-08 codes (that needs imputation at the most disaggregates ISCO-08 codes available in each case). The output of this code is:
    - "Data/Cleaned/ knn_algorithm/Match_internet_exercise.rds"
- **06-graphs.R:** this code uses outputs from all previous codes to create Figures 3-8 and 10 and Tables A3 andA4. It combines microdata after logit-lasso estimation to predict internet access at individual level (codes 01-03), merge GenAI exposure data and imputes exposure averages at the most disaggregated ISCO-08 codes available using KNN results for countries without 4-digit ISCO-08 information available. It prepares and reshapes data for creating each figure (see specific instructions for further details on how figures were created). The outcomes of this code are saved in "Outputs" and "Main" or "Annex" depending on the figure. Also, when processing welfare data, this code saved a dataset containing the average welfare and the share of employment exposed within six exposure categories by 2-digit ISCO-08 across countries, named

"Data/Cleaned/Welfare/ Welfare_at_2d.xlsx". This table serves as an input for later codes and figures.

- **03_Task_content_job_measures**
    - Codes: 00_preparadb_piaac_v1.do, 00_preparadb_step_v1.do, 01_piaac_v1.do, 01_step_v1.do takes raw microdata from "PIAAC" and "STEP" folder to cleaned and prepared the data for the GenAI exposure with task measures section of the paper; it merges data, keep variables of interest and create task measures for each country. The resulting databases by executing these codes are stored in "Data/Cleaned/ Tasks_outputs".
    - **02_preparadb_task_v1.do:** this code takes cleaned PIAAC data and merge GenAI exposure scores in Excel file "Exposure_Gradients.xlsx" and prepares task content measures. The result of executing this code is database "Cleaned/PIAAC/ PIAAC_4digit_with_4digit_AI_tasks_scores.dta" This dataset contains exposure scores and task measures at a 4-digit ISCO-08 level for PIAAC countries.
    - **02_task_measures_2d_v1.do** and **03_task_ai_2d_v1.do:** like the previous code, this takes cleaned PIAAC and STEP data and merge GenAI exposure scores, but at a 2-digit ISCO-08 level. The results of executing this code are databases "Data/Cleaned/PIAAC/tasks_exposure_2d_for_pawel.dta" and "Data/Cleaned/PIAAC/2d_step_piaac_for_pawel.dta". These datasets contain exposure scores and task measures at a 2-digit ISCO-08 level for PIAAC-STEP countries.
- **04_GenAI_exposure_task_content**

  Requirement: to successfully execute "Step_2_GenAI_and_task_content.Rmd", R must be launched by opening the project file:

    - "RP_Disruption_without_Dividend.Rproj"

    - **Step_2_GenAI_and_task_content.Rmd:** this code uses processed datasets to create Figures 9, 11–14 and Figure A15. It loads, prepares and reshapes data (see specific instructions for further details on how these figures were created), merge standardized task-intensity datasets (STEP/PIAAC), estimated GAM models and calculate GenAI exposure scores used in the figures. The outcomes of this code are saved in "Outputs" and "Main" or "Annex" depending on the figure.
- **05_Mlogit_model_annex**
    - **01-processing-micro-data.R:** this code follows a similar structure for the microdata preparation on codes "01-processing-data-with-internet.R" and "02-processing-data-without-internet.R" (see the code sections above for further details). It processes all data sources, keeps just the variables of

interest for a multinomial model estimation and a slightly different sample where all individuals are taken whether they have internet access information or not. The output of this code is saved in:

- "Data/Cleaned/mlogit_model_exercise/microdata_for_model.rds"

- **02-data-preparation-mlogit.R:** this code takes the output from the previous code and make final preparations for the model estimation. It filters countries with 4-digit ISCO codes available, merge GenAI exposure data and creates the dependent variable for the model. The output of this code is a .dta file and is saved in:

  - **"**Data/Cleaned/mlogit_model_exercise/mlogit_data_STATA.dta**"**

- **03-mlogit-model-estimation.do:** this do-file takes the output of the previous code as an input for the estimation of a multinomial logit to explore the relationship between three major exposure outcomes (not/low exposed, augmentation potential and automation potential) and socio-economic characteristics. It also estimates the average marginal effects on the probability of exposure outcomes. The model results are used to create Table A5 and Figure A16 (see specific instructions for further details on how these outcomes were created).

Outputs

The "Outputs" folder contains all Figures of the paper. It separates outputs into two folders: "Main" for the outputs in the main text, and "Annex" for the outputs in the appendix of the paper. For more specific details of each of the figures included in this paper, see the following section "Specific Instructions", considering also previous explanations in folder "Codes".

Specific Instructions:

Figure 1: The dataset for this figure is constructed using the estimation of average exposure measures at the most disaggregated ISCO codes for each country from household surveys, available in the folder "mean_exposure_isco_all", and the matching between ILOSTAT labor shares at a 2-digit ISCO level and the KNN results for countries without microdata from household surveys, available in:

- "Data/Cleaned/knn_algorithm/Match_ILO_country_exercise.rds"
- "Data/Raw/Other/Country_emp_by_ISCO2D_ILO.xlsx"

These results are obtained by executing the codes "01-estimating_ISCO_averages.R" and "02-knn-closest_match(ILO-exercise).R" from folder "ILO_country_exercise". Also, the assigned part of the R script contacts the WB WDI database, downloads the GDP per capita (USD 2021 PPP) (NY.GDP.PCAP.PP.KD) and the income group for each country. The data are extracted from WDI API in a dynamic way; it takes the latest non-missing value available on January 12th, 2026.

Thus, the figure is the result of plotting the national-average share of exposed employment (y-axis) and the log of GDP per capita (x-axis), colored by the country income group.

Figure 2: The dataset for this figure is the same as for Figure 1 (see explanation of the data above). The figure plots the relationship between the national average of employment exposed across all six exposure measures (not exposed, low exposure, and gradients one through four), on the y-axis, and the log of GDP per capita (x-axis).

Figure 3: The dataset for this figure is constructed by estimating the average exposure within the six exposure categories on all microdata available, adjusted by the probability of access to the internet. The microdata with the probability of access to the internet are obtained by executing the code "03-models-estimation.R". The exposure data are obtained by executing the code "04-estimating-avg-exposure-gradients.R" and are imputed to countries without ISCO at the 4-digit level using the results from "Match_internet_exercise.rds", obtained by executing the code "05-knn-closest_match(internet-exercise).R", and the figure is produced by executing the code "06-graphs.R". Thus, this figure plots the share of employment, with and without internet access, exposed and low or non-exposed, disaggregated by country. All codes mentioned above are part of the folder "Codes/Internet_access_exercise".

Figure 4: The dataset for this figure is the same as for figure 3 (see explanation of the data above), but instead of estimating the average of exposure at a national level, it contains the share of employment disaggregated by socio-economic characteristics across three major exposure categories by income group: employment exposed under Gradients 3 & 4, Gradients 1& 2, and those low or non-exposed, each disaggregated by internet access. In this figure results are disaggregated by 1-digit ISCO-08 occupations. This figure is obtained by executing the code "06-graphs.R" from "Internet_access_exercise".

Figures 5: The dataset for this figure is the same as for Figure 4. Results are disaggregated by gender across three major exposure categories (see explanation on Figure 4 for further details on the groups). This figure is obtained by executing the code "06-graphs.R" from "Internet_access_exercise".

Figure 6: The dataset for this figure is the same as for Figure 4. Results are disaggregated by age group (16-25**,** 26-35, 36-45, 46-55, 56-65), following PIAAC's age group classification, across three major exposure categories (see explanation on Figure 4 for further details on the groups). This figure is obtained by executing the code "06-graphs.R" from "Internet_access_exercise".

Figure 7: The dataset for this figure is the same as for Figure 4. Results are disaggregated by highest level of educational attainment (no education, primary, secondary, tertiary) across three major exposure categories (see explanation on Figure 4 for further details on the groups). This figure is obtained by executing the code "06-graphs.R" from "Internet_access_exercise".

Figure 8: The dataset for this figure is the same as for Figure 4. Results are disaggregated by sector, based on ISIC Rev. 4 sections as code in the GLD data, across three major exposure

categories (see explanation on Figure 4 for further details on the groups). This figure is obtained by executing the code "06-graphs.R" from "Internet_access_exercise".

Figure 9: The figure is based on the file "Data/Cleaned/Welfare/Welfare_at_2d.xlsx", which contains processed data on ISCO-08 2-digit share of employment exposed within six exposure categories and welfare, for sixteen countries from SEDLAC and WAEMU, and ISCO-08 2-digit labels from "Data/Cleaned/ mapping.csv".

This figure can be obtained by executing the assigned part of the R script: "Codes/ 04_GenAI_exposure_task_content/Step_2_GenAI_and_task_content.Rmd". It plots the distance from the median welfare for each country by 2-digit ISCO-08 occupation.

Figure 10: The dataset for this figure is the same as for Figure 4. Results are disaggregated by welfare quintiles, across three major exposure categories (see explanation on Figure 4 for further details on the groups). Welfare data is just available for countries in SEDLAC (household income per capita) and WAEMU (household expenditure per capita). Welfare quintile cuts are estimated for each country individualy. This figure is obtained by executing the code "06-graphs.R" from "Internet_access_exercise".

Figure 11: The figure is based on the file:

- "Data/Cleaned/ PIAAC/PIAAC_4digit_with_4digit_AI_tasks_scores.xlsx"

Which contains processed data on ISCO-08 4-digit GenAI exposure disaggregated by task intensity concepts, for seven countries from PIAAC. This figure can be obtained by executing the assigned part of the R script: "Codes/04_GenAI_exposure_task_content /Step_2_GenAI_and_task_content.Rmd". It plots the relationship between GenAI exposure (x-axis, at ISCO-08 4-digit), and four concepts of task intensity: 1. Computer Use 2. Routine Manual 3. Non-Routine Analytical 4. Non-Routine Interpersonal (y-axis).

Figure 12: The dataset for this figure is the same as for Figure 11 (see explanation of the data contents above), but it uses variables: value (value of task intensity measures) and MeanScore2025 (GenAI exposure score) for the estimation of a generalized additive model (GAM) to predict task intensity at four levels of GenAI exposure across countries. Thus, the figure plots the predicted task intensity at the four levels of exposure based on the GAM results with country specific smooth terms. These results can be obtained by executing the assigned part of the R script:

"Codes/04_GenAI_exposure_task_content/Step_2_GenAI_and_task_content.Rmd".

Figure 13: The figure is based on the file:

- "Data/Cleaned/PIAAC/ 2d_step_piaac_for_pawel.csv"

Which contains processed data on GenAI exposure (ISCO-08 4-digit) and task intensity measures across countries, a similar version of the dataset used in Figure 11, but computing task scores for more countries using four data sources: STEP 2012, STEP 2013, STEP 2015/16, PIAAC first wave, and PIAAC second wave. It relies on a GAM estimation, predicting task intensity by all six exposure levels and country income groups. Thus, this figure plots the predicted task intensity by exposure levels, stratified by country income groups. These

results can be obtained by executing the assigned part of the R script: "Codes/04_GenAI_exposure_task_content /Step_2_GenAI_and_task_content.Rmd".

Figure 14: The dataset for this figure is the same as for Figure 13 (see explanation of the data contents above). The plot compares the original GenAI country-level exposure scores (left) and an adjusted exposure score based on the prediction from a GAM between mean exposure scores at the 2-digit ISCO-08 level and standardized task intensity measures. (right). These results can be obtained by executing the assigned part of the R script: "Codes/04_GenAI_exposure_task_content /Step_2_GenAI_and_task_content.Rmd".

Appendix:

Table A1: The table is provided in the Excel file "List_countries_A1.xlsx" and can be obtained by executing the code "03-graphs.R" from "ILO_country_exercise". It contains the list of countries by income group used in Figures 1 & 2, which combines both microdata available from household surveys (see explanation of the Raw Data folder above) and the Excel file containing the latest value for all countries available of the labor share at a 2-digit ISCO-08 level in "Country_emp_by_ISCO2D_ILO.xlsx".

Table A2: The table is provided in the Excel file "Population_A2.xlsx" and can be obtained by executing the code "03-graphs.R" from "ILO_country_exercise". The assigned part of the R script contacts the WB WDI database, downloads the total population (SP.POP.TOTL), the total population ages 0-14 (SP.POP.0014.TO) and the employment to population ratio ages 15+ (SL.EMP.TOTL.SP.ZS). The data are extracted from WDI API in a dynamic way; it takes the latest non-missing value available on January 12th, 2026.

Table A3: The table is provided in the Excel file "List_countries_A3.txt" and can be obtained by executing the code "06-graphs.R" from "Internet_access_exercise". It contains the list of countries with their data source and the most disaggregated ISCO codes available used in Figures 3 through 8 and 10, combining all microdata available from household surveys (see explanation of the Raw Data folder above).

Table A4: The table is provided in the Excel file "Population_A4.xlsx" and can be obtained by executing the code "06-graphs.R" from "Internet_access_exercise". The assigned part of the R script contacts the WB WDI database, downloads the total population (SP.POP.TOTL), the total population ages 0-14 (SP.POP.0014.TO) and the employment to population ratio ages 15+ (SL.EMP.TOTL.SP.ZS). The data are extracted from WDI API in a dynamic way; it takes the latest non-missing value available on January 12th, 2026.

Table A5: The table is provided in the Excel file "Model_estimates_A5.xls" and can be obtained by executing the do-file "03-mlogit-model-estimation.do" from "Mlogit_model_annex". The dataset for this table uses "microdata_for_model.rds" as an input, a similar version of the processed microdata for the exercises using internet access but keeping all individuals on the working age whether they have internet access information available or not and keeping just variables of interest. The code "02-data-preparation-mlogit.R" makes final preparations to the data and creates the final dataset used to run the multinomial logit.

Figure A15: The assigned part of the R script contacts the WB WDI database, downloads the internet users per 100 people (IT.NET.USER.ZS) and the GDP per capita (USD 2021 PPP) (NY.GDP.PCAP.PP.KD). The data are extracted from WDI API in a dynamic way; it takes the latest non-missing value available on January 9th, 2026. The figure plots the relationship of internet users (y-axis) and the log of GDP per capita (x-axis) across countries with both information available. It can be obtained by executing the assigned part of the R script: "Codes/04_GenAI_exposure_task_content /Step_2_GenAI_and_task_content.Rmd".

Figure A16: The dataset used for this figure is the same as in Table A5 (see explanation above for further details on the dataset), used to run the model results. It plots the marginal effects on the probability of exposure outcomes (augmentation potential & automation potential) with respect to the omitted outcome (not exposed). This figure is obtained by executing the do-file "03-mlogit-model-estimation.do".

References

Gmyrek, P., Berg, J., Kamiński, K., Konopczyński, F., Ładna, A., Nafradi, B., Rosłaniec, K., Troszyński, M. 2025. Generative AI and Jobs: A Refined Global Index of Occupational Exposure, ILO Working Paper 140 (Geneva, ILO). https://doi.org/10.54394/HETP0387