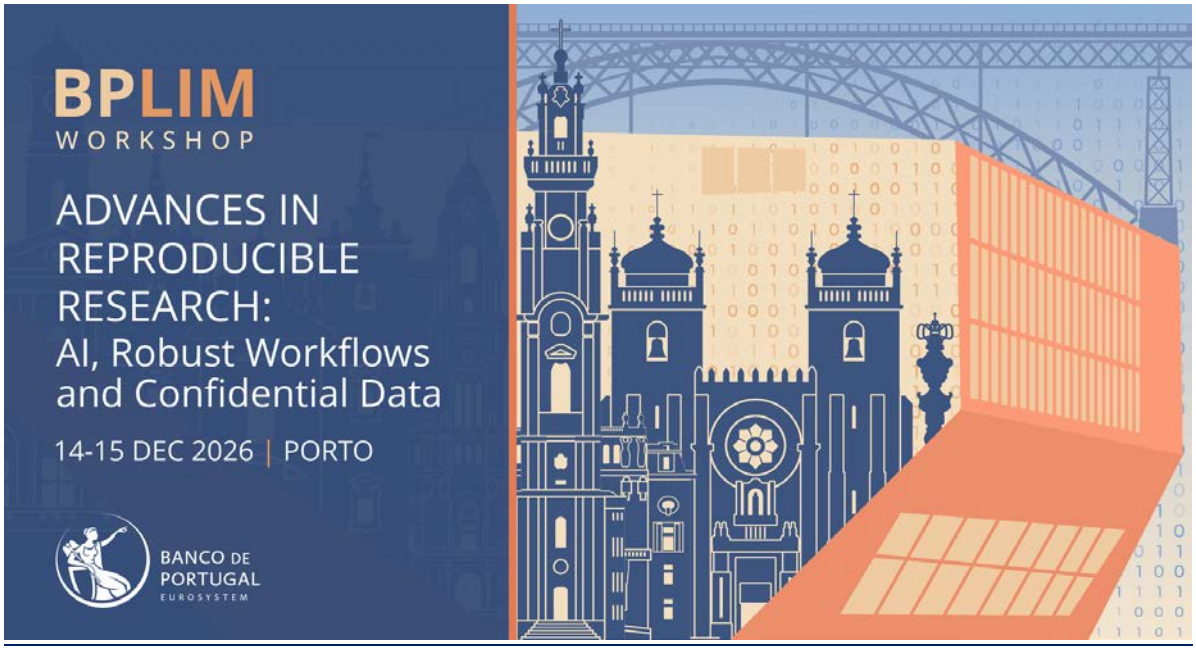
BANCO DE PORTUGAL
EUROSISTEMA



PROGRAMME

14 DEC 2026

14h00–14h15 **Welcome**

14h15–15h00 **Editorial Screening when Science is Cheap**
Gabriel Sekeres, Cornell University

New artificial intelligence tools, particularly agentic large language models, have greatly accelerated scientists ability to produce particular forms of evidence. We construct a simple model of the scientific publication process and derive conditions under which this acceleration (think of it as a shock to the cost of an additional result, broadly construed) necessarily leads to an increase in the quality of the unit of publication. Importantly, we show that current empirical social science does not meet these criteria. Intuitively, while AI agents improve an honest researcher's ability to produce evidence, they also increase the ability of a dishonest researcher to p-hack. The central issue is that social science involves "hidden" evidence, so merely producing more evidence does not necessarily improve the quality of the field. To resolve this issue, we develop an approach that tries to reveal as much information as possible. Instead of a central specification with many robustness checks, we propose to consider the entire specification surface, the set of all reasonable specifications for a given hypothesis and source of data. We show that the specification surface is an optimal mechanism for eliciting the true search process even if when omission is costless.

Finally, we show how to construct the specification surface for a given question and dataset, and in particular we demonstrate that constructing the surface is not meaningfully harder than current practice with the use of agentic systems.

15h00–15h45 **Dodo Review: Does the code back the claim?**

Miklós Koren, Central European University

Computational reproducibility establishes whether the reported results can be reproduced using the original data and code. Verifying the consistency between the artifacts in the replication package and the research claims made in the paper is an important additional step in the reproducibility process. Dodo Review is an AI-powered tool that identifies claims in academic papers and verifies them against the code, outputs, and other artifacts provided in their replication packages. For each claim, Dodo Review provides a verdict of true or false together with a justification and pointers to the relevant evidence in the code. When the available materials are insufficient to reach a conclusion, the tool explicitly identifies the claim as uncertain or unverifiable. The presentation will showcase these functionalities through practical examples, illustrating how Dodo Review can be used to assess whether research code supports the claims made in the accompanying paper.

15h45–16h15 ***Coffee Break***

16h15–17h00 ***DeadReckoning*: From Broken Research Code to a Reproducible Replication Package**

Florian Oswald, University of Turin, and Journal of Political Economy

Reproducibility has become a central requirement of empirical research. By the time a paper is accepted, the underlying research workflow has often evolved over several years across multiple machines, collaborators, software environments, and storage locations. Files have been moved, package versions have changed, intermediate scripts have been forgotten, and crucial links between results and code are frequently undocumented. Consequently, many researchers discover that reproducing their own published results from a clean environment is far more challenging than expected.

This presentation introduces *DeadReckoning*, a tool designed to help researchers make their research computationally reproducible. After checking whether the project contains confidential or restricted data, *DeadReckoning* works on a copy of the project, leaving the original files untouched. It reads the paper and traces figures and tables back to the scripts and data that produce them, flagging anything that cannot be

traced. The system then runs the code and fixes what it can, using a large language model to diagnose execution failures and propose repairs when needed. Once a functioning workflow has been established, *DeadReckoning* produces a working replication package and can also build and validate a Docker environment for submission. The presentation will demonstrate the tool through practical examples and discuss the challenges of recovering real-world research workflows.

17h00–17h45

Scalable Infrastructure for Validation of Computational Social Science Research (SIVACOR)

Lars Vilhuber, Cornell University, and American Economic Association

Scalable Infrastructure for Validation of Computational Social Science Research (SIVACOR) aims to build a scalable framework that enables users to confirm and demonstrate reproducible analysis. We build on the metadata schema developed within the TRACE project to build a system that generates a Trusted Research Object (TRO). Researchers will use the public-facing SIVACOR system to execute their analysis, generating a certified record of successful execution. The certificate assures researchers that their models are likely to run when executed by others, and can use the generated certificate to demonstrate this fact when submitting articles to a journal. Our target audience are social science researchers who ultimately need to publish in a journal that has a data and code verification process in place, but generally do not have the technical expertise to work in containerized environments or larger computational pipelines.
SIVACOR allows researchers to work in their preferred environments until the last moment, without needing to fully understand Docker or similar technologies (which remain little used in the social sciences). The system will provide a web interface where users can upload a ZIP file containing their code and data, and select from a set of curated computational environments. The SIVACOR backend will then automatically create a containerized environment, execute the analysis, and generate a TRO that includes the code, data, execution environment details, and results. This TRO can then be accessed by journal editors, replacing the often tedious and manual process of verifying the reproducibility of the submitted research.

17h45–18h30

ReRun: BPLIM's Reproducibility Tool

Gustavo Iglésias and Marta Silva, Banco de Portugal - BPLIM

Reproducibility remains particularly challenging when research relies on confidential data that cannot be shared with reviewers or data editors. ReRun was developed by BPLIM to address this challenge by providing a controlled, automated, and certifiable replication process. Building on the TRACE framework, it executes researchers' code on the original confidential data and generates a Transparent Research Object (TRO) that provides verifiable evidence of how the results were produced.

ReRun supports complex workflows involving multiple software applications and the use of containers to preserve the computational environment, including software versions and dependencies. It automatically assembles a replication package containing code, configuration files, logs, metadata, and execution records. The associated TRO provides digitally signed evidence of the integrity and provenance of these computational artifacts and their execution.
By combining automated workflow execution, documented computational environments, and digitally signed certification records, ReRun enables trusted data providers such as BPLIM to make verifiable evidence available of how research results were generated, without requiring confidential data to be shared or data editors to independently rerun the analysis.

15 DEC 2026

10h00 – 10h45 **Defensive Programming for Reproducible Research with Confidential Data**
Lars Vilhuber, Cornell University, and American Economic Association

Journals require that you share your code and data in a replication package at the end of your research project. That can be a challenge, when some parts of your research data are confidential.
Following some best practices from day 1 can not only help you prepare this package later, but also make you more productive researchers. Following some best practices before releasing a package can avoid costly revisions.
This presentation provides guidance on how to ensure the reproducibility and transparency of your research when data are confidential. It is not meant to be exhaustive, and it is not meant to be prescriptive. There are many ways to construct a replication package, and even more situations in which confidential data are housed.

10h45–11h15 ***Coffee Break***

11h15–12h00 **AI-Assisted Research with Confidential Data**
Miklós Koren, Central European University

The increasing use of AI tools in research raises particular challenges in environments that handle confidential or sensitive data. This presentation explores practical approaches to enabling AI-assisted research workflows while safeguarding confidential information. We will discuss solutions for detecting and redacting personally identifiable information and other sensitive content before interacting with external large language models, as well as the use of locally deployed models to support coding tasks within secure environments. Given the rapid evolution of these technologies, the presentation will focus on the opportunities, practical challenges, and lessons emerging from early experiments.

12h00–12h45 **Randomness in large language models: What researchers need to know (and report)**

Florian Oswald, University of Turin, and Journal of Political Economy

Large language models (LLMs) are increasingly used to generate data for research. Typical use cases are classifications, annotations, information extraction, and generation of numerical scores. Unlike conventional measurements, LLM outputs can vary across repeated requests even when the prompt and apparent model settings remain unchanged. This variation arises from deliberate sampling, silent model updates, numerical rounding, or expert routing. Setting a dedicated temperature parameter to zero removes deliberate sampling when that option is available, but it does not eliminate the other sources of randomness. Exact reproduction is therefore generally not possible when using proprietary application programming interfaces. Local execution of open-weight models offers greater control, but reproducibility still depends on the complete hardware and software stack. We illustrate these issues through sentiment classifications of corporate filings and examine their consequences for downstream regression results. We then propose a reporting standard for articles and replication packages, as well as guidance for data editors and authors. Together, these findings and recommendations establish that LLM outputs should be treated as draws from a distribution rather than as fixed measurements.

12h45–13h00 **Closing Remarks**

BIOGRAPHIES:

Florian Oswald, University of Turin, and Journal of Political Economy

Florian Oswald is an Associate Professor of Economics in the ESOMAS department of the University of Turin and an affiliate of Collegio Carlo Alberto. He currently serves as Data Editor of the Journal of Political Economy, having previously held the same position for The Economic Journal and The Econometrics Journal. His research focuses on urban economics, macroeconomics, and industrial organization, with a particular interest in computational methods.

Gabriel Sekeres, Cornell University

Gabriel Sekeres is a third-year PhD student in Economics at Cornell University, focusing on microeconomic theory. His work focuses on mechanism design and game theory applied to the context of AI, particularly problems of alignment and robust institutions.

Gustavo Iglésias, Banco de Portugal

Gustavo Iglésias is a Data Analyst at Banco de Portugal Microdata Research Laboratory (BPLIM), where he has worked since 2018. His work focuses on microdata management, statistical programming, research reproducibility, and the development of tools to support empirical research.

Lars Vilhuber, Cornell University, and American Economic Association

Lars Vilhuber is a Research Professor in the Economics Department of Cornell University and Executive Director of Cornell's Labor Dynamics Institute. He is the Data Editor of the American Economic Association and chairs the Scientific Committee of the French Secure Data Access Centre (CASD). His work focuses on labor economics, data confidentiality, research reproducibility, and transparency in scientific research.

Marta Silva, Banco de Portugal

Marta Silva is a Research Economist at the Banco de Portugal Microdata Research Laboratory (BPLIM) since 2016. She is also an Invited Assistant Professor at the School of Economics and Management of the University of Porto since 2020, where she teaches practical approaches to data management and reproducible research in economics. Her research interests include labor economics, applied microeconometrics, and policy evaluation.

Miklós Koren, Central European University

Miklós Koren is a Professor of Economics at Central European University and a Senior Research Fellow at the HUN-REN Centre for Economic and Regional Studies. He served as Data Editor of the Review of Economic Studies from 2019 to 2025. His research focuses on international trade, economic growth, and development.