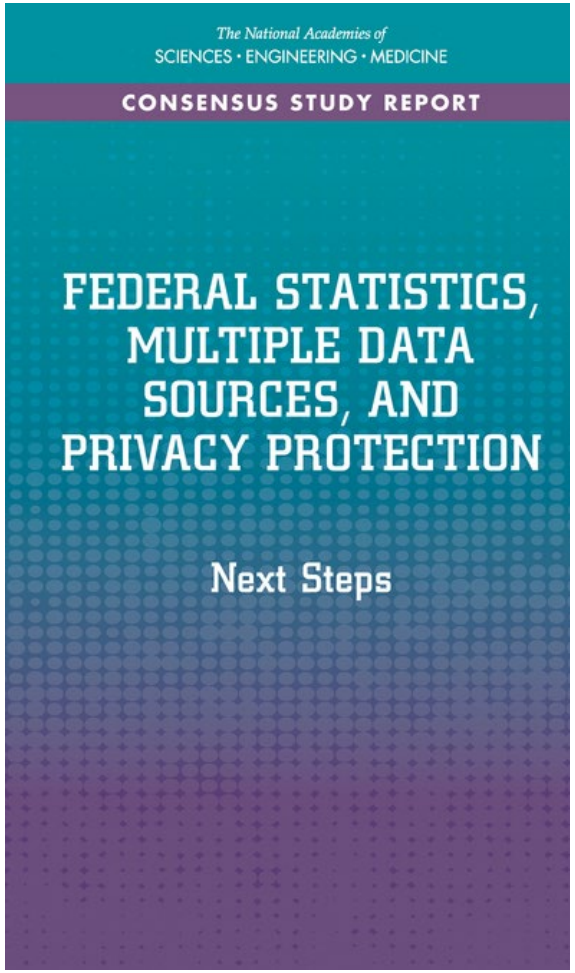
# Improving Official Statistics with AI and AI with Official Statistics

Frauke Kreuter
University of Maryland & LMU Munich
Washington DC 24th, 2026




AAPOR
AMERICAN ASSOCIATION FOR PUBLIC OPINION RESEARCH

**Responsible AI Integration in Survey Research**

David M. Rothschild (Microsoft Research)*, Jenny Marlar (Gallup)*, Ashley Amaya (Pew), Soubhik Barari (NORC at the University of Chicago), Trent Buskirk (Old Dominion University), Curtiss Cobb (Meta), Jen Gennai (T3), Sunshine Hillygus (Duke), Ramya Korlakai Vinayak (University of Wisconsin-Madison), Masha Krupenkin (University of Maryland), Sunghee Lee (University of Michigan), Darby Steiger (SSRS), Brock Webb (US Census Bureau)

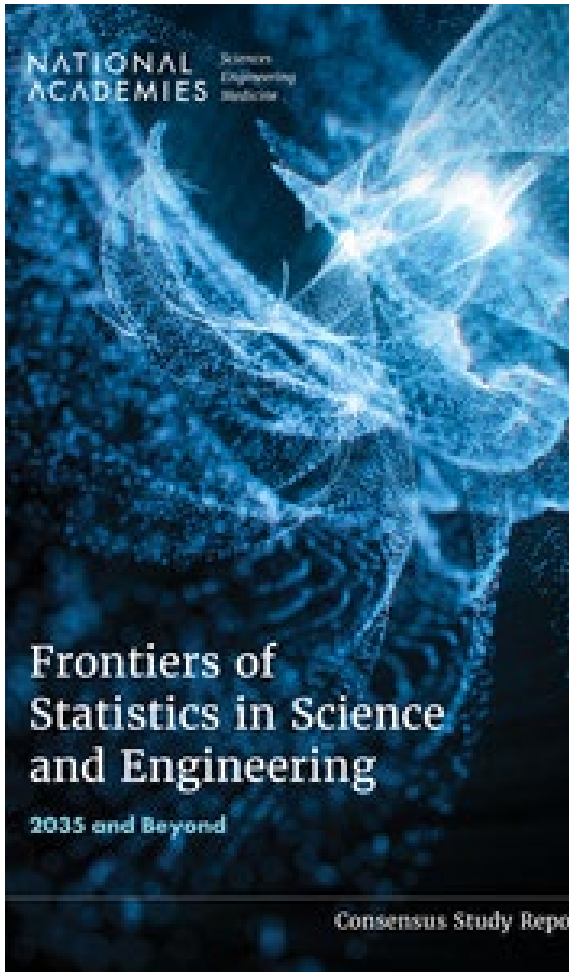
*Co-Chairs

This report was commissioned, reviewed, and accepted by the AAPOR Executive Council as a service to the profession. The opinions expressed in this report are those of the authors and do not necessarily reflect the views of the AAPOR Executive Council. The authors, who retain the copyright to this report, grant AAPOR a non-exclusive perpetual license to the version on the AAPOR website and the right to link to any published versions.

*The Task Force would like to thank Rene Bautista, Stephen Blumberg, Mario Callegaro, Stephanie Eckman, Emily Geisen, Krista Jenkins, Michael Link, Brenna Muldivan, Stanley Presser and Henning Silber for their invaluable review and feedback. The Task Force would also like to thank Hannah Cha for her support of the report and development of the disclosure tool.*

Suggested citation: Rothschild, D., Marlar, J., Amaya, A., Barari, S., Buskirk, T., Cobb, C., Gennai, J., Hillygus, D., Korlakai Vinayak, R., Krupenkin, M., Lee, S., Steiger, D., Webb, B.(2026). *Responsible AI Integration in Survey Research.* American Association for Public Opinion Research (https://aapor.org/wp-content/uploads/2026/05/Responsible-AI-Integration-In-Survey-Research.pdf).

Artificial Intelligence (AI), particularly large language models (LLM) and other generative systems, is rapidly transforming survey research across the full research life cycle: from question design and data collection to analysis and dissemination. While algorithmic tools have long influenced survey methods, recent advances in scale, flexibility, and accessibility raise novel methodological, ethical, and governance challenges that existing standards do not fully address. This report, produced by the AAPOR Task Force on Responsible AI Integration in Survey Research, provides a structured framework for understanding where and how AI is being used in surveys, evaluates associated benefits and risks, along with necessary research and development, through the lens of total survey error (TSE) and human–subject protections, and offers guidance for responsible practice. We propose principles for evaluation and human oversight and use that to introduce a practical and tractable disclosure framework designed to promote transparency, informed interpretation, and reproducibility of AI-enabled survey research. Together these contributions aim to support innovation while safeguarding data quality, trust, and credibility of survey-based research.



# Leveraging AI for Official Statistics



QUESTIONNAIRE DESIGN



SYNTHETIC DATA



INTERVIEWER

# Leveraging Official Statistics for AI



QUESTIONNAIRE DESIGN



SYNTHETIC DATA



INTERVIEWER

# tl:dr

Needs prompts that include textbooks and good articles as context.

Given good prompting, certainly a valuable help.

Very good at translating to different reading levels.




**SURV605** *(Perm Req)* **AI Assisted Surveys**

Credits: 2 Grading Method: Regular, Audit

**Prerequisite:** SURV400 or equivalent.
**Recommended:** The students should have at least one subscription to advanced LLM models such as Gemini 3 Pro, C
example. We will also experiment with open source models such as meta.ai (Llama 4).

The course introduces the students to the latest literature and practices on AI assisted surveys. For each step of the survey
well, or not, AI tools can assist the researcher in designing, fielding, analyzing, and summarizing survey findings. Before moving
we will discuss key topics such as how LLMs generate content, prompt engineering, and the Total Error framework.

Hide Sections Online

PLS1 Mario Callegaro, Olga Kononykhina Seats (Total: 15, **Open: 5**, )
**Meets September 3, 2026 - October 22, 2026**
Th 11:00am - 12:00pm ONLINE
Class time/details on ELMS ONLINE

*Golden ID students are not eligible for this section. Students must have access to a standalone web camera and a headset to participate in the weekly online meetings.*

*To apply for permission and to request a seat, click here.*



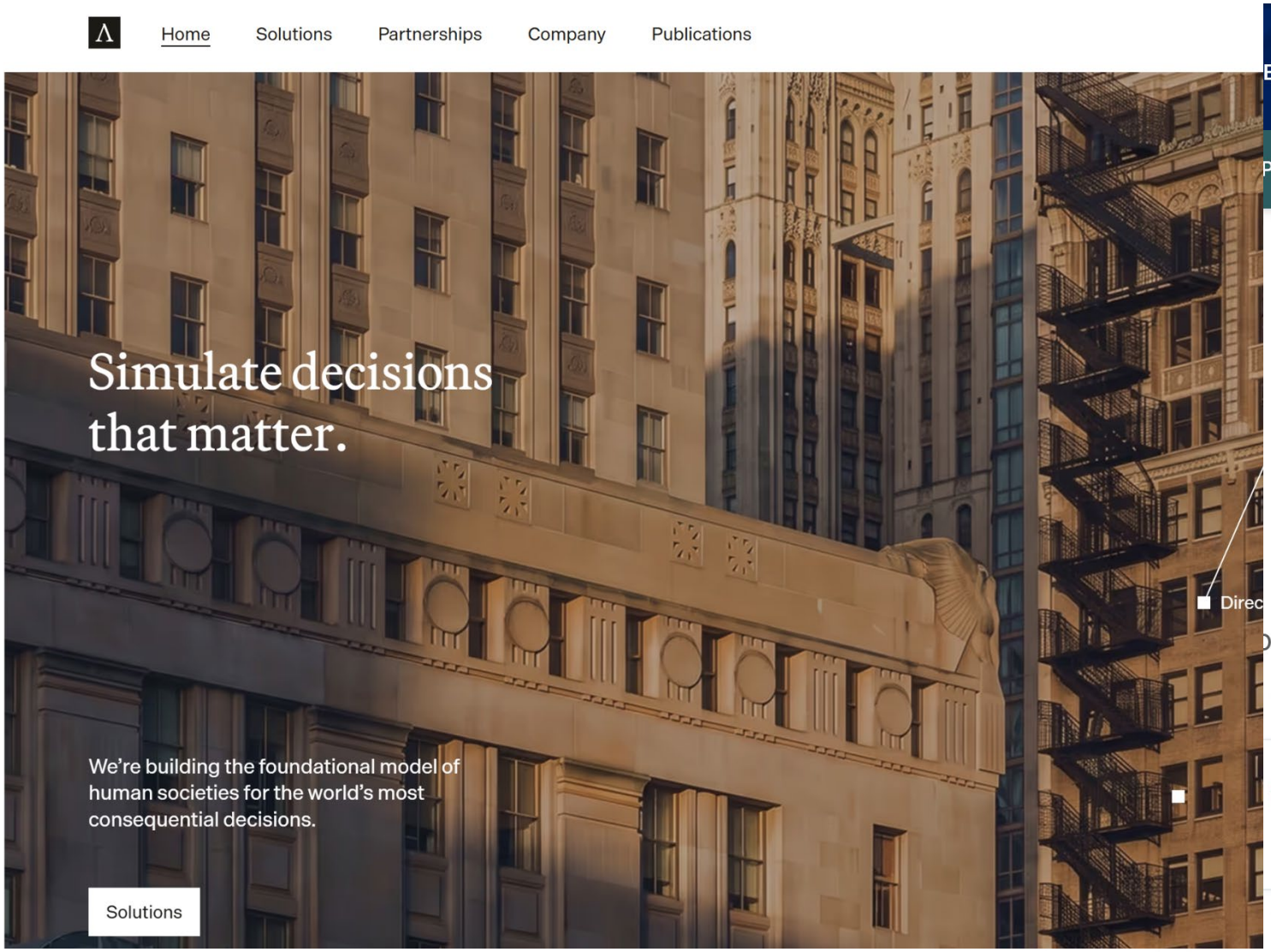
QUESTIONNAIRE DESIGN



SYNTHETIC DATA



INTERVIEWER




ERS INVESTORS ABOUT IPSOS

Python About Donate

# POLICY ENGINE

Direc

pen-source tax and benefit analysis. Model policy reforms across all 50 states.

*the poverty impact of expanding the Child Tax Credit*

Enter PolicyEngine

# tl:dr

Learning to create and use synthetic data. Imputation is promising – what is needed is an evaluation comparing different imputation techniques.

Missing data always easier to handle when the mechanism for missingness is random/known.

(Potential) biases due to training data and instruction prompts should be kept in mind.

## 30 years of synthetic data

Jörg Drechsler[1,2] and Anna-Carolina Haensch[2,3]

[1]Institute for Employment Research, Germany
[2]The Joint Program in Survey Methodology, University of Maryland, College Park, USA
[3]Ludwig-Maximilians-Universität, Munich, Germany

**Abstract**

The idea to generate synthetic data as a tool for broadening access to sensitive microdata has been proposed for the first time three decades ago. While first applications of the idea emerged around the turn of the century, the approach really gained momentum over the last ten years, stimulated at least in parts by some recent developments in computer science. We consider the upcoming 30th jubilee of Rubin's seminal paper on synthetic data (Rubin, 1993) as an opportunity to look back at the historical developments, but also to offer a review of the diverse approaches and methodological underpinnings proposed over the years. We will also discuss the various strategies that have been suggested to measure the utility and remaining risk of disclosure of the generated data.

### 1 Introduction

We live in a data-driven world today. Data are collected whenever we interact with our environment, whenever we use our loyalty card in the supermarket, measure our physical activities through wearables, when we check the online weather forecast for our weekend trip, or when we stay in contact with our friends using social media. In the public sector, the ever-growing importance of data is reflected in concepts such as evidence-based policy, and open data movements (see, for example,

Statistical Science https://doi.org/10.1214/24-STS927

QUESTIONNAIRE DESIGN



SYNTHETIC DATA



INTERVIEWER

QUESTIONNAIRE DESIGN

SYNTHETIC DATA

INTERVIEWER



# Leveraging Official Statistics for AI

“The bias I am most nervous about is the bias of the human feedback raters“

Sam Altman
March 25 2023 “The Lex Fridman Podcast“



Building up a much broader and more ingenious stable of evaluations, along with interpretability analysis to cross-check them, would be hugely valuable, and a lot of progress could be made on this in 1-2 years.

Dario Amodei September 12 2026 BlogPost



Foto: mashable.com

# *Experts - and soon expert (AI) systems – make invisible value judgements affecting our lives*

## EXAMPLE

Employment Agency



## RESEARCH EVIDENCE



Prediction models are **less accurate** for minority groups

*Kern et al. (2024) | ACM Journal on Responsible Computing | *Own work*



Training data has poor **measurement quality**

*Eckman, Plank & Kreuter (2024) | ICML | *Own work*



**Annotator feedback** needs to be adjusted to population

*Eckman et al. (2026) | NLP Perspectives | *Own work*

# WHY IS THIS URGENT?

*In 2025, >1.600 AI systems in public institutions across the EU*

Federal Ministry of Labour and Social Affairs

"JobBot"
Labor Chatbot



LMU KLINIKUM

"AVA"
Vaccination Bot



PENSIONS MYNDIGHETEN

"Penni"
Pensions Bot



UNIVERSITY OF COPENHAGEN

"Muni" Services
Chat 37 Cities



RÉPUBLIQUE FRANÇAISE

"Albert"
GenAI Platform



...

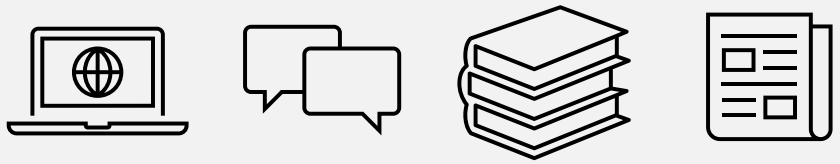
*... and likely a large number of "shadow AI"*

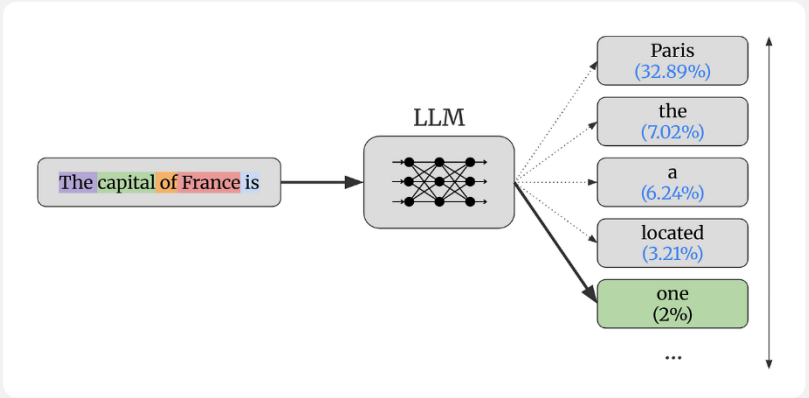
# *Post training can lead to a real change in AI*

## *...with official statistics as key player out of the crisis*

### PRE-TRAINING

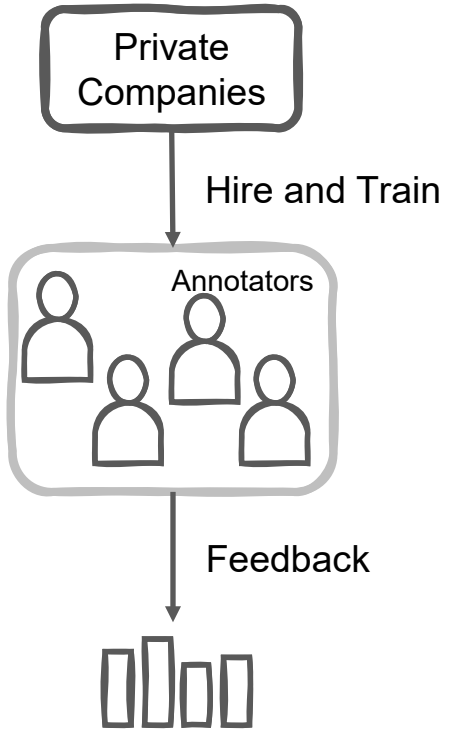
**What?** ► Models learn to **continue phrases**
**How?** ► Massive amounts of **data & compute** power






### POST-TRAINING

**What?** → Models learn to **reply conversationally**
**How?** → **Human feedback** / preferences



# *Fine-tuning on (and human feedback from) representative longitudinal data makes LLMs more equitable*

## STATUS QUO OF LLM POST-TRAINING

**Orchestrator**

Private Companies

*hires*

**Choice of annotators (Raters)**

Annotators

Convenience sample

**Training feedback**

Conceals minority voices

Common Answer

Conducted mostly at model release

## TARGET PICTURE WITH AIM-SOCIETY

Public Stakeholders

*Samples representatively*

Survey Methodology

2026
2025
2024

Representative data with targeted measurement

Privacy preserving and GDPR compliant

Full variance-covariance matrix used

Distribution

Updated with ongoing data collection

# tl;dr

Real opportunity to get new stakeholders for data collected in official statistics.

AI ready data has been tested by IMF, discussed within UN. Next step is getting them benchmark ready.

Partnerships with academic teams and continued education.

