Working title: Data Descriptor for HMC Data Set

Angelica Henestrosa

Leibniz-Institut für Wissensmedien, Tübingen

October 17, 2025

Abstract

Since the emergence of large language models (LLMs) in 2022, generative AI has rapidly expanded into mainstream applications, leading, for example, to the integration of Apple Intelligence into customer devices in 2024. This integration into personal technology marks a significant shift and a further reduction in barriers to use, bringing advanced AI capabilities into everyday devices and making them accessible to private individuals. Thus, the use of generative AI–consciously or unconsciously–along with interaction through LLM-powered (voice) assistants and engagement with AI-generated content is expected to increase significantly. However, data that link this usage to psychological variables and track it over time remain scarce. This longitudinal study comprises the data from an American sample across six waves at two-month intervals between September 2024 and July 2025. It examines user behavior, attitudes, knowledge, and perceptions related to generative AI. Thus, this data set allows for future research on psychological and behavioral dynamics of AI use over time, offering insights into user engagement and the individual factors connected to it.

1 Background and Summary

The introduction of transformer architectures in 2017 marked a major breakthrough in natural language processing (NLP), enabling significant advances in machine learning (ML) and the development of large language models (LLMs). These models, trained on vast corpora of text data, have demonstrated unprecedented capabilities in generating coherent and contextually relevant language. A milestone in public engagement with generative AI (GenAI) was the release of ChatGPT in November 2022, which made LLMs widely accessible to non-expert users. Since then, millions of individuals have interacted with conversational agents and other GenAI tools, often regularly integrating them into everyday tasks such as writing, coding, learning, and decision-making (LIT). This widespread proliferation of AI technologies, coupled with their increasingly diverse applications and personalized user experiences, raises the questions on how psychological factors shape and might explain differences in AI adoption and usage. As AI systems become more adaptive and embedded in everyday life, understanding the determinants of usage intensity, behavioral patterns, and types of use becomes essential. Moreover, the field of AI is evolving at a fast pace, and user characteristics such as attitudes and trust are subject to change over time. Therefore, longitudinal research that captures temporal fluctuations in user traits and behaviors is crucial.

Therfore, this longitudinally designed data set aims to capture the evolving perceptions of opportunities and risks associated with AI, perceived capabilities of AI systems, attitudes toward AI, trust in AI, willingness to delegate tasks to AI, areas of application, (to be continued) and the interrelationships among these constructs over time and get some hints on causality. Longitudinal studies are more likely to find changes if there is a potential change trigger (Zhao et al., 2024).

Central questions are whether predictors of technology acceptance as well as technology use change over time, whether the perception of AI-Tools as tools vs. agents (if so: what type of role/relationship) changes over time, whether this perception is related to concepts like credibility, trustworthiness, or task delegation, and whether factors such as social presence of perceived anthropomorphism mediate such processes. Long-term effects of delegating tasks to AI Tools on outcomes like perceived self-efficacy (writing skills), loneliness, or cognitive self-esteem and explore the moderating role of personality can also be explored.

This project is a joint project from the human-computer interaction group at the Leibniz-Institut für Wissensmedien in Tübingen (IWM). There are several (how many should we mention?) preregistrations from group members focusing on their individual subquestions. For an overview of the work packages and their research questions, please visit our repository https://gitea.iwm-tuebingen.de/HMC/data. Thus, this data descriptor may be used to examine research questions across the individual work packages, the possibility to extract and analyze specific subgroups or individual trajectories ignored in the work packages. Because the data set was collected shortly before the public release of Apple Intelligence on consumer devices, it offers a timely snapshot of user attitudes and behaviors at a pivotal moment in AI adoption. This context enhances the relevance of the data for understanding emerging patterns in human-AI interaction. Moreover, the findings may provide early indicators of how psychological variables such as trust, perceived usefulness, and willingness to delegate tasks relate to AI usage, potentially offering prognosis of similar developments in other countries.

2 Methods

2.1 e.g.: Participants and Data Collection

To examine those changes and relationships, an American sample mainly consisting of AI users (specify) was invited to participate in this survey at two-month intervals between September 2024 and July 2025.

This study targets an US-American sample due to Apple announcing to release its new AI platform Apple Intelligence in autumn 2024 (in the US due to the stricter regulations in the EU) and we expect many people to be exposed to this AI on their Apple devices. Data collection started at the end of August 2024?? (six waves, roughly one year).

* Prolific * Invitation * time and intervals * retention rate * second sample -> invitation of wave1 participants * focus on users -> exclusion of nousers without intention * ethics approval

2.2 e.g.: Measurements

* List of all measures by wave

We collected sociodemographic information, including, age, gender, educational level, and household income from all participants at wave 1.

3 Data Records

Data records for each of the six waves are available in CSV format at https://gitea.iwm-tuebingen.de/HMC/data together with the R scripts for data anonymization and data cleaning.

In a first step, the data was anonymized by removing participants' Prolific IDs and unused variables as well as variables only containing `NA` resulting from faulty questionnaire programming were removed. The results are six files (one for each wave) with the primary data containing the single items of each scale measured. Furthermore, variable names were harmonized and subjects excluded that filled in the survey several times. The final data sets are ready for analyses after taking some additional data preparation steps for building the scales (if desired).

Figure 1 shows the folder structure and files contained in the repository of the data records. This repostirory is generated from the local project folder that all project collaborators can access. All files are text files or PDFs with the exceptions of the codebook which is an EXCEL file. However, an export of the information contained in the EXCEL codebook to a MARKDOWN file is also included, for faster readability online and to ensure that all files are in non-proprietary formats.

Furthermore, a codebook explaining variable abbreviations and coding and containing references and information about the waves in which the variable was collected is available at https://gitea.iwm-tuebingen.de/HMC/data/src/branch/main/03_data/item_reference.md.

Table 1 provides an overview of the demographic variables over all six waves. Education and income were collected on six-point scales. Answering options for education are

1. Some high school or less
2. High school diploma or GED

```
https://gitea.iwm-tuebingen.de/HMC/data
|-- 01_project_management
|   |-- workpackages
|   |   |-- workpackages.md
|-- 02_material
|   |-- AI_Trends_Wave1_Survey.pdf
|   |-- AI_Trends_Wave2_Survey.pdf
|   |-- AI_Trends_Wave3_Survey.pdf
|   |-- AI_Trends_Wave4_Survey.pdf
|   |-- AI_Trends_Wave5_Survey.pdf
|   |-- AI_Trends_Wave6_Survey.pdf
|-- 03_data
|   |-- 01_raw_data
|   |   |-- anonymization.R
|   |-- 02_anonymized_data
|   |   |-- cleaning.R
|   |-- 03_cleaned_data
|   |   |-- HMC_wave1_cleaned.csv
|   |   |-- HMC_wave2_cleaned.csv
|   |   |-- HMC_wave3_cleaned.csv
|   |   |-- HMC_wave4_cleaned.csv
|   |   |-- HMC_wave5_cleaned.csv
|   |   |-- HMC_wave6_cleaned.csv
|   |-- HMC_codebook.xlsx
|   |-- item_reference.md
|   |-- README.md
|-- README.md
```

Figure 1: Folder structure of the repository containing the data records.

3. Some college, but no degree
4. Associates or technical degree
5. Bachelor's degree
6. Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS etc.)"

and for income

1. Less than $25,000
2. $25,000-$49,999
3. $50,000-$74,999
4. $75,000-$99,999
5. $100,000-$149,999
6. $150,000 or more.

The rate of users of AI systems increases over the six waves from about 76% to almost 90% in the sixth wave.

	Total N	User	Male	Female	Other	Age M(SD)	Education M(SD)	Income M(SD)
wave 1	1007	76.07%	500	494	13	38.68 (11.11)	4.37 (1.34)	3.51 (1.58)
wave 2	768	80.34%	375	384	8	39.37 (11.08)	4.33 (1.32)	3.50 (1.57)
wave 3	658	82.83%	318	332	6	39.86 (11.00)	4.30 (1.33)	3.51 (1.56)
wave 4	611	82.49%	282	323	5	40.13 (11.04)	4.22 (1.35)	3.43 (1.56)
wave 5	564	85.99%	259	300	4	40.43 (11.06)	4.19 (1.33)	3.42 (1.56)
wave 6	514	89.30%	238	270	5	40.36 (11.12)	4.15 (1.33)	3.36 (1.53)

Table 1: Demographic variables per wave

4 Technical Validation

Wave 1 was conducted shortly before iOs 18?? was published. -> were there any other external events potentially influencing the survey?

* Analysis of sample differences across waves -> was the sample equally distributed regarding sociodemographic characteristics?

* attention check * bot detection question * forced to respond

5 Usage Notes (optional)

Maybe here elaborate on limitations: * no data on no-users for wave 1-3 * not representative for age/gender/education/region due to focus on users * online survey: inattentive participants, fatigue effects especially in wave 1 and 6 (more variables) * rentention rate/dropout rate across waves

6 Code Availability

The primary cleaned data and accompanying R code for data anonymization and cleaning for all six waves is available at https://gitea.iwm-tuebingen.de/HMC/data. The repository and all material can be downloaded directly or cloned as a Git repository. All additional R pacckages used for data cleaning (like, e. g., `dplyr`, `qualtRics`, or `openxlsx`) are available on CRAN (https://cran.r-project.org/) and can be freely downloaded there. However, the scripts are mainly provided to make transparent which steps haven been take for data annymization and data cleaning. The data files and codebook can be downloaded and sued without having to rerun any of the scripts. We provide the data on item level here, so that they can be used for any kind of analysis. The codebook provides information needed to aggregate items into scales, e. g. which items belong to one scale and which items should be inversed before being included into the scale.

References

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, *15*(2), 1–38. https://doi.org/10.1145/3639372

Author Contributions

Competing Interests

Acknowledgements