Protocol for a Systematic Review on Transparency Practices in LLM-based Persuasion Studies

Friederike Stock Philipp Lorenz-Spreen Dirk Wulff

June 2025

1 Background

Large Language Models (LLMs) are increasingly utilized and investigated within the domain of persuasive communication, for instance, in generating personalized messages or advertisements [1]. Recent developments include interactive tools designed to produce personalized arguments to drive attitude and behavior change. One notable application involves addressing conspiracy theories: LLM-based interactive persuasion tools have demonstrated the capacity to durably reduce belief in various conspiracy theories [2], which are typically resistant to correction once established [3]. Furthermore, interactive tools have proven effective in increasing charitable donations [4], and numerous potential future applications exist, ranging from promoting climate action to encouraging health-conscious behaviors.

However, the absence of robust regulation, coupled with the potentially significant societal impact of LLM-based persuasion, raises substantial concerns regarding transparency. The nature of transparent communication (or its deficiency) carries both ethical and practical implications. As the deployment of persuasive LLM tools in behavioral studies becomes increasingly accessible, these tools can be adopted by diverse actors with varied motivations, including political agendas and commercial objectives [5]. Full transparency is crucial as it empowers the public to discern which tools they wish to engage with and facilitates necessary regulation. Conversely, a lack of transparency may foster suspicion among study participants [6] and the broader society, potentially eroding trust in LLM-based persuasion tools. Such erosion could impede research on LLM-based persuasion and its potential for addressing critical societal challenges.

This systematic review aims to comprehensively document and analyze current transparency practices in empirical studies that utilize LLMs for persuasion.

2 Research Questions and Objectives

The primary objectives of this systematic review are:

1. To identify and describe the common practices regarding transparency in published studies investigating LLM-based persuasion.
2. To determine the extent to which participants in these studies are informed about their interaction with an AI.
3. To assess the availability of key methodological details relevant to transparency, such as prompts, interaction instructions, and the specific LLMs used.
4. To document whether and how the persuasive intent of the LLM interaction (if existing) is disclosed to participants, either prior to interaction or during debriefing.

This review will focus on characterizing the existing landscape of transparency practices, not on evaluating the effectiveness of the persuasion attempts themselves.

3 Search Strategy

Our search strategy will employ two major literature databases: Web of Science (WoS) and Scopus. Both databases are recommended as primary sources for systematic reviews [7]. The search will include preprints, which are indexed by Scopus. No time limitations for publication date will be imposed. Only original research articles published in English will be included. We will also manually screen the reference list of the review by Rogiers et al. (2024) [1] for potentially relevant studies (backward citation searching). Rogiers et al. (2024) is the only review on persuasive AI that we are aware of at the moment of preregistration; however, should other review articles become available throughout the process, we will also screen their references until screening is complete.

Additionally, to identify further relevant studies, including gray literature or very recent publications, we will contact authors and researchers in the field via the following mailing lists: European Association for Decision Making, Society for Judgment and Decision Making, Cognitive Science Society, css-net.

We anticipate that the risk of publication bias significantly distorting the findings of this review is minimal, as our primary interest lies in the reported research practices concerning transparency, rather than the statistical significance or direction of experimental outcomes of the primary studies.

3.1 Inclusion Criteria

Studies will be included if they meet all of the following criteria:

- **Population:** Involves human participants.

- **Intervention/Phenomenon of Interest:** Involves a persuasion attempt. Persuasion is defined as as an intentional, goal-directed, message-based process [8].
- **Technology:** The persuasion element is explicitly powered or delivered by an LLM or a system described as AI-driven generative text model.
- **Study Type:** Empirical studies (e.g., experiments, quasi-experiments, observational studies reporting primary data).
- **Language:** Published in English.

3.2 Exclusion Criteria

Studies will be excluded if they meet any of the following criteria:

- The persuasion attempt is in the context of marketing, i.e. the targeted behavior is purchasing or the targeted attitude is regarding a brand or product.
- Conceptual, theoretical, or methodological articles without primary empirical data on LLM persuasion.
- Review articles, meta-analyses, commentaries, or editorials.
- Studies where AI/LLM is used for data analysis only, and not as the agent of persuasion.
- Full text not available in English.

3.3 Query Strings

The following query strings will be adapted for each database:

Query (Web of Science):

```
TS=(persuas* OR "attitude change" OR "behavior change" OR \chang* attitude" OR \chang* behav
```

Query (Scopus):

```
TITLE-ABS-KEY ( persuas* OR "attitude change" OR "behavior change" OR "chang* attitude" OR '
```

4 Procedure

4.1 Study Selection

Study selection will be performed in two stages using Covidence, a systematic review management software. We will pilot the screening criteria with a small subset of articles (e.g., 25-50 abstracts) among the review team to ensure consistent application before commencing full screening.

Stage 1: Title and Abstract Screening

- Titles and abstracts of retrieved records will be screened independently by one human reviewer and an LLM (Meta Llama 3.3 70B Instruct). The LLM-based review will initially be piloted on a subset of articles and validated against the ratings of a second human rater.
- Disagreements between the human and the LLM-based review will be resolved by discussion, or by a second human reviewer if consensus cannot be reached.

Stage 2: Full-Text Screening

- Articles deemed potentially eligible based on title and abstract will undergo full-text screening.
- Full texts will be independently assessed for eligibility by one human reviewer.
- In case of uncertainty, inclusion will be determined by a second human reviewer. Reasons for exclusion at this stage will be documented.

A PRISMA flow diagram will be generated to illustrate the study selection process.

4.2 Data Extraction

Data from included studies will be extracted by one human reviewer using a pre-defined data extraction form. The extracted data will include answers to questions in the form, and verbatim text of prompts and participant instructions. A human reviewer will extract answers to the questions and extract the text. An LLM will be used to independently answer the core questions, based on the manuscript and the text extracted by the human reviewer. A second human reviewer will independently verify a subset of the extracted data of 10 articles to assess interrater reliability. Any discrepancies will be resolved through discussion or consultation with a second reviewer.

We will pilot the data extraction form on a small sample of included studies (e.g., 3-5 articles) and refine it as necessary before full extraction.

If relevant information is missing from the manuscript or its supplementary materials, we will contact the corresponding authors once via email and allow a response period of two weeks.

The extracted data will be made publicly available, for example, as a supplementary file to the published review or on a repository like OSF.

Data Extraction Items (for all included articles, based on full-text and supplementary materials):

1. **General Study Information:**
   (a) Authors, year of publication, publication type (journal article, preprint, conference paper).

(b) Stated study objectives related to persuasion.

(c) Country/region where the study was conducted.

(d) Demographic characteristics of the participant sample (age, gender, and education).

2. **Persuasion Context:**

   (a) Is the persuasion interactive? (Yes/No/Partially/Not Reported)

   (b) If yes or partially, provide a brief description of the interaction.

   (c) Domain/topic of persuasion (e.g., health, political, social issue, other (specify)).

3. **Transparency Criteria:**

   (a) **Participant Awareness of AI:**

      i. Were participants informed before or during the interaction that they were interacting with an AI/LLM? (Yes, explicitly / Yes, implicitly / No / Not Reported)

      ii. If yes, how was this communicated? (Extract verbatim if possible)

   (b) **Prompt:**

      i. Is the specific prompt(s) (or key elements thereof) used to guide the LLM available? (Yes/No/Partially/Not Reported)

      ii. If yes, extract the prompt(s).

      iii. Where is the prompt located? (Main text / Supplement / Appendix / External repository / Available on request / Not available)

      iv. Does the reported prompt explicitly state a persuasion goal or persuasive strategy? (Yes/No/Unclear/Not Applicable)

      v. Does the reported prompt explicitly discourage revealing the persuasion goal or persuasive strategy to the interaction partner? (Yes/No/Unclear/Not Applicable)

   (c) **Participant Instructions for Interaction/Task:**

      i. Are the instructions given to participants regarding the interaction/task available? (Yes/No/Partially/Not Reported)

      ii. If yes, extract the instructions.

      iii. Do the instructions reveal the persuasive nature or intent of the interaction? (Yes, explicitly / Yes, implicitly (e.g., "debate an AI to convince it") / No / Not Reported)

   (d) **Debriefing:**

      i. Was a debriefing procedure reported? (Yes/No/Not Reported)

ii. If yes, did the debriefing explicitly disclose the AI's involvement (if not disclosed before)? (Yes/No/Not Applicable/Not Reported)

iii. If yes, did the debriefing explicitly disclose the persuasive intent? (Yes/No/Not Applicable/Not Reported)

(e) **LLM Specification:**

i. Which specific LLM(s) or AI system was used? (e.g., GPT-3.5, GPT-4, Claude 2, Llama 2, custom model) (Open text field – specify if "Not Reported")

4. **Study Design and Setting:**

(a) Study design (e.g., Randomized Controlled Trial, Quasi-experiment, Observational Study).

(b) Setting (Lab study / Online study / Field study / Other - specify / Not Reported).

5. **Ethical Considerations Reported by Primary Study:**

(a) Did the study report obtaining ethical approval (e.g., IRB, Ethics Committee)? (Yes/No/Not Reported)

5 Data Synthesis

The extracted data will be synthesized narratively. We will use descriptive statistics (e.g., frequencies, percentages) to summarize the prevalence of different transparency practices across the included studies. Key aspects to be summarized will include:

- The proportion of studies reporting each transparency element (e.g., disclosure of AI interaction, prompt availability, debriefing details).
- Common methods of disclosure and types of information provided.
- Differences in transparency practices based on study characteristics (e.g., interactive vs. non-interactive, study setting), if discernible patterns emerge.

We do not plan to conduct a quantitative meta-analysis of persuasion effectiveness, as the focus is on transparency practices. The synthesis will aim to provide a comprehensive overview of the current state of transparency in LLM-based persuasion research.

6 Ethics and Transparency of this Review

This systematic review will synthesize data from publicly available research. As such, it does not involve direct interaction with human participants or the

collection of primary sensitive data. Therefore, ethical approval from a research ethics committee is not required for this review itself. We will ensure that all cited works are appropriately credited.

All prompts used for the LLM review process will be made available.

7 Dissemination Strategy

The findings of this systematic review will be disseminated through:

- Publication in a peer-reviewed academic journal.
- Presentation at relevant academic conferences.
- A preprint publication, if appropriate and permitted by the target journal.
- The extracted data will be made available (e.g., via OSF or as supplementary material to the publication), as stated in the Data Extraction section.

8 Amendments to the Protocol

Any necessary amendments to this protocol that arise during the review process will be documented with justification, including the date of the change.

References

[1] Alexander Rogiers et al. *Persuasion with Large Language Models: a Survey.* 2024. arXiv: 2411.06837 [cs.CL]. URL: https://arxiv.org/abs/2411.06837.

[2] Thomas H Costello, Gordon Pennycook, and David G Rand. "Durably reducing conspiracy beliefs through dialogues with AI". In: *Science* 385.6714 (2024), eadq1814.

[3] Cian O'Mahony et al. "The efficacy of interventions in reducing belief in conspiracy theories: A systematic review". In: *PLoS One* 18.4 (2023), e0280902.

[4] Mary Phuong et al. *Evaluating Frontier Models for Dangerous Capabilities.* 2024. arXiv: 2403.13793 [cs.LG]. URL: https://arxiv.org/abs/2403.13793.

[5] Cameron R. Jones and Benjamin K. Bergen. *Lies, Damned Lies, and Distributional Language Statistics: Persuasion and Deception with Large Language Models.* 2024. arXiv: 2412.17128 [cs.CL]. URL: https://arxiv.org/abs/2412.17128.

[6] Ralph Hertwig and Andreas Ortmann. "Deception in experiments: Revisiting the arguments in its defense". In: *Ethics & behavior* 18.1 (2008), pp. 59–92.

[7] Michael Gusenbauer and Neal R Haddaway. "Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources". In: *Research synthesis methods* 11.2 (2020), pp. 181–217.

[8] Marco Dehnert and Paul A Mongeau. "Persuasion in the age of artificial intelligence (AI): Theories and complications of AI-based persuasion". In: *Human Communication Research* 48.3 (2022), pp. 386–403.