Plan Overview

A Data Management Plan created using DMPonline

Title: Highly Realistic Synthetic Media Online: Diffusion and Moderation Strategies

Creator: Marilù Miotto

Principal Investigator: Marilù Miotto

Contributor: Ting Li, Anna Priante

Affiliation: Erasmus University Rotterdam

Funder: Netherlands Organisation for Scientific Research (NWO)

Template: Data Management Plan v4.5

ORCID iD: 0009-0007-8612-9225

Project abstract:

As AI tools spread, generating highly realistic synthetic media (deepfakes) has become increasingly accessible. What once required specialized technical skills can now be achieved with widely available tools, leading to a proliferation of user-generated deepfake content. Importantly, this content can be innocuous, such as memes, fictional stories, or educational material, but it can also be extremely harmful, including disinformation, fraudulent advertisements impersonating public figures, and non-consensual sexual imagery.

Thus, this project aims to study deepfakes, deepfake-related discussions, and possible moderation strategies. In this context, moderation presents a particularly challenging task. While regulation increasingly requires the labeling of deepfake content, it is important that labels and other moderation strategies limit the spread of harmful content without negatively affecting non-harmful content, including authentic content appearing alongside moderated deepfakes.

This project is articulated around three distinct papers. The first paper investigates the efficacy of labels in reducing users' willingness to share disinformative political deepfakes, while also accounting for potential second-order effects on authentic content, particularly news, appearing alongside labeled deepfakes. The second paper takes a longitudinal approach to investigate the diffusion of different deepfake frames across Reddit communities, focusing on the interaction between these frames and community-specific moderation rules. It assesses the efficacy of moderation policies in limiting the diffusion of harmful frames while examining potential second-order effects, such as the migration of these frames to less-moderated communities within the same platform. The third paper further investigates how moderation policies evolve within online communities, examining how changes to subreddit rules on deepfake content interact with the diffusion and evolution of deepfake-related frames.

Together, these papers aim to provide a broader understanding of the effectiveness of deepfake moderation strategies and their potential unintended consequences.

ID: 149364

Start date: 01-09-2023

End date: 31-08-2027

Last modified: 23-10-2026

Grant number / URL: 14000022.003

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Highly Realistic Synthetic Media Online: Diffusion and Moderation Strategies

General

Please tick the following boxes if you agree to act according to the following terms:

- I will answer all questions truthfully and to the best of my knowledge
- I will discuss the data management plan with my research team
- I will check and, if necessary, update my data management plan a minimum of once a year

Support in writing a data management plan is available through the faculty Data Stewards. Which research support professional is available for you?

- Data Steward of my own faculty - RSM

Scientific research must be conducted in line with existing guidelines on good research practices and integrity. Please tick the boxes if you have read and understand these guidelines and will act accordingly.

- The European Code of Conduct for Research Integrity (ALLEA, 2017)
- The Netherlands Code of Conduct for Research Integrity (VSNU, 2018)

Administration & Project Description

Project title

Highly Realistic Synthetic Media Online: Diffusion and Moderation Strategies

Project start date as intended

2023-09-01

Project duration in months as intended

48

Funding body (if applicable)

Public Values in the Algorithmic Society (AlgoSoc) - NWO

Grant number (if applicable)

024.005.017

Date of DMP Version 1

2024-09-01

Current DMP - Version [if other than version 1]

5

Current DMP - Date [if other than version 1]

2026-10-01

List the name and affiliation of all members of the research team.
List the researcher responsible for research data management first.
For PhD projects, please indicate the Promotor(s) and/or Daily Supervisor(s) with a (!)

	Name	Email	ORCID	Research Institution
1	Marilù Miotto	miotto@essb.eur.nl	https://orcid.org/0009-0007-8612-9225	Erasmus Graduate School of Social Sciences and the Humanities
2	(!) Ting Li	tli@rsm.nl	https://orcid.org/0000-0001-7291-3050	Erasmus Research Institute of Management
3	(!) Anna Priante	priante@rsm.nl	https://orcid.org/0000-0003-1536-3755	Erasmus Research Institute of Management
4				
5				
6				
7				
8				
9				
10				

Briefly summarize the project background and research question(s) to help others understand the purpose for which the data are being collected or created

As previously described, the project focuses on moderating deepfake content and deepfake-related discussions, with particular attention to potential unintended effects. More broadly, it addresses the following overarching research questions: *How effective are different moderation strategies in limiting harmful deepfake content and discussions? What unintended effects can these strategies produce on other content, the diffusion of deepfake-related discussions, and the norms governing online communities?*

The first study investigates how labeling deepfake content affects users' perceptions of the labeled content itself, as well as their perceptions of other content appearing alongside it, and how these changes in perception ultimately shape sharing behavior. It therefore examines the effectiveness of labels in limiting the spread of harmful content while accounting for potential spillover effects on authentic content.

The second study investigates how deepfake-related discussions evolve online, focusing specifically on Reddit, a social media platform hosting discussions on a wide range of topics, including technology. In particular, it distinguishes between different ways of discussing and framing deepfake technology, differentiating between harmful and non-harmful frames. Adopting a longitudinal approach, the study examines the diffusion of these frames across the platform and how moderation rules within individual sub-communities affect their diffusion. Importantly, it considers the effects of moderation at both the sub-community and platform levels. At the sub-community level, it investigates whether moderation limits the spread of harmful discussions within individual communities. At the platform level, it focuses on potential migration effects, examining whether discussions disappear from the platform following moderation or instead migrate to less-moderated communities.

Finally, the third study also relies on Reddit discussions but focuses specifically on how moderation norms and rules are negotiated within individual sub-communities. Rather than examining the platform as a whole, it investigates how external events related to deepfake technology, such as the release of new technologies or major political events, affect communities' understanding of deepfakes and, consequently, their approach to moderating related content and discussions. In doing so, it examines how moderation norms evolve in response to changing technological and social contexts.

Specify the research type and briefly describe the methodology, the types of data to be generated and/or collected, and the tools used for data collection

Study 1. Experiment (online survey). Participants view a series of posts in a social media-like setting, evaluate each post's credibility, and report different measures of engagement (e.g., intentions to share, comment, or like). The survey also includes control questions concerning news consumption, familiarity with AI, and demographic characteristics. The survey is built on Qualtrics and distributed through Prolific (all participant data are anonymized). RStudio is used to analyze the data.

Studies 2 and 3. Secondary data (publicly available Reddit data). Both studies rely on publicly available Reddit data (already obtained through an API and made public by third parties). The second study uses user posts and interactions to reconstruct the community network and evaluate diffusion patterns. Textual data are analyzed to identify different deepfake-related frames through topic modeling and to examine community-specific moderation strategies. The third study focuses on discussions within selected sub-communities, using natural language processing (NLP) techniques to extract recurring conversation themes and analyze sentiment, with the aim of understanding how community norms and moderation rules evolve over time. Python is used to

process the data, either on a personal laptop or on Google Colab when a GPU is needed. For NLP methods, we test a selection of LLMs for test screening (from the original data) and proceed with the one that performs best (LLM APIs used: OpenAI, Mistral, DeepSeek, Gemini); topic modeling uses BERTopic.

Specify the (financial and time) resources needed for data management in this project

The research costs, including participant compensation for the online experiment, are covered by AlgoSoc, an NWO-funded project. During data processing, data will be stored on personal external drives and subsequently transferred to the EUR Yoda Research Portal for secure storage. The time required for data management is allocated within the PhD candidate's research activities.

Preparation: Legal Arrangements and Policy

1. With whom will you need to make legal arrangements?

- With research participants

2. List the agreements that you will initiate and with whom will you make them.

Who	Type of agreement
Research Participants	Informed Consent

3. List the agreements or other data management policies that you need to uphold but did not initiate. If you are reusing existing data, list the terms of use under which you may re-use them.

Who	Type	Version and Date
EUR	RDM policy of Erasmus University Rotterdam (EUR)	Version 1.0 [August 14th, 2020]
EUR	Internet and ICT facilities policy	[July 13th, 2021]
NWO	Funding agreement	[October, 2022]

4. Do you need to obtain ethical approval for your research project?

- Yes, I have obtained ethical approval

5. If you have obtained ethical approval, list the reference number

ETH2324-0732 (RSM, IRB-E)

During research: Collecting and analyzing

6. Specify what data you will be collecting and indicate format, estimated size, and whether this is data that you will be generating or existing data that you will be re-using.

Type	Format	Estimated size	Generate	Re-use	Data Classification (optional)
Digital survey data	.csv	1-5 GB	Yes	No	Internal
Reddit API data	.zstd	500 GB	No	Yes	Internal
Reddit filtered data	.csv	10-50 GB	Yes	Yes	Internal

7. Will you be collecting or re-using (sensitive) personal data?

- Yes – Personal data that is non-sensitive -->Consult your faculty's Privacy Officer

8. If you collect or re-use (sensitive) personal data, how will you protect the privacy of participants when sharing your data?

- I will fully anonymize the data

9. Will you be collecting or re-using non-personal sensitive data?

- No

10. Where will you store your data during the project? You can select multiple options.

- EUR SURF Yoda
- EUR OneDrive
- Other (please specify where the data will be stored in additional information box)

Personal SSD

12. What hardware and software do you use? Select all applicable options.

- EUR supported software as found in the software catalog
- EUR supported hardware [e.g. @wEURk laptop, @wEURk workstation]

13. If you use private hardware, software, or freeware, please specify what and for what reason:

Question not answered.

14. Are regular backups made of your data?

- Yes, manually (please specify WHO makes the backups and HOW OFTEN backups are made in the additional information box).

When data is not on EUR-supported tools, backups are done manually weekly, by the PhD candidate.

15. Who manages access to the data?

- Researcher responsible for research data management

16. Who will have access to the data (during the project)?

- Only researchers as indicated under ‘Administration & Project description’

17. How are you going to make sure your data will be accessible in case of staff changes, illness, etc?

- I have not yet discussed this with the research team or checked with my department or faculty

While I haven't discussed it yet, as a PhD candidate and this being part of my dissertation, I'm the only person that is actually working on the data.

18. Have you and your research team agreed on a way to name and order project folders and files?

- Yes - I am working on the documentation

19. Have you and your research team agreed on how to handle versioning of files?

- Yes - I am working on the documentation

Research Publication: Data sharing and re-use

20. What data (and code) will be shared in a research data repository?

- All data (and code) underlying published papers / reports

21. Please specify why you are unable to share (all) data (and code)

The raw data contains a Prolific ID that might be linked to the participants. So, when sharing the data, it's essential to anonymize it. Otherwise, the data and code can be shared.

For publicly available Reddit data, we will make the code and aggregated data available to avoid sharing users' usernames (though Reddit usernames are almost always pseudonyms).

22. List the data (and code) that you plan to share in a research data repository. Also list the information / documentation / metadata that you will include to make the data package self-explanatory and re-usable in the future (for other researchers and yourself)

Data	Format	Size
Anonymized survey data (study 1)	.csv	< 1 GB
Python code for data pre-processing (study 1)	.py	< 1 GB
R code for data analysis (study 1)	.r	< 1 GB
ReadMe text file (study 1)	.rtf (probably, or .txt)	< 1 GB
Python code for data filtering (study 2 & 3)	.py	< 1 GB
Python code for text classification (study 2 & 3)	.py	< 1 GB
LLM output for text classification (study 2 & 3)	.csv	1 - 5 GB
Python code for network construction (study 2)	.py	< 1 GB
Aggregated network data for diffusion analysis (study 2)	.csv	5 - 10 GB
ReadMe text file (study 2)	.rtf (probably, or .txt)	< 1 GB
ReadMe text file (study 3)	.rtf (probably, or .txt)	< 1 GB

23. In which repository will you place the metadata, data, and/or code that are associated with your paper?

- EUR Data Repository (EDR)

24. What metadata standard will you use to document your research?

- I do not know
- DDI (Data Documentation Initiative)

Not discussed yet, possibly DDI (following EUR policies and previous feedback advise).

25. Will you place any restrictions on re-using of data?

- No

28. Under what license will you make your data available for re-use?

- Creative commons (e.g. CC0 or CC-BY, please specify in Q.29)

29. Please specify which license

Don't know yet, possibly CC-BY (or CC-BY-NC-SA)

After research: Archiving

30. You may be obliged to destroy some data before archiving. Do any of such obligations apply to you?

- Yes - Other (please specify in the additional information box).

Prolific IDs that might allow tracing back to the participant's identity (Study 1).

31. List the data and all documentation you will be archiving. These data constitute your archival package.

Data	Format	Size
Informed Consent forms (copy; Study 1)	.pdf	<1 GB
Copy of the survey (Study 1)	.pdf	<1 GB
Digital survey data (Study 1)	.csv	1-5 GB
Digital survey data processing and analysis code (Study 1)	.r	<1 GB
Reddit filtered data (Studies 2; 3)	.csv	10-50 GB
Reddit data processing code (Studies 2; 3)	.py	<1 GB
Reddit analysis data (Studies 2; 3)	.csv	1-5 GB
Ethical approval copy	.pdf	<1 GB
DMP (final version)	.pdf	<1 GB

32. Where will you be archiving your data?

- I do not know