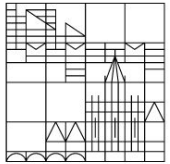
Reproducible Research and Ethical Considerations

Indira Sen

17.06.24

Universität Konstanz



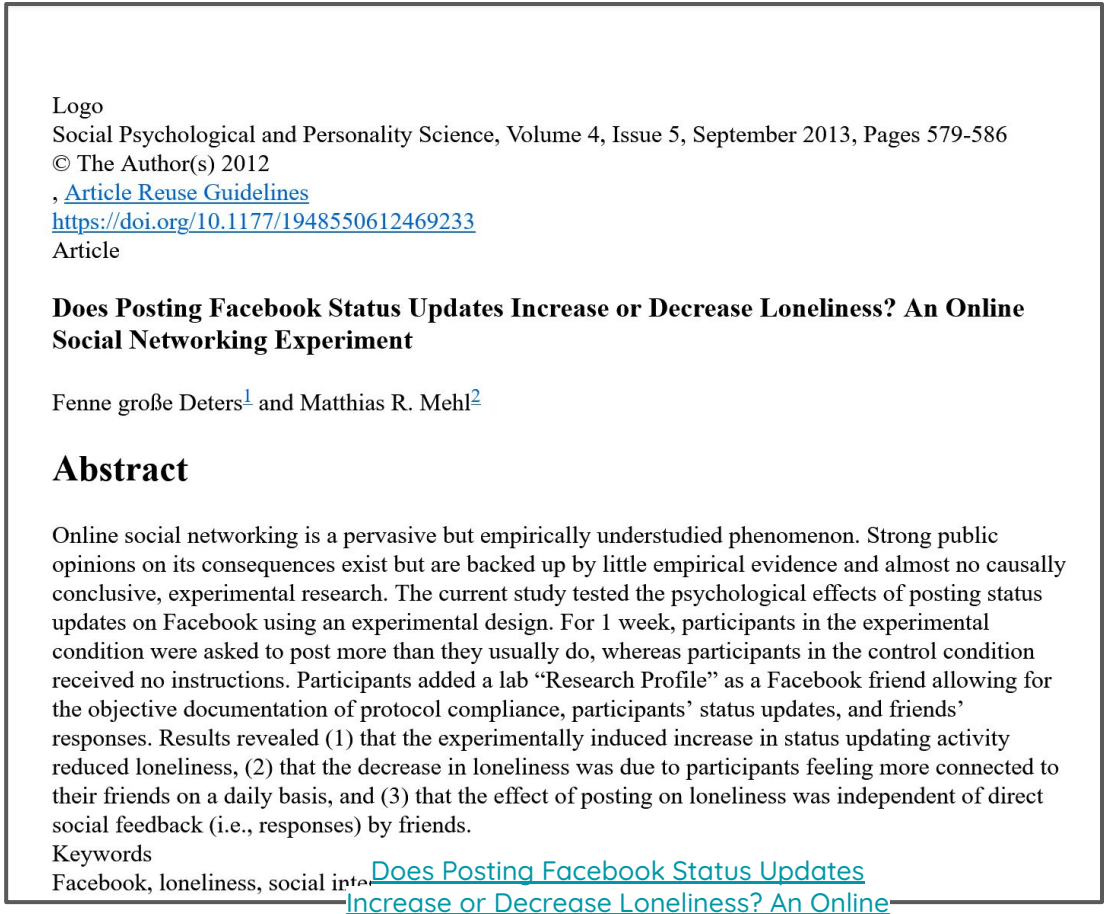
Agenda

- Recap
- Today:
  - Reproducible research pipeline
    - Data sharing
    - Documentation
    - Best Practices
  - Ethical considerations for CSS
    - Ethics Frameworks
    - Dual use
  - Filling out an ethics checklist

Recap: Causal Inference

Causal Inference for Experiments

- In the lab
  - High internal validity
  - Might lack external validity



Logo
Social Psychological and Personality Science, Volume 4, Issue 5, September 2013, Pages 579-586

https://doi.org/10.1177/1948550612469233
Article

Does Posting Facebook Status Updates Increase or Decrease Loneliness? An Online Social Networking Experiment

Fenne große Deters[1] and Matthias R. Mehl[2]

Abstract

Online social networking is a pervasive but empirically understudied phenomenon. Strong public opinions on its consequences exist but are backed up by little empirical evidence and almost no causally conclusive, experimental research. The current study tested the psychological effects of posting status updates on Facebook using an experimental design. For 1 week, participants in the experimental condition were asked to post more than they usually do, whereas participants in the control condition received no instructions. Participants added a lab "Research Profile" as a Facebook friend allowing for the objective documentation of protocol compliance, participants' status updates, and friends' responses. Results revealed (1) that the experimentally induced increase in status updating activity reduced loneliness, (2) that the decrease in loneliness was due to participants feeling more connected to their friends on a daily basis, and (3) that the effect of posting on loneliness was independent of direct social feedback (i.e., responses) by friends.

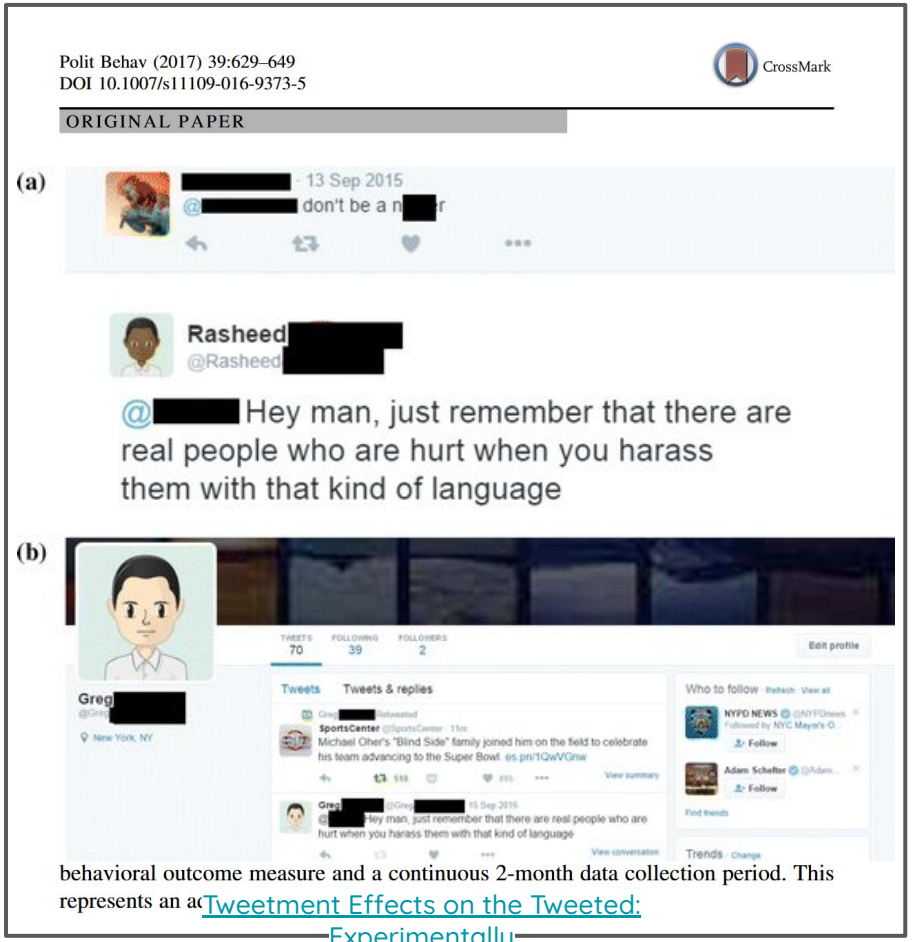
Keywords
Facebook, loneliness, social inte

Does Posting Facebook Status Updates Increase or Decrease Loneliness? An Online Social Networking Experiment

Causal Inference for Experiments

- In the lab
  - High internal validity
  - Might lack external validity

- **In the “wild” or field experiments**



[Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment]

Propensity score matching is quite popular in text + CSS

Paper	Treatment	Outcome(s)	Confounder	Text data	Text rep.	Adjustment method
Johansson et al. (2016)	Viewing device (mobile or desktop)	Reader's experience	News content	News	Word counts	Causal-driven rep. learning
De Choudhury et al. (2016)	Word use in mental health community	User transitions to post in suicide community	Previous text written in a forum	Social media (Reddit)	Word counts	Stratified propensity score matching
De Choudhury and Kiciman (2017)	Language of comments	User transitions to post in suicide community	User's previous posts and comments received	Social media (Reddit)	Unigrams and bigrams	Stratified propensity score matching
Falavarjani et al. (2017)	Exercise (Foursquare checkins)	Shift in topical interest on Twitter	Pre-treatment topical interest shift	Social media (Twitter, Foursquare)	Topic models	Matching
Olteanu et al. (2017)	Current word use	Future word use	Past word use	Social media (Twitter)	Top unigrams and bigrams	Stratified propensity score matching
Pham and Shen (2017)	Group vs. individual loan requests	Time until borrowers get funded	Loan description	Microloans (Kiva)	Pre-trained embeddings + neural networks	A-IPTW, TMLE
Kiciman et al. (2018)	Alcohol mentions	College success (e.g. study habits, risky behaviors, emotions)	Previous posts	Social media (Twitter)	Word counts	Stratified propensity score matching
Sridhar et al. (2018)	Exercise	Mood	Mood triggers	Users' text on mood logging apps	Word counts	Propensity score matching
Saha et al. (2019)	Self-reported usage of psychiatric medication	Mood, cognition, depression, anxiety, psychosis, and suicidal ideation	Users' previous posts	Social media (Twitter)	Word counts + lexicons + supervised classifiers	Stratified propensity score matching
Sridhar and Getoor (2019)	Tone of replies	Changes in sentiment	Speaker's political ideology	Debate transcripts	Topic models + lexicons	Regression adjustment, IPTW, A-IPTW
Veitch et al. (2019)	Presence of a theorem	Rate of acceptance	Subject of the article	Scientific articles	BERT	Causal-driven rep. learning + Regression adjustment, TMLE
Roberts et al. (2020)	Perceived gender of author	Number of citations	Content of article	International Relations articles	Topic models + propensity score	Coarsened exact matching
Roberts et al. (2020)	Censorship	Subsequent censorship and posting rate	Content of posts	Social media (Weibo)	Topic models + propensity score	Coarsened exact matching

Table 1: Example applications that infer the causal effects of treatment on outcome by measuring confounders (unobserved) from text data (observed). In doing so, these applications choose a representation of text (text rep.) and a method to adjust for confounding.

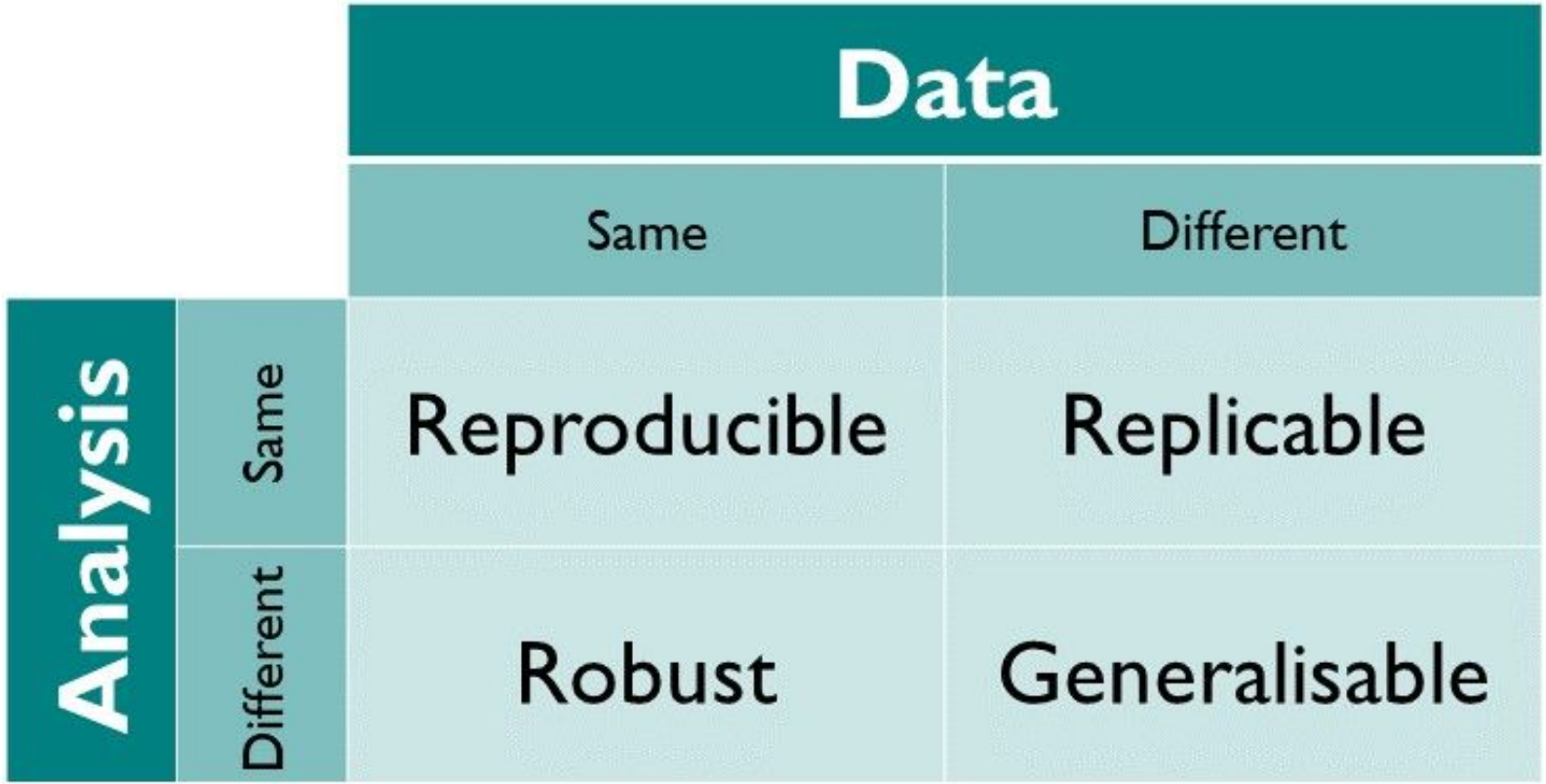
Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates

Comparison of different approaches

Approach	Key assumptions	Strengths	Limitations
Regression-based approach[1]	Error term is uncorrelated with all regressors	Easy to implement. Well-developed literature on mediating/moderating and nested models.	No clear distinction between treatment and covariates.
Propensity score approach	Strong ignorability (i.e., selection on observables only)	Estimates causal effect at a given time. Explicit consideration of all variables that relate to treatment assignment. Sensitivity analysis for violation of strong ignorability assumption.	Diagnostics for adequacy of propensity score model and methods for estimating mediating/moderating /nested effects are still in early stages. Requires large sample sizes.
RD approach	Strong ignorability	Allows estimation of treatment effect if treatment assignment changes discontinuously on the basis of some Z.	Requires making some assumptions and extrapolations for control units in the range of Z, where we do not have any control units, and vice versa.
Dummy endogenous variable approach	Error terms of selection and outcome equations are linearly related and bivariate normal	Allows error terms of selection and outcome equations to be correlated.	Linearity and bivariate normality assumptions are not testable. Both researchers and managers have difficulty in conceptualizing and understanding these assumptions.
IV approach	Exclusion restriction[2]	Allows estimation of causal effects when treatment variable is endogenous.	Exclusion restrictions are not testable and rarely justifiable. Large standard errors if sample sizes are small or instruments are weak. Assumes a constant treatment effect for all individuals.

From Association to Causation Via a Potential Outcomes Approach

Reproducible Research Pipeline



https://book.the-turing-way.org/reproducible-research/overview/overview-definitions

Why is reproducibility and replicability important?

- Cornerstone of good science: reproducibility of empirical results is an essential part of the scientific method

- Ensure trustworthiness, rigor, and transparency of your findings

- Selfish reasons:
  - Better for you communicate what you did
  - Prevent mistakes:
    - Science is **iterative** and **self-correcting** —> honest mistakes happen
    - Providing reproducible materials out means that others can critically analyze and correct mistakes

A Beginner's Guide to Conducting Reproducible Research

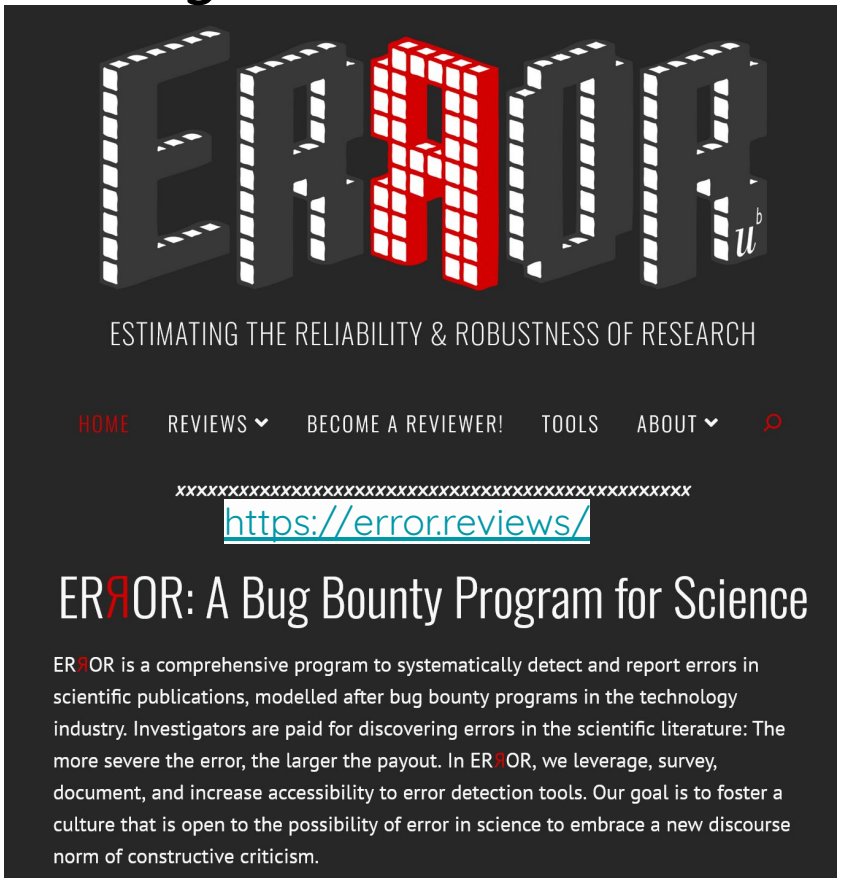
Reasons behind the reproducibility crisis

- Fraud
- Technical debt:
  - “There’s a deadline, I’ll do this later”
  - “It’s just exploratory right now…”
- Lack of awareness of good practices
- Lack of incentives

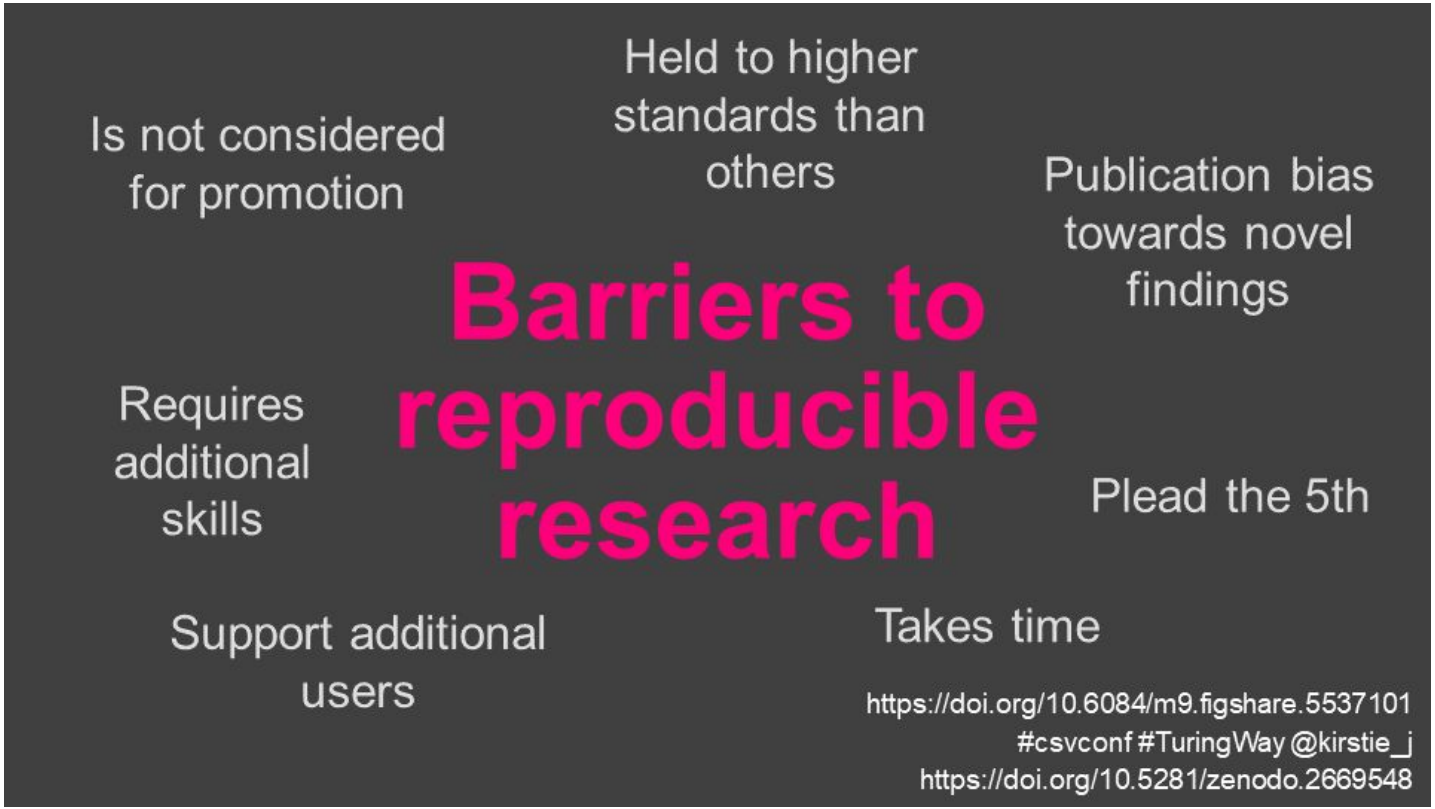
https://datacolada.org/109

Reasons behind the reproducibility crisis

- Fraud
- Technical debt:
  - “There’s a deadline, I’ll do this later”
  - “It’s just exploratory right now…”
- Lack of awareness of good practices
- Lack of incentives



Common barriers to reproducibility



Common barriers to reproducibility

- Data
  - Inaccessible data
  - Poor or insufficient documentation
- Proprietary analysis tools, e.g., models or software
- Lack of reliability measures
  - **TIP:** always include error bars, confidence intervals
- Underpowered experiments
  - **TIP:** try to do a power analysis if you already have some exploratory results

Common barriers to reproducibility for CSS research

- Social media data sharing
  - Dataset decay
  - Tensions with welfare, privacy, and informed consent
  - Secondary data brokers providing little information about **data provenance**

Dataset	Lang	Size	Avail	Construct	Sharing	Hosting
Waseem2016 (Waseem and Hovy 2016)	EN	6,909	91/54%	HS	IDs	GitHub
WaseemHovy2016 (Waseem and Hovy 2016)	EN	16,791	64/50%	HS (sexism, racism)	IDs	GitHub
BenevolentSexism (Jha and Mamidi 2017)	EN	7,205	-/33%	HS (sexism)	IDs	GitHub
Davidson2017 (Davidson et al. 2017)	EN	25,296	•	HS & OL	Tweets	GitHub
Golbeck2017 (Golbeck et al. 2017)	EN	35,000	•	Harassment	Tweets on req	SH
Ross2017 (Ross et al. 2016)	GER	477	•	HS (immigrants)	Tweets	GitHub
Bohra2018 (Bohra et al. 2018)	EN,HI	4,067	69/69%	HS	IDs	GitHub
ElSherief2018 (ElSherief et al. 2018)	EN	28,498	-/32%	HS	IDs	GitHub
Founta2018 (Founta et al. 2018)	EN,HI	79,894	61/39%	HS	IDs / Tweets on req	SH / Zenodo
GermEval2018 (Wiegand, Siegel, and Ruppenhofer 2018)	EN,HI	8,541	•	OL	Tweets	GitHub
IberEval2018 (Fersini, Rosso, and Anzovino 2018)	EN	3,977	•	Sexism, misogyny	Tweets on req	SH
Rezvan2018 (Rezvan et al. 2018)	EN/HI	24,189	•	Harassment	Tweets on req	SH
Ribeiro2018 (Ribeiro et al. 2018)	EN	4,972	•	HS	Network(req)	GitHub / Kaggle
HASOC19 (Mandl et al. 2020)	EN,HI,GER	7,005	•	HS & OL	Tweets	SH
HatEval (Basile et al. 2019)	EN,SP	13,000	•	HS	Tweets	SH
OLID (Zampieri et al. 2019)	EN	14,100	•	OL	Tweets	SH
Ousidhoum2019 (Ousidhoum et al. 2019)	EN,AR	5,647	•	HS	Tweets	GitHub
Toosi2019 (Toosi 2019)	EN	31,961	•	HS sentiment	Tweets	Kaggle
ALONE (Wijesiriwardene et al. 2020)	EN	688	•	HS/OL (Toxicity)	Tweets on req	SH
Gomez2020 (Gomez et al. 2020)	EN	149,823	48/44%	Multimodal HS	Tweets, IDs, Img	SH
MeTooMA (Gautam et al. 2020)	EN	9,973	78/78%	HS	IDs	Dataverse
CMSB (Samory et al. 2021)	EN	2,743	87/85%	Sexism	Tweets on req	Datorium
Covid2021 (Wich, Räther, and Groh 2021)						
SWAD						GitHub

[The End of the Rehydration Era The Problem of Sharing Harmful Twitter Research Data]

Common barriers to reproducibility for CSS research

- Social media data sharing
  - Dataset decay
  - Tensions with welfare, privacy, and informed consent
  - Secondary data brokers providing little information about **data provenance**
- Use of black-box services
  - Little information about service development (e.g., ChatGPT, Perspective API, Botometer)

[On the Challenges of Using Black-Box APIs for Toxicity Evaluation in Research]

On the Challenges of Using Black-Box APIs for Toxicity Evaluation in Research

Luiza Pozzobon[†] Cohere For AI luiza@cohere.com
Beyza Ermiş Cohere For AI beyza@cohere.com
Patrick Lewis Cohere patrick@cohere.com
Sara Hooker Cohere For AI sara@cohere.com

Abstract

Perception of toxicity evolves over time and often differs between geographies and cultural backgrounds. Similarly, black-box commer-

Automatic toxicity detection tools, which often use machine learning algorithms to quickly analyze large amounts of data and identify patterns of toxic language, are a popular and cost-effective method

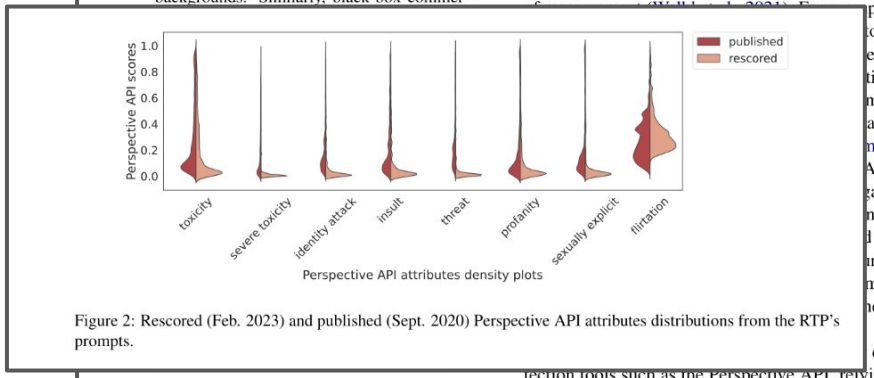


Figure 2: Rescored (Feb. 2023) and published (Sept. 2020) Perspective API attributes distributions from the RTP's prompts.

legal frameworks

research ethics

changing platform
access options / ToS

user expectations /
privacy

Tensions

platforms as
black boxes

data access

data sharing

interdisciplinarity

methods as
black boxes

publishing
practices

missing data

different conclusions when it comes to addressing specific challenges

E.g., in the context of research ethics:

- ? Big vs. small data
- ? Users as authors vs. users as research subjects
- ? Particularly vulnerable groups (e.g., activists) vs. professional / public accounts (e.g., politicians)
- ? Different practices in quoting from user accounts based on disciplinary requirements

Weller, Katrin, and Katharina E. Kinder-Kurlanda. 2017. "To Share or Not to Share?: Ethical Challenges in Sharing Social Media-based Research Data." In Internet Research Ethics for the Social Age, edited by Michael Zimmer, and Katharina E. Kinder-Kurlanda, 115-129. New York u.a.: Peter Lang.

different conclusions when it comes to addressing specific challenges

Or in the context of data sharing:

- ? balancing between following principles of good scientific practice and between respecting legal constraints
- ? Perceived ethical obligations *towards the scientific community*
- ? Not sharing data to protect users vs. sharing to include users

Weller, Katrin, and Katharina E. Kinder-Kurlanda. 2017. "To Share or Not to Share?: Ethical Challenges in Sharing Social Media-based Research Data." In Internet Research Ethics for the Social Age, edited by Michael Zimmer, and Katharina E. Kinder-Kurlanda, 115-129. New York u.a.: Peter Lang.

social media data aren‘t „ordinary“ research data

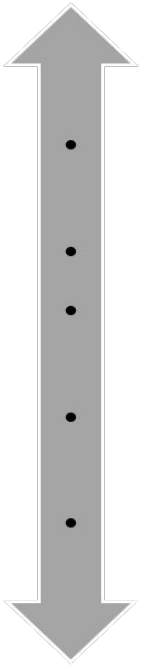
Perceived ethical obligations *towards social media users*

“It’s all public, it doesn’t belong to us, we don’t create the data, we don’t evoke it, I mean it’s natural. I don’t think you have the right to really keep other people from it, no.”

“We share datasets with everybody, actually. We don’t feel we own that.”

how much should I share?

Most reproducibility



What is being shared?

- whole dataset plus additional research information (e.g. scripts)
- whole dataset
- whole dataset, but without direct identifiers (pseudonymization)
- parts of the dataset removed (anonymization)
- changed dataset (e.g. only tweet IDs)

Most privacy

Weller, Katrin, and Katharina E. Kinder-Kurlanda. 2016. "A manifesto for data sharing in social media research." In Proceedings of the 8th ACM Conference on Web Science (WebSci '16), 166-172. New York: ACM.

How can you ensure your results are reproducible?

- Reproducible workflow from the very onset
  - Version control
  - Related: back up stuff regularly
  - Clear data management plan
  - Documentation

- When using social media data
  - Make sure the servers where your data is stored is secure

https://book.the-turing-way.org/reproducible-research/reproducible-research

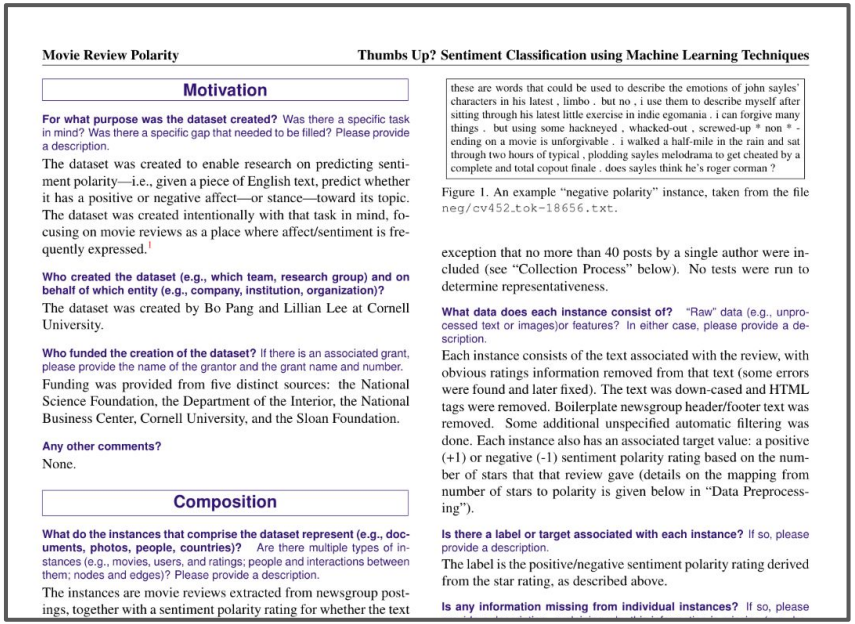
How can you ensure your results are reproducible?

- When using black-box software
  - Save everything! For example, save the labels and not just the final analysis
  - Make a note of when the black-box software was used

- Training and testing models
  - Functions are your friends; make everything as modular as possible
  - Clear definition of inputs and outputs
  - Save the models
  - Save the model outputs!

Getting ready to publish your data and materials

- Make your data as accessible as possible
  - For sensitive cases, ensure privacy and anonymity (more later)
  - Consider uploading to data archives like Zenodo or Harvard Data Archive instead of github
  - Include a license: https://choosealicense.com/licenses/
  - For particularly sensitive cases, consider gated access
  - For newly created data resources (e.g., a newly annotated dataset) add a data sheet or data statement



Movie Review Polarity **Thumbs Up? Sentiment Classification using Machine Learning Techniques**

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity—i.e., given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. The dataset was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are movie reviews extracted from newsgroup postings, together with a sentiment polarity rating for whether the text

these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up * non * - ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example "negative polarity" instance, taken from the file `neg/cv452_tok-18656.txt`.

exception that no more than 40 posts by a single author were included (see "Collection Process" below). No tests were run to determine representativeness.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and later fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) sentiment polarity rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in "Data Preprocessing").

Is there a label or target associated with each instance? If so, please provide a description.

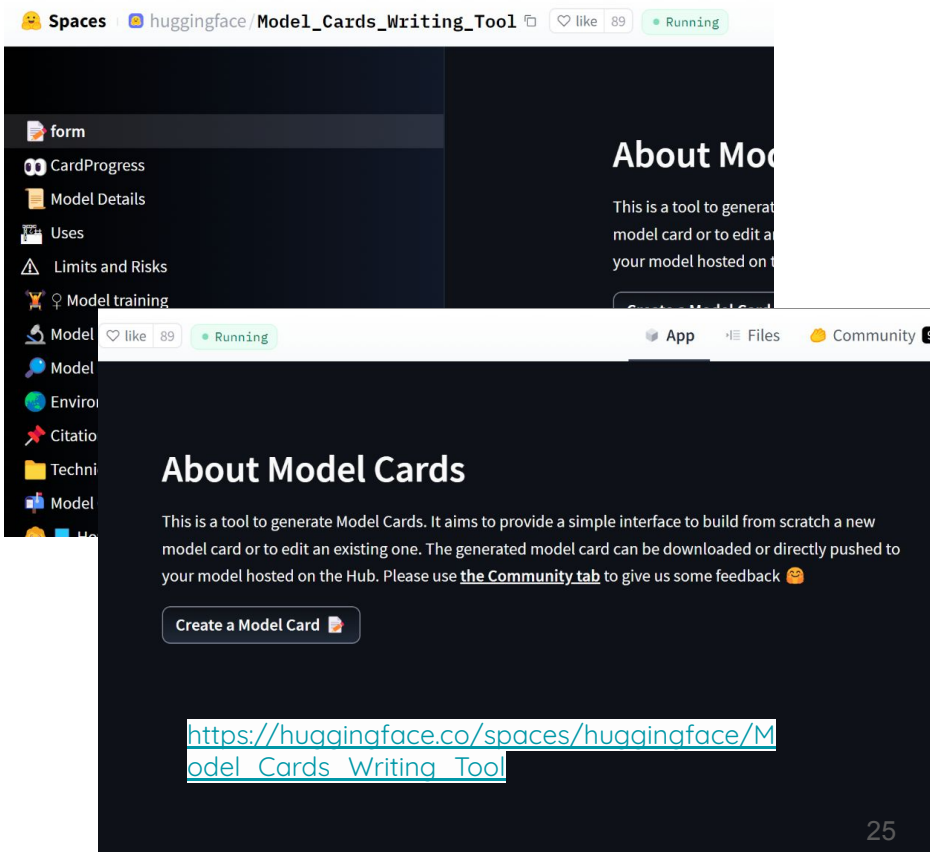
The label is the positive/negative sentiment polarity rating derived from the star rating, as described above.

Is any information missing from individual instances? If so, please

Datasheets for Datasets

Getting ready to publish your data and materials

- Respect data licensing agreements when using secondary dataset sources
  - Point to existing datasets rather than uploading a copy of it
  - Clearly describe how to access the data and how to preprocess it
  - Even better: upload the preprocessing script
- Code Documentation
  - Model cards

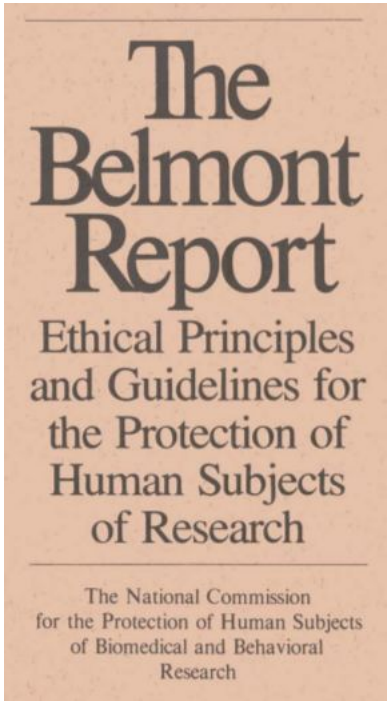


Reproducible Research Resources

- The Turing Way's Guide to Reproducible Research: https://book.the-turing-way.org/reproducible-research/reproducible-research
- Course of effective research workflows: https://www.brendanmichaelprice.com/workflow/ focusing on many details like
  - Readable code
  - Collaborating well
- Schoch et al'2023 "Computational Reproducibility In Computational Social Science"

Ethical Considerations

The Belmont Report



Highly influential report by a national commission in the US

Three overarching principles:

1. **Autonomy/Respect for persons:** respect for individual autonomy, and particularly protection of persons with diminished autonomy

2. **Beneficence and Nonmaleficence:** maximize benefits and minimize harms

3. **Justice:** benefits and burdens should be justly divided (e.g. equally, or according to needs, or efforts, or contributions, or merit)

The Belmont Report

Involving Human Subjects

- Mainly pertinent for experiments
- Tricky when testing interventions
  - E.g., whether people fall for misinformation or not
- Demographics of human subjects
- Consideration when exposing them to harmful content

Handling and Presenting Harmful Text in NLP Research

Hannah Rose Kirk
University of Oxford /
The Alan Turing Institute
United Kingdom
hannah.kirk@oii.ox.ac.uk

Abeba Birhane
Mozilla Foundation /
University College Dublin
Ireland
abeba@mozillafoundation.org

Bertie Vidgen
The Alan Turing Institute
United Kingdom
bvidgen@turing.ac.uk

Leon Derczynski
IT University of Copenhagen
Denmark
ld@itu.dk

Abstract

Text data can pose a risk of harm. However, the risks are not fully understood, and how to handle, present, and discuss harmful text in a safe way remains an unresolved issue in the NLP community. We provide an analytical framework categorising harms on three axes: (1) the harm type (e.g., misinformation, hate speech or racial stereotypes); (2) whether a harm is *sought* as a feature of the research design if explicitly studying harmful content (e.g., training a hate speech classifier), versus *unsought* if harmful content is encountered when working on unrelated problems (e.g., language generation or part-of-speech tagging); and (3) who it affects, from people (mis)represented in the data to those handling the data and those publishing on the data. We provide advice for practitioners, with concrete steps for mitigating

ful content is *sought* when researchers deliberately investigate phenomena such as hate speech, extremism or misinformation. In other cases, researchers are working in seemingly unrelated domains (e.g., language generation, part-of-speech tagging or semantic search) but may still encounter *unsought* harmful content, especially if the data are scraped from internet sources (Luccioni and Viviano, 2021; Dodge et al., 2021; Kreutzer et al., 2022). Third, different groups are harmed by text content during the research process and may suffer immediate, representational or vicarious harms. These groups include *data subjects* (i.e., people represented in the data); *data handlers and researchers*, (i.e., those who collect, annotate or audit the data, and produce research outputs) (Pyevich et al., 2003; Vidgen et al., 2019; Newton, 2020); and *readers and*

utputs).

Handling and Presenting Harmful Text in NLP Research

Dual Use

Need to think about potential misuses and harms of your work

Area	vulnerability	harms
NLP Applications	4.3	crime, oppression, manipulation
Ethics and NLP	4.1	ethics washing, surveillance, plagiarism, oppression
Psycholinguistics	4.0	cyber bullying, oppression
Generation	3.8	crime, cyber bullying, manipulation, oppression, ethics washing
Dialogue Systems	3.8	surveillance, crime
MT and Multilinguality	3.3	surveillance, crime
ML for NLP	3.2	military application, manipulation, oppression
Resources and Evaluation	3.1	ethics washing, surveillance, manipulation
Interpretability	3.0	ethics washing, manipulation
Information Extraction	2.8	surveillance, censorship
IR and Text Mining	2.7	surveillance

Table 1: Average score for vulnerability across ACL areas (with at least three answers) the participants work on and their associated harms.

Thorny Roses: Investigating the Dual Use Dilemma in Natural Language Processing

Lucie-Aimée Kaffee[1], **Arnav Arora**[2], **Zeerak Talat**[3], **Isabelle Augenstein**[2]
[1]Hasso Plattner Institute, Germany, [2]University of Copenhagen, Denmark
[3]Mohamed Bin Zayed University of Artificial Intelligence, United Arab Emirates
lucie-aimee.kaffee@hpi.de, aar@di.ku.dk, z@zeerak.org, augenstein@di.ku.dk

Abstract

Dual use, the intentional, harmful reuse of technology and scientific artefacts, is an ill-defined problem within the context of Natural Language Processing (NLP). As large language models (LLMs) have advanced in their capabilities and become more accessible, the risk of their intentional misuse becomes more prevalent. To prevent such intentional malicious use, it is necessary for NLP researchers and practitioners to understand and mitigate the risks of their research. Hence, we present an NLP-specific definition of dual use informed by researchers and practitioners in the field. Further, we propose a checklist focusing on dual-use in NLP, that can be integrated into existing conference ethics-frameworks. The definition and checklist are created based on a survey of NLP researchers and practitioners.[1]

Introduction

s usage of NLP odels) increases

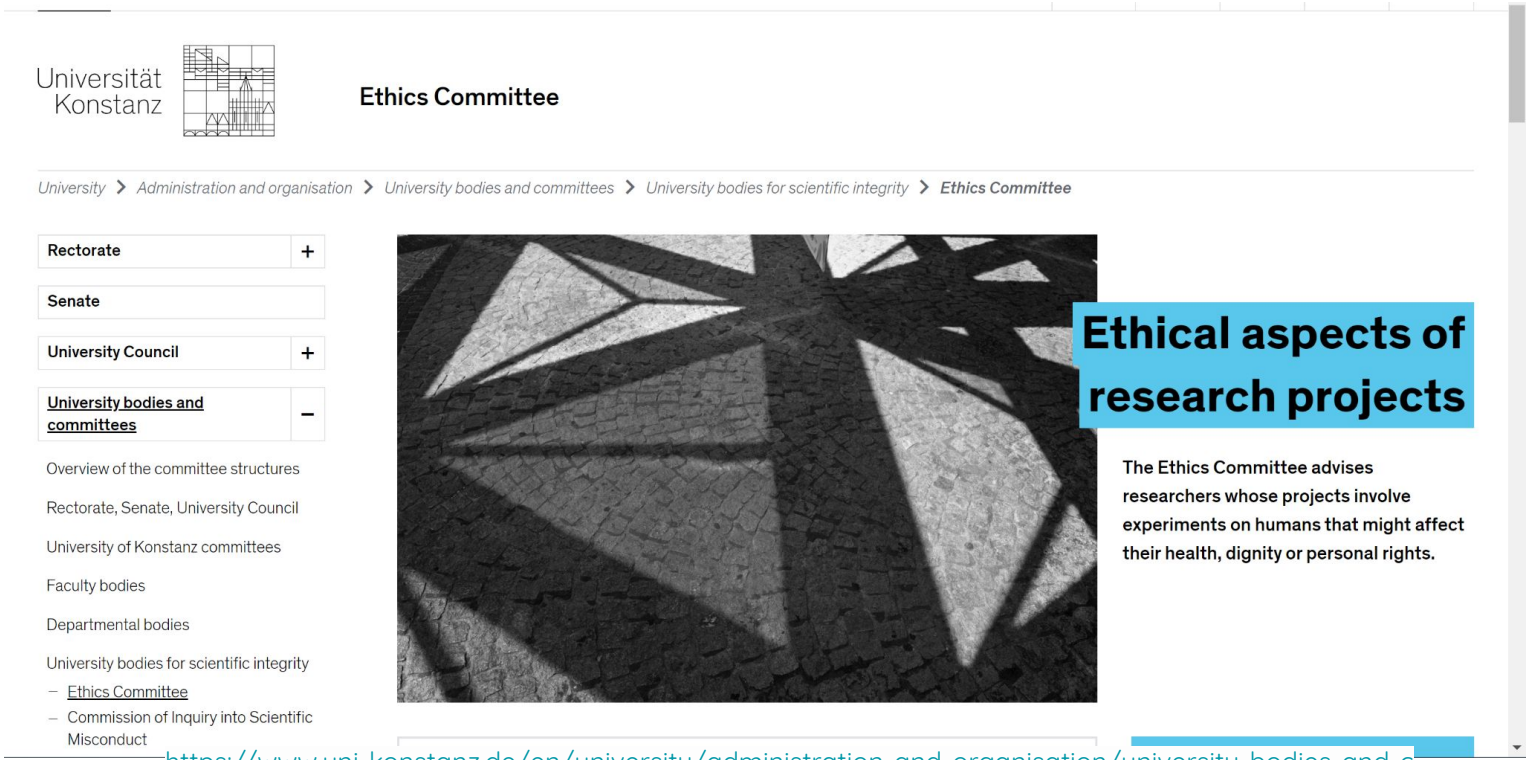
research artefacts. This is reflected in contemporary ethical review processes which emphasise the impacts of research on individual subjects, rather than the wider social impacts of conducted research.

While very few research projects have malicious motivations, some are reused to harm any but particularly marginalised groups of society. This presents a crucial gap in the ethics of artificial intelligence and NLP on malicious reuse, or *dual use*, which has been particularly absent in literature.

Concerns of dual use of Artificial Intelligence (AI) have been discussed by prior work (e.g., Shankar and Zare, 2022; Kania, 2018; Schmid et al., 2022; Urbina et al., 2022; Ratner, 2021; Gamage et al., 2021). However, NLP technologies are rarely included in such considerations. As LLMs are being incorporated into a wide range of consumer-facing products, the dual use considera- ingly critical for research and ance, a online mental health ChatGPT to respond to

Thorny Roses: Investigating the Dual Use Dilemma in Natural Language Processing

Ethics Review and Uni Konstanz



https://www.uni-konstanz.de/en/university/administration-and-organisation/university-bodies-and-committees/university-bodies-for-scientific-integrity/ethics-committee/

Ethics Review and Uni Konstanz

- which methods, technical devices, psychotherapy approaches, etc. will be used,
- how exactly data protection requirements will be fulfilled,
- are any invasive methods used,
- are any medication/s administered, how have the used dosages/application routes been decided,
- which risks for the human test subjects can be expected, and which precautions you will take to reduce risks or avoid/minimise problems that might arise during the experiment, give information in tabular format with the following aspects
    - Potential hazard causes
    - Effects of the hazard / severity of consequences
    - Your assessment of respective hazard probability
    - Measures to minimize risks

CSS/NLP/Machine Learning Resources

Conference guidelines:

- EMNLP [Empirical Methods for NLP]: https://2021.emnlp.org/call-for-papers/ethics-faq
- NAACL [North American Association of Computational Linguistics]:
  - https://2021.naacl.org/ethics/faq/
  - https://2021.naacl.org/ethics/review-questions/
- ACL [Association of Computational Linguistics]: https://www.aclweb.org/portal/content/acl-code-ethics
- Neurips: https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist

Responsible NLP checklist: https://aclrollingreview.org/responsibleNLPresearch/

Reproducibility / Ethics checklist:

- https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf
- https://2020.emnlp.org/blog/2020-05-20-reproducibility
- ICWSM [International Conference of Web and Social Media]: https://www.overleaf.com/latex/templates/aaai-icwsm-2024-paper-checklist/vxbztbhhrbch

Next week(s)

- 24.06: First Project Guidance Session (in-person, same room)
- 27.06: Deadline for submitting project proposal
- 1.07: Project Guidance and Discussion
- 8.07: Project Guidance and Discussion
- 15.07: Project Guidance and Discussion
- 22.07 1:30-3:00 PM [online]: Midway Presentations
- 29.07: Project Guidance and Discussion
- 09.08 10-12 AM [online]: Final Presentations
- 26.08: Final Reports Due