Researching Responsible and Trustworthy Natural Language Processing

Replication, Reproducibility, and Generalisation

Emily Allaway

20 October 2025

School of Informatics
University of Edinburgh
eallaway@ed.ac.uk

Overview

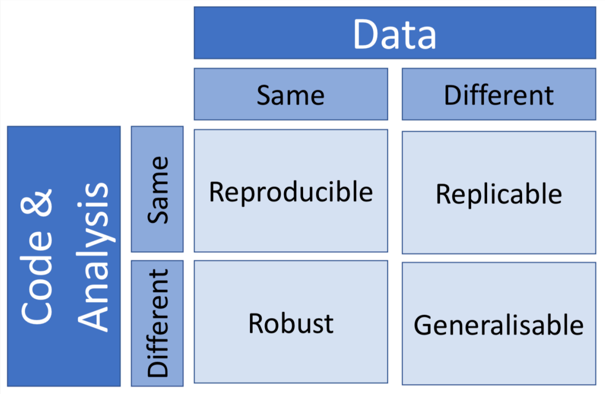
Overview

Replication and Reproducibility

Generalisation

What do we expect of scientific results?

Terms



From: Pineau et al. (2021) and https://github.com/WhitakerLab/ReproducibleResearch

Issue in NLP

Table 1
Distribution of data and code availability in both 2011 and 2016.

	2011: data		2016: data		2011: code		2016: code	
Data / code available	**116**	**75.8%**	**196**	**86.3%**	**48**	**33.1%**	**131**	**59.3%**
- working link in paper	98	64.1%	179	78.9%	27	18.6%	80	36.2%
- link sent	11	7.2%	15	6.6%	17	11.7%	50	22.6%
- repaired link sent	7	4.6%	2	0.9%	4	2.8%	1	0.5%
Data / code unavailable	**37**	**24.2%**	**31**	**13.7%**	**97**	**66.9%**	**90**	**40.7%**
- sharing impossible	19	12.4%	14	6.2%	46	31.7%	42	19.0%
- no reply	17	11.1%	12	5.3%	43	29.7%	32	14.5%
- good intentions	0	0.0%	2	0.9%	5	3.4%	12	5.4%
- link down	1	0.7%	3	1.3%	3	2.0%	4	1.8%

From Wieling et al. (2018)

Efforts to improve reproducibility

The ARR Responsible NLP Research checklist, based on:

✓ **For all reported experimental results**

- ☐ Description of computing infrastructure
- ☐ Average runtime for each approach
- ☐ Details of train/validation/test splits
- ☐ Corresponding validation performance for each reported test result
- ☐ A link to implemented code

✓ **For experiments with hyperparameter search**

- ☐ Bounds for each hyperparameter
- ☐ Hyperparameter configurations for best-performing models
- ☐ Number of hyperparameter search trials
- ☐ The method of choosing hyperparameter values (e.g., uniform sampling, manual tuning, etc.) and the criterion used to select among them (e.g., accuracy)
- ☐ Expected validation performance, as introduced in §3.1, or another measure of the mean and variance as a function of the number of hyperparameter trials.

Dodge et al. (2019)

3. ☐ **Safe use** of data is ensured. *(Check all that apply)*

   3.1. ☐ The data does not include any protected information (e.g. sexual orientation or political views under GDPR), or a specified exception applies. *See Section ___.*

   3.2. ☐ The paper is accompanied by a data statement describing the basic demographic and geographic characteristics of the population that is the source of the language data, and the population that it is intended to represent. *See ___.*

   3.3. ☐ If applicable: the paper describes whether any characteristics of the human subjects were self-reported (preferably) or inferred (in what way), justifying the methodology and choice of description categories. *See Section ___.*

   3.4. ☐ The paper discusses the harms that may ensue from the limitations of the data collection methodology, especially concerning marginalized/vulnerable populations, and specifies the scope within which the data can be used safely. *See Section ___.*

   3.5. ☐ If any personal data is used: the paper specifies the standards applied for its storage and processing, and any anonymization efforts. *See Section ___.*

   3.6. ☐ If the individual speakers remain identifiable via search: the paper discusses possible harms from misuse of this data, and their mitigation. *See Section ___.*

Rogers et al. (2021)

Limitations & Risks

Limitations (week 3), Risks (now)

- Examples
  - potential malicious or unintended harmful effects & uses
  - environmental impact
  - fairness
  - privacy
  - security
- Consider particularly
  - Dual use
  - Variety of stakeholders impacted
  - Relevant mitigation strategies

From A2: https://aclrollingreview.org/responsibleNLPresearch/

Exercise - why are items on the checklist

For each item on the checklist, discuss

- Why is the information useful?
- What area of reproducible research does it contribute to?

B - Scientific Artifacts

B1. Cite creators?

B2. Licenses and terms of use?

B3. Intended use?

B4. Uniquely identifying information or offensive content?

B5. Documentation of artifacts?

B6. Statistics of artifacts?

C - Compute Experiments

C1. Parameters and compute information?

C2. Hyperparameters?

C3. Descriptive statistics of results?

C4. Packages and settings?

D - Human Annotations & Participants

D1. Instructions to participants?

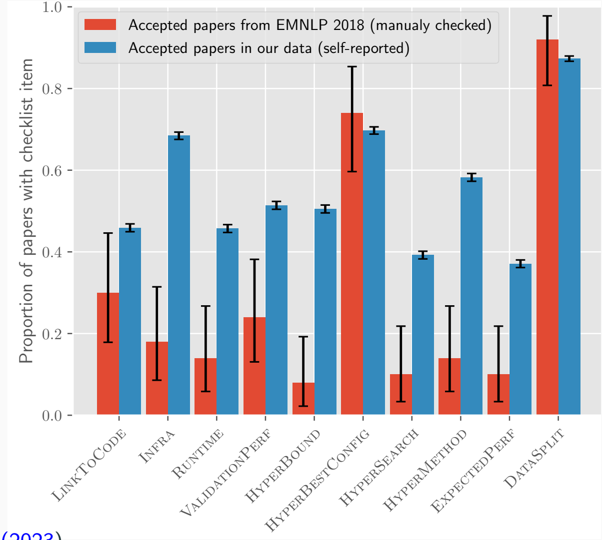
D2. Recruitment information?

D3. Consent?

D4. Ethics approval?

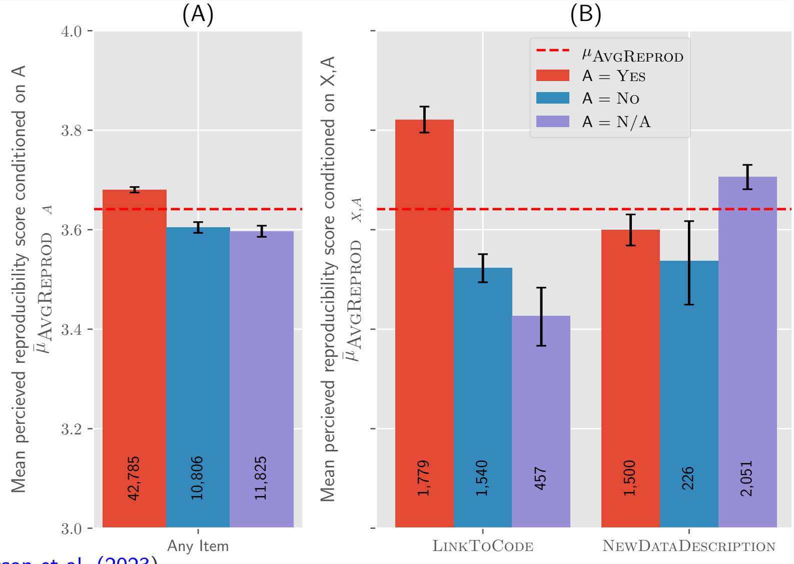
D5. Annotator population characteristics?

Is the checklist effective?



From Magnusson et al. (2023)

Perceived usefulness of the checklist



From: Magnusson et al. (2023)

Potentially bad faith responses

Response	Conference	Submissions	ACCEPT
YES	EMNLP 2020	134 (4.5%)	-9.9%
	EMNLP 2021	238 (7.3%)	-6.7%
	NAACL 2021	79 (6.4%)	-3.3%
	ACL 2021	213 (7.3%)	-8.2%
NO	EMNLP 2020	1 (0.0%)	-39.7%
	EMNLP 2021	0 (0.0%)	-
	NAACL 2021	0 (0.0%)	-
	ACL 2021	0 (0.0%)	-

From: Magnusson et al. (2023)

Work still difficult to reproduce

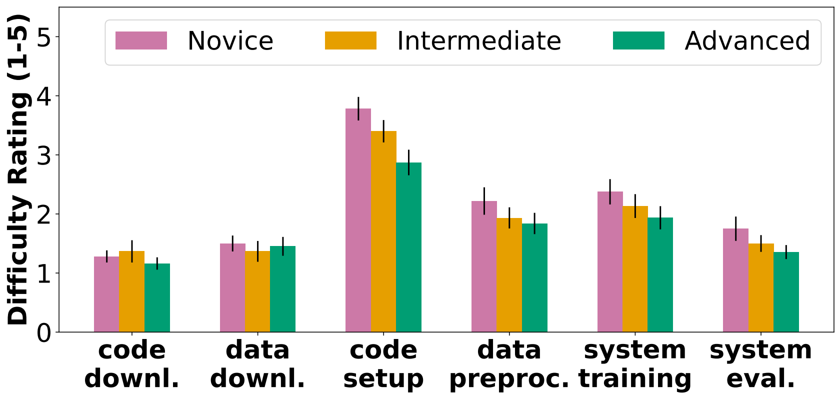


Figure 2: Mean reproducibility difficulty rating (1-5, 5 being most difficult) for each step of experiments

From Storks et al. (2023)

Code is a major blockers to reproduction

Reproducibility Blocker	**Frequency**
Insufficient Code Dependency Specification	38
Difficult-to-Access External Resources	27
Unclear Code Usage Documentation	17
Pre-Existing Bugs in Code	16
Difficult-to-Read Code	11
Other	30

From Storks et al. (2023)

Challenges doing reproducible research

- *Takes time!*

- Held to higher standards
- Openness to mistakes
- Publication bias towards novel findings
- IP/confidentiality issues

From Data et al. (2017)

Exercise - recommendations

Magnusson et al. (2023) and Storks et al. (2023) both make recommendations for the checklist:

1. Checklist responses made public
2. Extra time allowed for submitting the checklist & accompanying items

Suggested ACLRC Addition	**Frequency**
Standards for Documentation Clarity	22
Full Specification of Code Dependencies	18
Demonstration of Code Usage	9
Provision of Support for Issues	8
Standards for Code Clarity	5

Should these be implemented?

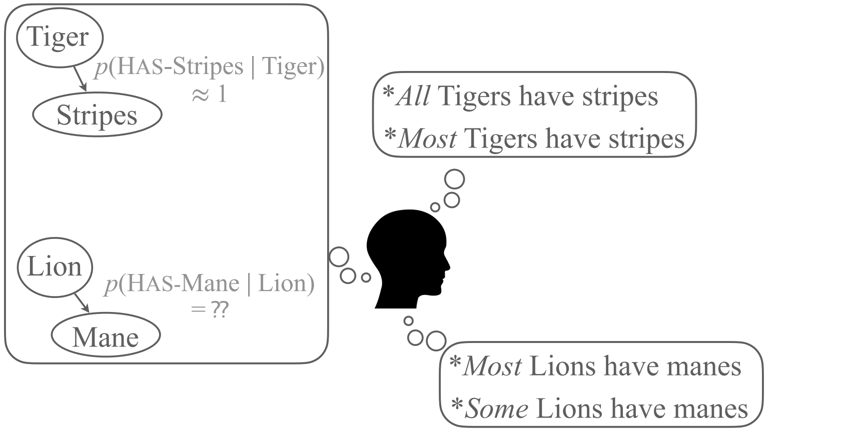
Further reading

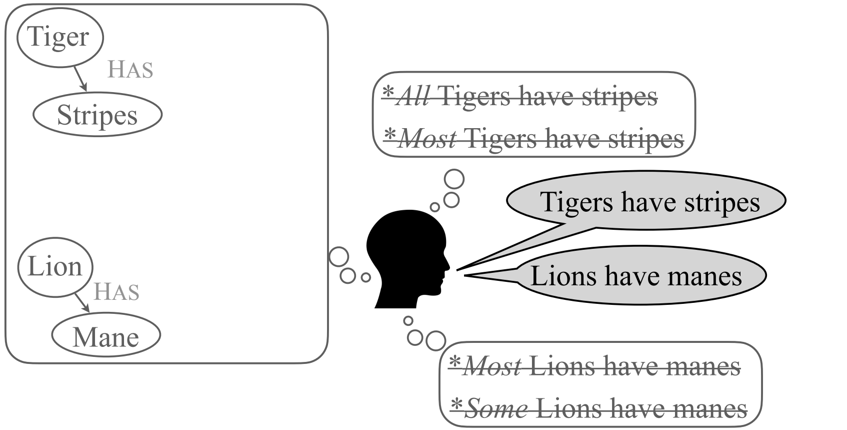
- Reproducibility, correctness, and buildability: The three principles for ethical public dissemination of computer science and engineering research (Rozier and Rozier, 2014)
- Three Dimensions of Reproducibility in Natural Language Processing (Cohen et al., 2018)
- A practical taxonomy of reproducibility for machine learning research (Tatman et al., 2018)

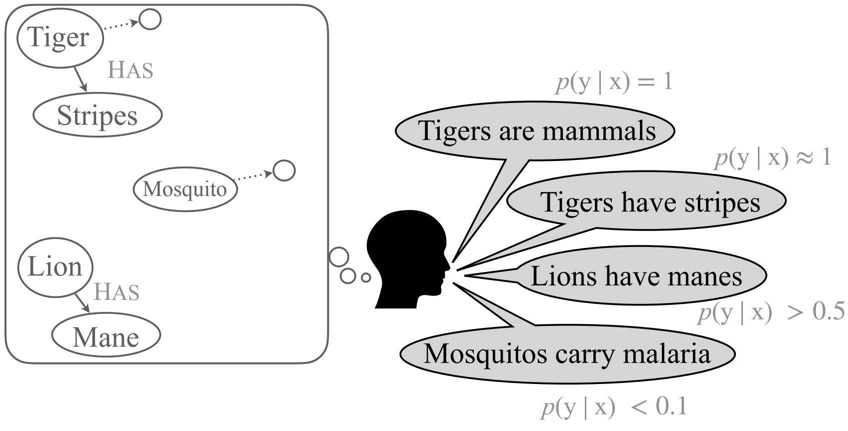
Overview

Replication and Reproducibility

Generalisation


Tiger
p(HAS-Stripes | Tiger) ≈ 1
Stripes
Lion
p(HAS-Mane | Lion) = ??
Mane
**All* Tigers have stripes
**Most* Tigers have stripes
**Most* Lions have manes
**Some* Lions have manes

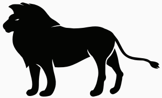

Tiger
HAS
Stripes
Lion
HAS
Mane
~~*All Tigers have stripes~~
~~*Most Tigers have stripes~~
Tigers have stripes
Lions have manes
~~*Most Lions have manes~~
~~*Some Lions have manes~~


Tiger
HAS
Stripes
Mosquito
Lion
HAS
Mane
p(y | x) = 1
Tigers are mammals
p(y | x) ≈ 1
Tigers have stripes
Lions have manes
p(y | x) > 0.5
Mosquitos carry malaria
p(y | x) < 0.1

Generic overgeneralization

{ **All** / Most / Some } **Lions have manes**

This lion has a mane



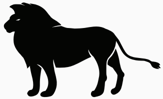
Leslie et al. (2011)

Generic overgeneralization

All
Most
Some

Lions have manes

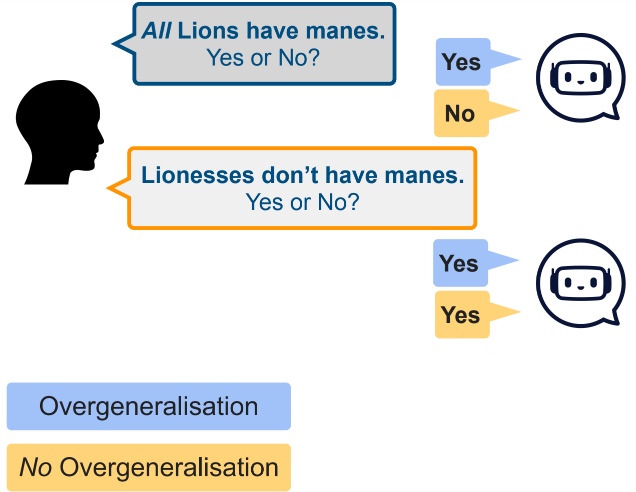
This lion has a mane



Exception

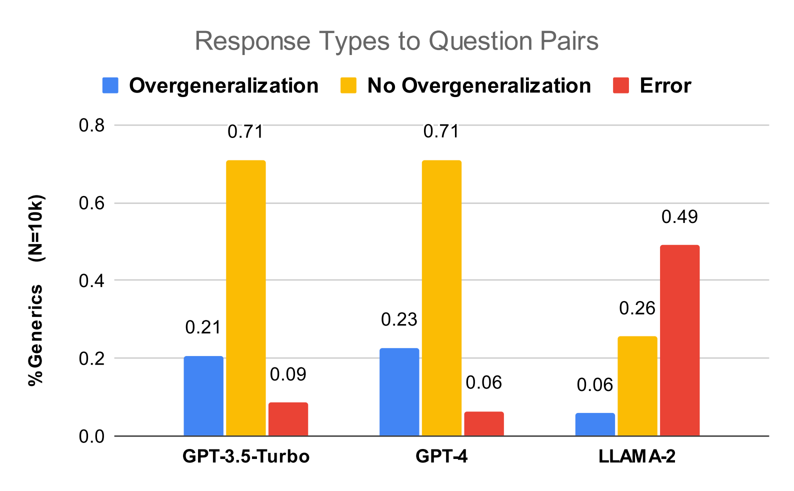
Lionesses don't have manes

Probing for overgeneralization



From Allaway et al. (2024).

Consistent overgeneralization in LLMs



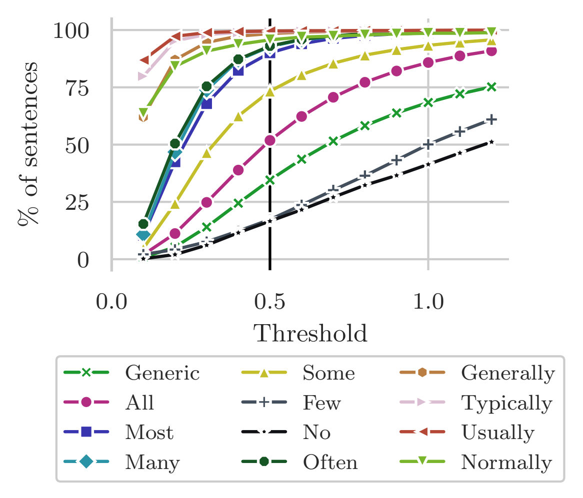
From Allaway et al. (2024).

Claims as generalisations

Many ways to express generalisations

- Quantifiers
  - *Most* models perform well on the MT tasks
  - We find that *generally* our method produces fewer hallucinations than the baselines.
- Scoping
  - *When using beam search*, models produce more fluent outputs.
- Generics
  - Our model outperforms or performs comparably to the baselines.

Generic generalisations



Generics are not like quantified statements (Calderón et al., 2025).

Hasty generalisations

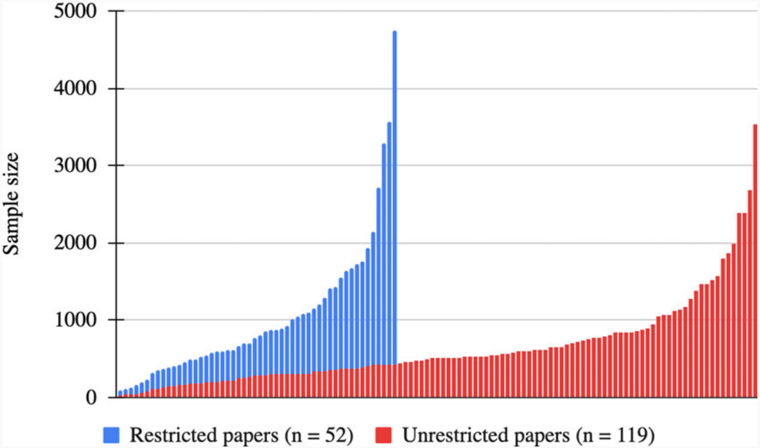


Figure 2. Full distribution of sample sizes within each group of x-phi articles.

From Peters and Lemeire (2024)

Hasty generalisations

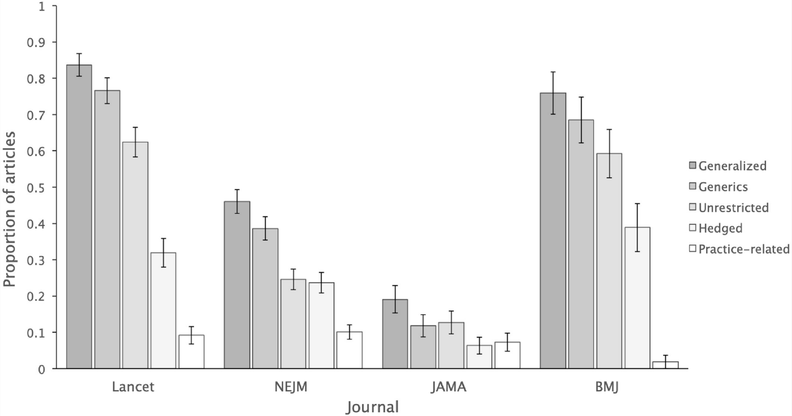


Fig 2. The proportion of articles with generalized, generic, unrestricted, hedged, or practice-related result claims (derived by dividing the number of articles with these claims with the total number of articles of each journal). Error bars indicate standard error for the variability in proportion estimates.

From Peters et al. (2024)

Exercise - evaluating claim soundness

1. Rate each claim on a scale of 1-5 (least to most) in terms of *robustness*, *replicability*, and *generalisability*.
2. Find and identify where the evidence for each claim is located.
    - Highlight table or figure labels (e.g., Figure 1) or sentences.
    - Use a separate color for the evidence corresponding to each claim.
3. Consider whether there is any *non-supporting* evidence.
    - If so, highlight or mark this evidence (e.g., cells in a table).
4. Rerate the claims, taking into account the evidence you have seen.
    - If your rating is different from initially, please describe why briefly.
5. Discuss the claims and your ratings with your partner and come to a consensus for each aspect.

Exercise - evaluating claim soundness

Observations?

- How did evidence impact your perception?
- Did non-supporting evidence influence your perception?
- Did either of these things surprise you?

References i

Emily Allaway, Chandra Bhagavatula, Jena D. Hwang, Kathleen McKeown, and Sarah-Jane Leslie. 2024. Exceptions, instantiations, and overgeneralization: Insights into how language models process generics. *Computational Linguistics*, 50:1211–1275.

Gustavo Cilleruelo Calderón, Mahrad Almotahari, Emily Allaway, Barry Haddow, and Alexandra Birch. 2025. A new path from language models to semantic theory. In *In Submission to ARR*.

K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E. Hunter. 2018. Three dimensions of reproducibility in natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Scientific Data, Kirstie Whitaker, and The Ludic Group LLP. 2017. Kirstie Whitaker - Better Science through Better Data 2017 keynote scribe images.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Sarah-Jane Leslie, Sangeet Khemlani, and Sam Glucksberg. 2011. Do all ducks lay eggs? The generic overgeneralization effect. *Journal of Memory and Language*, 65(1):15–31.

Ian Magnusson, Noah A. Smith, and Jesse Dodge. 2023. Reproducibility in NLP: What have we learned from the checklist? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12789–12811, Toronto, Canada. Association for Computational Linguistics.

Uwe Peters and Olivier Lemeire. 2024. Hasty generalizations are pervasive in experimental philosophy: A systematic analysis. *Philosophy of Science*, 91(3):661–681.

Uwe Peters, Henrik Røed Sherling, and Benjamin Chin-Yee. 2024. Hasty generalizations and generics in medical research: A systematic review. *Plos one*, 19(7):e0306749.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research*, 22(164):1–20.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'just what do you think you're doing, dave?' a checklist for responsible data use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kristin Yvonne Rozier and Eric WD Rozier. 2014. Reproducibility, correctness, and buildability: The three principles for ethical public dissemination of computer science and engineering research. In *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering*, pages 1–13. IEEE.

Shane Storks, Keunwoo Yu, Ziqiao Ma, and Joyce Chai. 2023. NLP reproducibility for all: Understanding experiences of beginners. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10199–10219, Toronto, Canada. Association for Computational Linguistics.

Rachael Tatman, Jake VanderPlas, and Sohier Dane. 2018. A practical taxonomy of reproducibility for machine learning research.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics*, 44(4):641–649.