Researching Responsible and Trustworthy Natural Language Processing

Replication, Reproducibility, and Generalisation

Emily Allaway

19 October 2026

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

Overview

Replication and Reproducibility

Generalisation

Overview

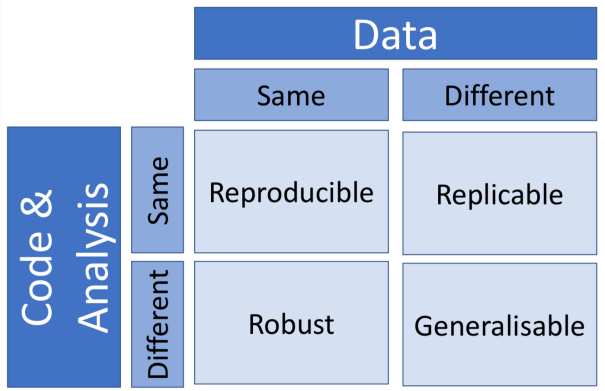
Replication and Reproducibility

Generalisation

What do we expect of scientific results?

Support the claims
Properly scoped to where you tested
If reproducibility is relevant, there should be enough information to reproduce
Verifiable — enough information about methods and also data provenance
Transparent about limitations
Metrics should be reliable, make sense for the given task
Construct validity — are you measuring what you say you're measuring

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 2), 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?
- What type of rigor does this relate 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?

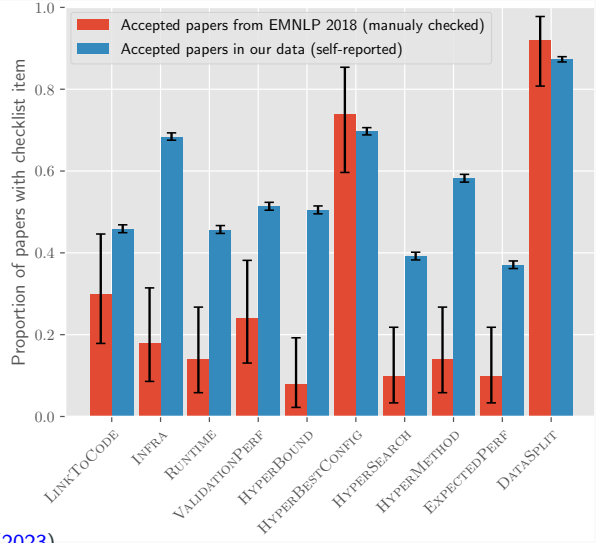
D2. Recruitment information?

D3. Consent?

D4. Ethics approval?

D5. Annotator population characteristics?

Is the checklist effective?



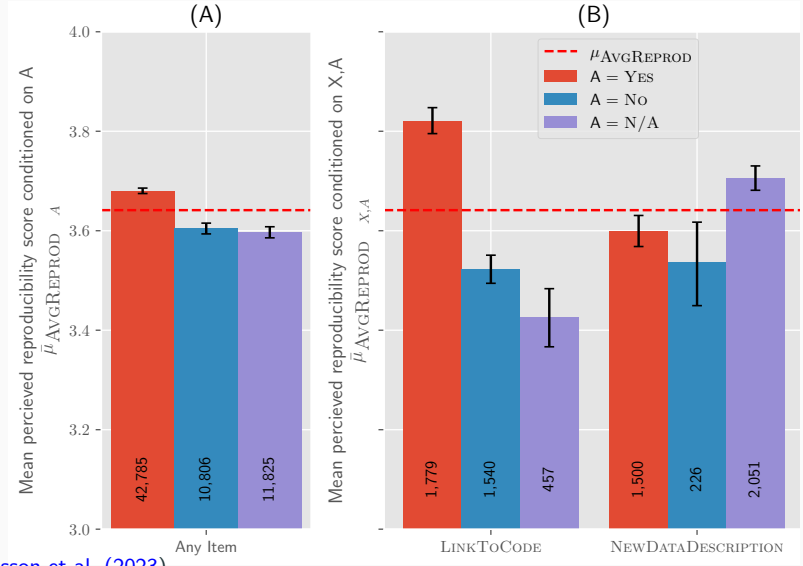
From Magnusson et al. (2023)

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)

Perceived reproducibility from the checklist



From: Magnusson et al. (2023)

Work still difficult to reproduce

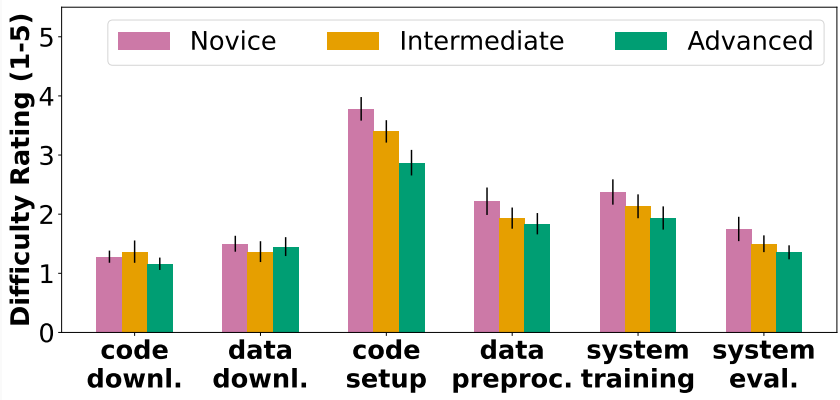


Figure 2: Mean reproducibility difficulty rating (1-5 5 being most difficult) for each step of experiments

From Storks et al. (2023)

Code is a major blocker 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?

Overview

Replication and Reproducibility

Generalisation

What makes generalisable research?

Something that doesn't rely on awkward/unrealistic assumptions
Expand it easily to different areas
Qualitative research: theoretical constructs are transferrable across contexts
Statistically: model is reliable when applied to unknown data/texts

Roles of reasoning in generalisation

- Induction plays a role in forming & empirically corroborating hypotheses
  - Identify similarities within a group $\rightarrow$ generalisation
- Deduction plays a role in hypothesis formation
  - Identify a particular instance using accepted facts
  - How we take & build on existing research

From Bunge (1960) and Rothchild (2006).

Generalisation through reasoning: deduction

Syllogisms and the application of rules (deduction)

A	**B**	**C**
Major premise (or rule)	*Minor premise (or case)*	*Conclusion (or result)*
All metals are conductors of electricity	Copper is a conductor of electricity	Copper is a metal

From Gobo and Marcheselli (2023).

Generalisation through reasoning: deduction

Syllogisms and the application of rules (deduction)

A	B	C
Major premise (or rule)	*Minor premise (or case)*	*Conclusion (or result)*
All metals are conductors of electricity	Water is a conductor of electricity	Water is a metal
Adults smoke	I smoke	I am an adult
The human being eats salad	The goat eats salad	The goat is a human being

From Gobo and Marcheselli (2023).

Generalisation and deduction

A real (simplified) example

- LLMs are good at generating synthetic data
- GPT-4o-mini is an LLM
- $\Rightarrow$ GPT-4o-mini is good at generating synthetic data.

Generalisation through reasoning: induction

Generalisation on the basis of case examination (Bunge, 1960).

- Scientists generalise: study population → broader population

From Peters et al. (2022).

Generalisation through reasoning: induction

Generalisation on the basis of case examination (Bunge, 1960).

- Scientists generalise: study population → broader population

- Scoping & generalisation bias
  "Healthy nutrition increases people's life expectancy
  if they do not have an intolerance with certain foods important to a healthy diet,
  if they do not face other biological or mental constraints (e.g., orthorexia), *if*
  they do not eat too much healthy food ..."

From Peters et al. (2022).

Generalisation through reasoning: induction

Generalisation on the basis of case examination (Bunge, 1960).

- Scientists generalise: study population → broader population

- Scoping & generalisation bias
  "**Healthy nutrition increases people's life expectancy**
  if they do not have an intolerance with certain foods important to a healthy diet, *if* they do not face other biological or mental constraints (e.g., orthorexia), *if* they do not eat too much healthy food ..."

From Peters et al. (2022).

Generalisation and sample size

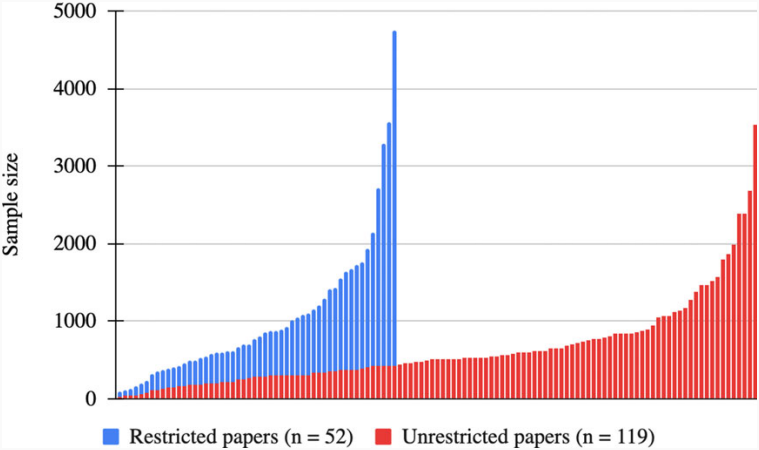
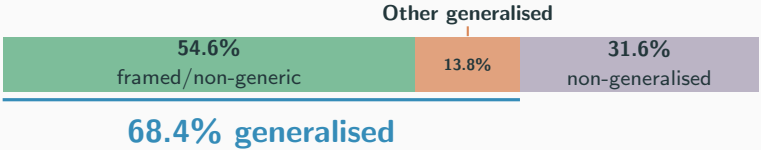


Figure 2. Full distribution of sample sizes within each group of x-phi articles.

From Peters and Lemeire (2024)

Generalisation in NLP papers

In NLP papers from 4 conferences, considering sentences about paper *outcomes*



From Diao et al. (2026).

Generalisation as a common issue in NLP

R2. Inappropriate scope of the claims	The authors evaluate a sample that does not represent the population about which the claim is made. E.g., a few QA benchmarks !="reasoning" or "understanding", LLMs of a certain size != LLMs.
R4. Misleading or inappropriate framing, overclaiming	E.g., concluding from benchmark evaluation that LLMs generally 'understand' language, without validating that construct.
R5. Inappropriate or missing statistical significance assessment	Ideally, at least the main experimental results should be accompanied by appropriate information about their statistical significance (error bars, confidence intervals, statistical significance tests), details about how this was computed, and discussion of factors of variability and any assumptions. Effect size estimation is also very welcome.

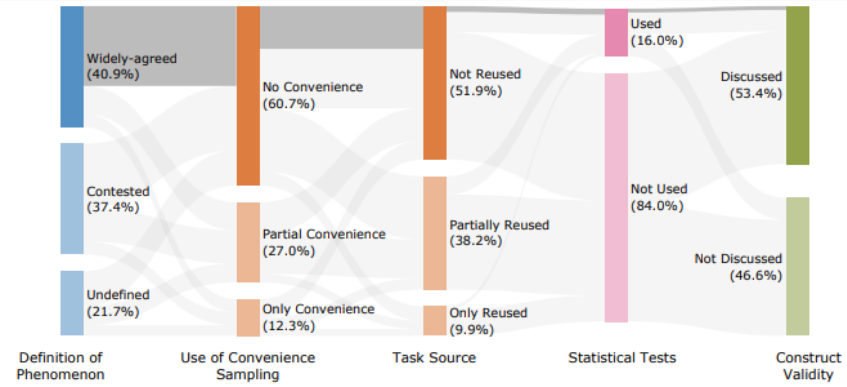
From ARR Reviewer Guidelines

What elements of rigor related to generalisation?

transparency, as much as possible
don't overhype your own claims
situate what you are doing within what already exists — can help with hype maybe
provide a robust theoretical explanation of why your results are generalisable and when they are/arent

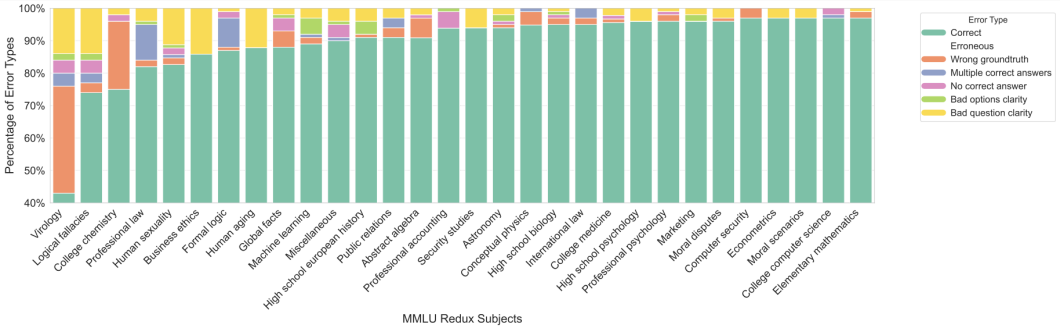
Generalisation and rigor – construct validity

Construct validity: whether a test actually measures what it says it does



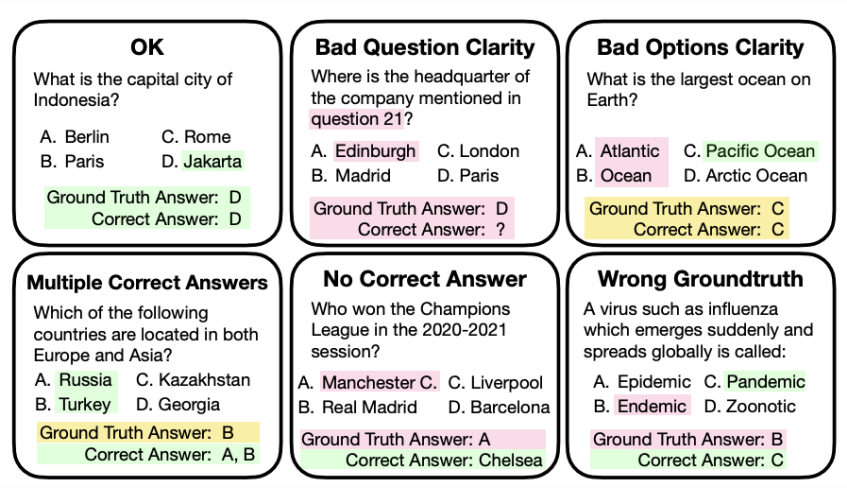
Results from analysis of 445 benchmark papers (Bean et al.).

Errors in MMLU can be problematic



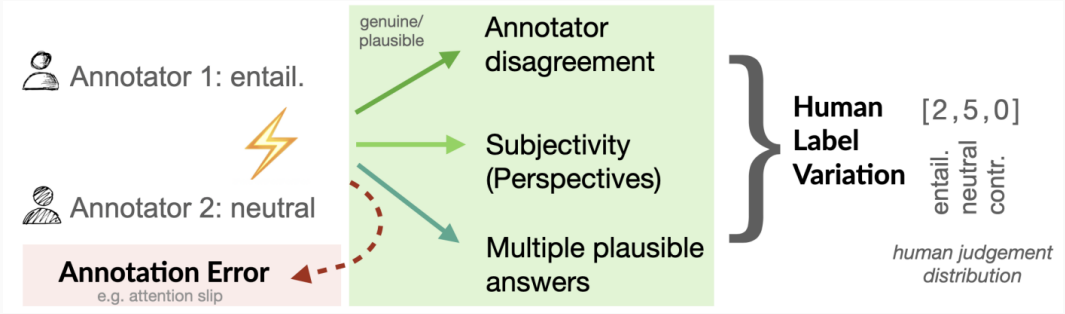
From Gema et al. (2025).

Errors are not always errors



From Gema et al. (2025).

Errors or human variation?



From Plank (2022).

So how do we do generalisable research then?

really well defined research questions
place greater value on metaresearch (literature reviews, survey papers)
shared definitions and normative practices
selecting subject where you know it will work for the specific but will likely offer generality
publishing raw data not just aggregated data
if you have a lot of disagreement, consider reevaluating your task or maybe your research method
interdisciplinarity can bring other perspectives on the same problem
do a lot of error analysis and looking at your data and results
“don’t generalise” —> be careful about the population
interrogating your assumptions and latent biases
don’t use language as a proxy for culture, language populations and geography are not the same thing
theres a tendency to see language as truth rather than language as labels

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)

References i

Andrew M Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, et al. Measuring what matters: Construct validity in large language model benchmarks, 2025. *URL https://arxiv. org/abs/2511.04703*.

Mario Bunge. 1960. The place of induction in science. *Philosophy of Science*, 27(3):262–270.

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.

Chenxin Diao, Nataliya Stepanova, and Emily Allaway. 2026. How broad is that claim? mapping generalization in nlp research. *Accepted to EMNLP*.

References ii

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.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. 2025. Are we done with MMLU? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096, Albuquerque, New Mexico. Association for Computational Linguistics.

Giampietro Gobo and Valentina Marcheselli. 2023. *Science, Technology and Society*. Springer.

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, Alexander Krauss, and Oliver Braganza. 2022. Generalization bias in science. *Cognitive science*, 46(9):e13188.

References iii

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

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.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 10671–10682.

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.

Irving Rothchild. 2006. Induction, deduction, and the scientific method. In *Society for the Study of Reproduction*, pages 1–11. Case Western Reserve University ^ eOhio Ohio.

References iv

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.