EACL 2021 Tutorial: Reviewing Natural Language Processing Research

Biases and Heuristics in Peer Review

Anna Rogers, University of Copenhagen

We want to do peer review well, but...

Peer review is a very difficult task!

Rogers, A., and Augenstein, I. (2020). What Can We Do to Improve Peer Review in NLP? In Findings of EMNLP, (Online: ACL), pp. 1256–1262.

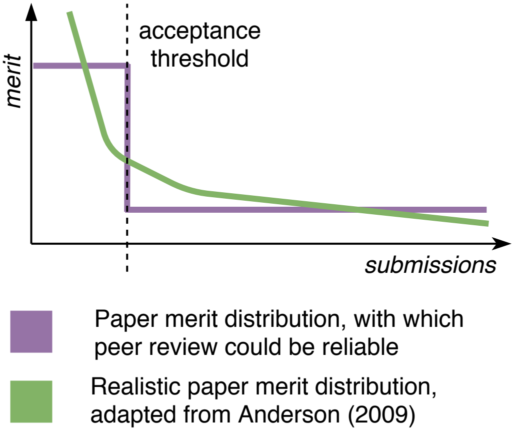
Goals of peer review

- quality control
- selecting impactful, important publications

What can we realistically expect from peer review?

- quality control ✘
- selecting impactful, important publications ✘

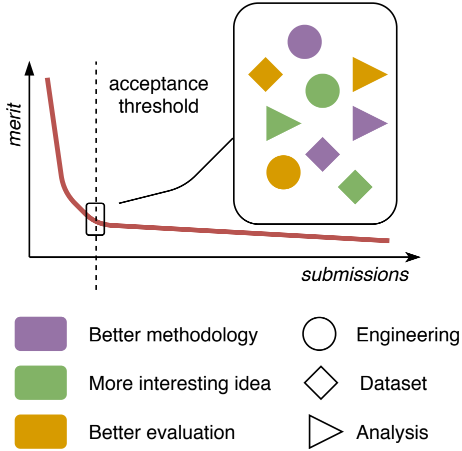
Why peer review is a difficult task



🟪 Paper merit distribution, with which peer review could be reliable

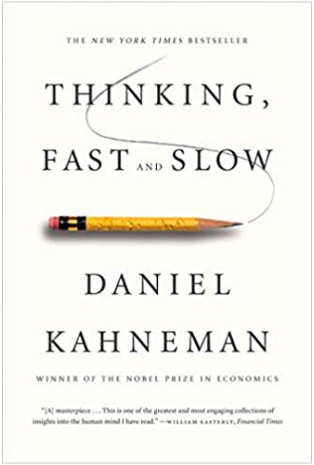
🟩 Realistic paper merit distribution, adapted from Anderson (2009)

Why peer review is a difficult task



How do people reason in high-uncertainty situations?

Biases to the rescue!



THE NEW YORK TIMES BESTSELLER

THINKING, FAST AND SLOW

DANIEL KAHNEMAN

WINNER OF THE NOBEL PRIZE IN ECONOMICS

"[A] masterpiece . . . This is one of the greatest and most engaging collections of insights into the human mind I have read." —WILLIAM EASTERLY, *Financial Times*

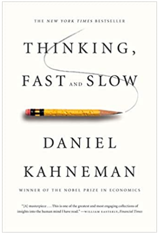
Implicit bias

”Bias that results from the tendency to process information based on unconscious associations and feelings, even when these are contrary to one’s conscious or declared beliefs”

Dictionary.com

Substitute questions

"This is the essence of intuitive heuristics: when faced with a difficult question, we often answer an easier one instead, usually without noticing the substitution."

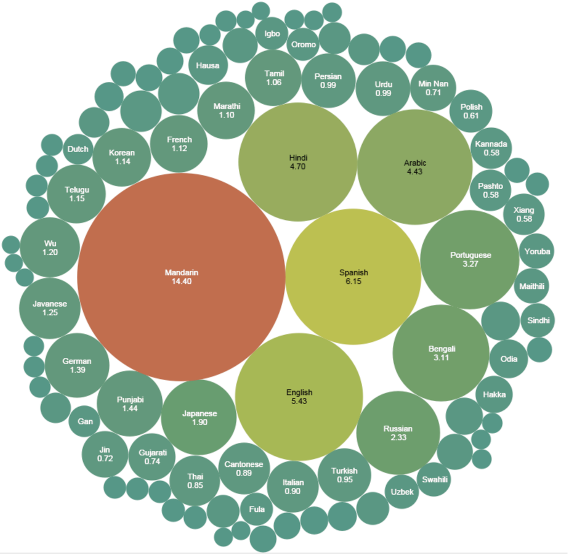

THE NEW YORK TIMES BESTSELLER
THINKING,
FAST AND SLOW
DANIEL
KAHNEMAN
WINNER OF THE NOBEL PRIZE IN ECONOMICS
"[A] masterpiece . . . This is one of the greatest and most engaging collections of insights into the human mind I have read." —WILLIAM EASTERLY, Financial Times

Language heuristic: definition

Difficult question:
Is this paper good?

→

Easy question:
Is it well-written?



Bubble chart of languages by proportion of native speakers worldwide (2007 estimates). Jroehl, CC BY-SA 4.0, via Wikimedia Commons

Language heuristic: issues

- non-native speakers of English systematically at disadvantage;
- papers with weaker content may be rated higher*than papers with weaker language!

As long as the paper as readable, make the effort to look at the content rather than language.

*Church, K. W. 2020. Emerging Trends: Reviewing the Reviewers (Again). *Natural Language Engineering*. https://www.cambridge.org/core/journals/natural-language-engineering/article/emerging-trends-reviewing-the-reviewers-again/10CDC1D71E1AEB21456CFBDA187CBCB6.

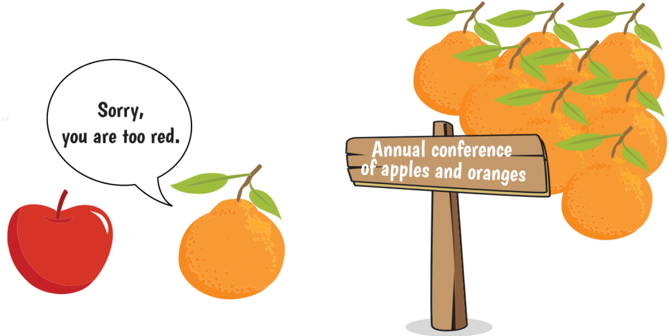
“Preferred methodology” heuristic: definition

Difficult question:
Is this paper good?

➔

Easy question:
Is this paper doing things the way I would do them?

Interdisciplinary field?



“Preferred methodology” heuristic: issues

- NLP is an inherently interdisciplinary field, *not* linguistics and *not* machine learning;
- if experimentalists dismiss theoretical papers, position papers, surveys, and the people working on the latter dismiss experimental work, we won’t get anywhere.

If the paper is in the scope of CFP, but you a priori disagree with the methodology or do not see this type of contribution as “research”, reviewing will be a waste of your and the authors’ time. Ask to reassign it.

Confirmation bias: definition

Difficult question:
Is this paper good?

➔

Easy question:
Does the result confirm my view of the issue?

Confirmation bias: example

- Study* of medical researchers who had previously reported results either for or against the clinical effectiveness of TENS therapy method;
- asked to review a fictitious paper reporting a positive result on this therapy, and deliberately including both strong and weak methodology points;
- higher evaluation by researchers who had already believed this therapy to work!

*Ernst, E., Resch, K. & Uher, E. 1992. Reviewer Bias. *Annals of Internal Medicine.* `https://www.acpjournals.org/doi/abs/10.7326/0003-4819-116-11-958_2`.

Confirmation bias: issues

- ignoring useful information;
- slowing down progress;
- backfiring effect: faced with disconfirming evidence, humans may strengthen their beliefs rather than adjust them.

Imagine your own paper being reviewed by your opponents on this issue, and give it the fair chance that you'd like to have yourself. Is the methodology solid? Do these results help to resolve the issue (either way)?

State-of-the-art (SOTA) heuristic: definition

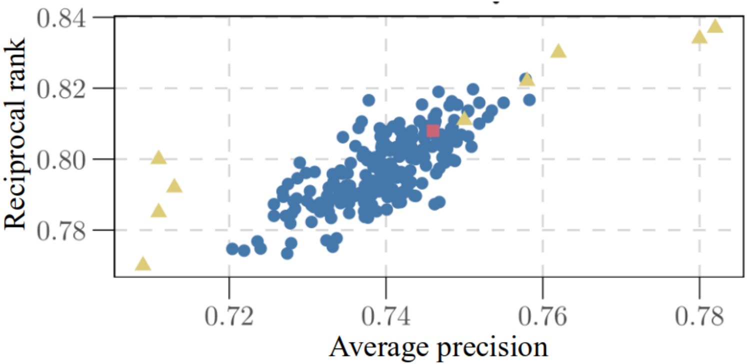
Difficult question:
Is this paper good?

→

Easy question:
Are the results SOTA?

SOTA heuristic: are the improvements really significant?

Variation between random seed runs for sample models (indicated by shapes) on TrecQA dataset*



*Crane, M. 2018. Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results. *Transactions of the Association for Computational Linguistics.* `https://aclweb.org/anthology/papers/Q/Q18/Q18-1018/`.

SOTA heuristic: issues

- the competition is no longer feasible for small labs*;
- puts everybody in a hamster wheel, with results already outdated by the time the paper is reviewed;
- encourages unreproducible, cherry-picked, brittle results;
- discourages improvements in other areas†;
- disadvantages data and theoretical work.

SOTA results are neither necessary nor sufficient for a valuable research contribution.

*Rogers, A. 2019. How the Transformers Broke NLP Leaderboards. *Hacking semantics.* https://hackingsemantics.xyz/2019/leaderboards/.

†Rogers, A. 2020. Peer Review in NLP: Reject-If-Not-SOTA. *Hacking semantics.* https://hackingsemantics.xyz/2020/reviewing-models/; Ethayarajh, K. & Jurafsky, D. 2020. Utility Is in the Eye of the User: A Critique of NLP Leaderboards. *arXiv:2009.13888 [cs].* http://arxiv.org/abs/2009.13888.

Bias towards positive results: definition

Difficult question:
Is this paper good?

→

Easy question:
Is this paper providing a positive result?

Bias towards positive results: example

75 psychologists reviewed* the same fictitious study with varied results (positive, negative and mixed):

- **Results section not shown:** “Very good. Well done. If the Results and Discussion... are as well written... I definitely recommend publication.”
- **Positive results:** ”An excellent paper.., it definitely merits publishing. I find little to criticize. The topic is excellent and very relevant, the design is quite adequate, and the style is very good.”
- **Negative results:** “There are so many problems with this paper, it is difficult to decide where to begin.”

*Mahoney, M. J. 1977. Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System. *Cognitive therapy and research.*

Bias towards positive results: NLP flavor

Author: This doesn't works.
Reviewer: Hmm, is there a bug?

Author: This works.
Reviewer: Great. ~~Hmm, did you get lucky?~~

Both cases require the same judgement about whether you believe that the implementation is correct. As a reviewer, you are not expected to reproduce the paper, but conferences now often include reproducibility checklist.

Bias towards positive results: issues

- in NLP: further conflating performance with advancement of the state of knowledge (see SOTA heuristic);
- ignoring useful information;
- slowing down progress.

When reading the paper, **first look at the methodology and design of the study and decide whether it is sound and will yield a useful piece of information. Then read the results.** Whether they are negative or positive, the question is how useful it'd be for them to be widely known.

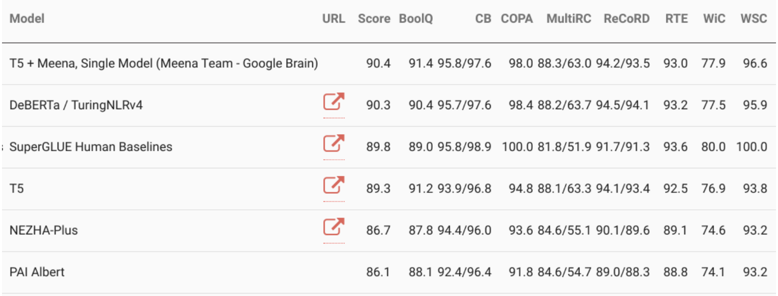
“Resource paper” heuristic: definition

Difficult question:
Is this paper good?

→

Easy question:
Is it an engineering paper?

SuperGLUE is “solved”, language is not!



Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC
T5 + Meena, Single Model (Meena Team - Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6
DeBERTa / TuringNLRv4	↗	90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9
SuperGLUE Human Baselines	↗	89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0
T5	↗	89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8
NEZHA-Plus	↗	86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2
PAI Albert		86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2

“Resource paper” heuristic: issues

- We desperately need non-game-able datasets!
- That requires breakthroughs in annotation and data methodology...
- which requires publication incentives...
- which won’t happen, if the centerpiece of a paper has to be a model, and data-focused papers get recommended to go to LREC or workshops.

Data & annotation methodology *can be* valuable contributions, and reviewing them requires extra expertise. Reviewers who have worked only on modeling should ask to reassign the paper.

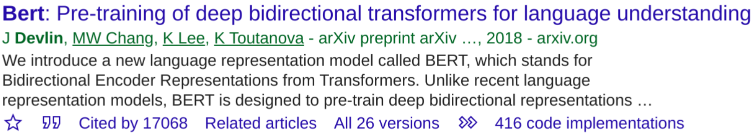
“Niche” heuristic: definition

Difficult question:
Is this paper good?

→

Easy question:
How many people would be interested in it?

“Niche” heuristic example: does the paper involve BERT?



Bert: Pre-training of deep bidirectional transformers for language understanding

J **Devlin**, MW Chang, K Lee, K Toutanova - arXiv preprint arXiv …, 2018 - arxiv.org

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations …

☆ ❞ Cited by 17068 Related articles All 26 versions » 416 code implementations

“Niche” heuristic: issues

- encourages the whole field to do incremental work in the same trendy, popular directions (word2vec craze ➔ BERTology craze ➔ ...?);
- marginalizes everything else into workshops/Findings;

Breakthroughs in niche topics are still breakthroughs.

“Niche language” heuristic: definition

Difficult question:
Is this paper good?

➔

Easy question:
Is it based on English?

“Does it generalize?” fallacy: most monolingual results probably do not transfer between languages!

Author: This works for Japanese.
Reviewer: How do we know that it generalizes to other languages?

Author: This works for English.
Reviewer: Great.

“Niche language” heuristic: issues

- there is little incentive for early-career researchers to try to publish resources for other languages;
- English became the “default” language*, and everything else is marginalized;
- misrepresents NLP progress: we actually only achieved much success with things that are easy in English.

Important work doesn’t *have* to be on English. If there are details on a language you don’t speak, review the parts you can review, and flag the issue for the chairs.

*Bender, E. M. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient.* `https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/`.

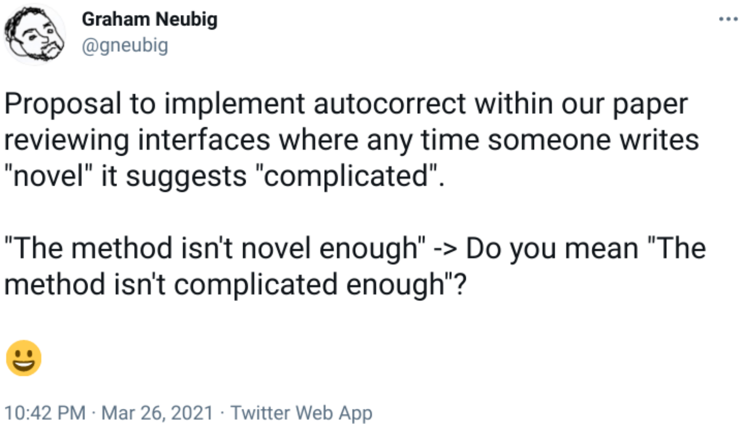
“Too simple” heuristic: definition

Difficult question:
Is this paper good?

→

Easy question:
Does it look like it was a lot of work?

“Too simple” heuristic: novelty != complexity



Graham Neubig
@gneubig

Proposal to implement autocorrect within our paper reviewing interfaces where any time someone writes "novel" it suggests "complicated".

"The method isn't novel enough" -> Do you mean "The method isn't complicated enough"?

😀

10:42 PM · Mar 26, 2021 · Twitter Web App

“Too simple” heuristic: issues

- encourages complex solutions, whether they are needed or not;
- more complexity ➔ potentially more brittleness and reproducibility issues.

The goal is to solve the problem, not to solve it in a fancy way.

“Too risqué” heuristic: definition

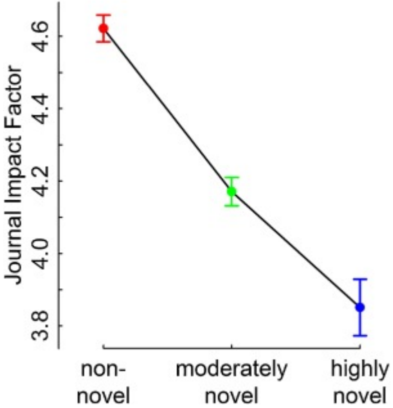
Difficult question:
Is this paper good?

→

Easy question:
Does it have an established precedent?

“Too risqué” heuristic:

Novel papers get into journals with lower impact factors*!



*Wang, J., Veugelers, R. & Stephan, P. 2017. Bias against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators. *Research Policy.* `https://www.sciencedirect.com/science/article/pii/S0048733317301038`.

“Too risqué” heuristic: issues

- favors “unobjectionable” rather than novel research, likely incremental;
- further amplifies the “trendy” topics.

If an idea does not have a clear precedent, it is likely to be judged more harshly. Try to give it a fair chance.

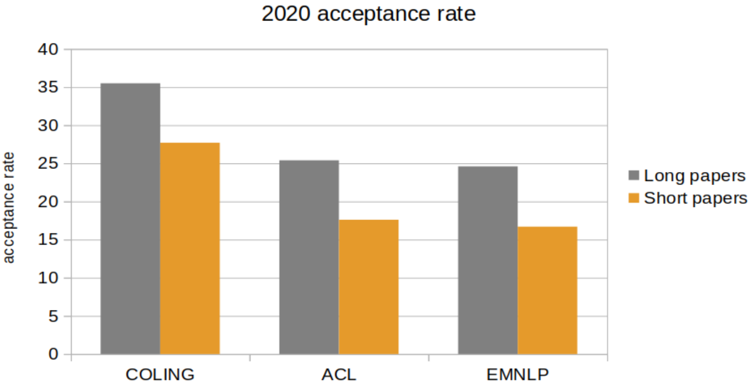
“Could something be added?” heuristic: definition

Difficult question:
Is this paper good?

→

Easy question:
Is it easy to think of something that could be added to this paper?

“Could something be added?” heuristic: short papers are impossible to publish!



“Could something be added?” heuristic: issues

- disadvantages short papers;
- disadvantages smaller labs, which may not have the resources or manpower to preemptively produce “just in case” experiments for a 40-page appendix.

No paper is perfect, and it is *always* possible to add more experiments. Does this paper do enough to make its point convincingly?

Social bias in peer review

Difficult question:
Is this paper good?

→

Easy question:
Is the paper by people who are likely to do good research?

To be discussed in the “Anonymity” section of this tutorial.

Social bias: issues

- decreases the chances for marginalized groups (by gender, race etc.)
- decreases the chances for unknown labs and researchers;
- increases the chances for well-known research groups;
- incentivizes the PR 'arms race'.

If you know who the authors of the paper are, ask to reassign it. The point of bias is that it is unconscious and we cannot control it, even if it feels like we are being impartial.

Church, K. W. Emerging Trends: Reviewing the Reviewers (Again). en. *Natural Language Engineering* **26,** 245–257. ISSN: 1351-3249, 1469-8110. https://www.cambridge.org/core/journals/natural-language-engineering/article/emerging-trends-reviewing-the-reviewers-again/10CDC1D71E1AEB21456CFBDA187CBCB6 (2020) (Mar. 2020).

Ernst, E., Resch, K. & Uher, E. Reviewer Bias. *Annals of Internal Medicine* **116,** 958–958. ISSN: 0003-4819. https://www.acpjournals.org/doi/abs/10.7326/0003-4819-116-11-958_2 (2021) (June 1992).

Crane, M. Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results. en-us. *Transactions of the Association for Computational Linguistics* **6,** 241–252. https://aclweb.org/anthology/papers/Q/Q18/Q18-1018/ (2019) (2018).

Rogers, A. How the Transformers Broke NLP Leaderboards. en. June 2019. https://hackingsemantics.xyz/2019/leaderboards/ (2019).

Rogers, A. Peer Review in NLP: Reject-If-Not-SOTA. en. Apr. 2020. https://hackingsemantics.xyz/2020/reviewing-models/ (2020).

Ethayarajh, K. & Jurafsky, D. Utility Is in the Eye of the User: A Critique of NLP Leaderboards. *arXiv:2009.13888 [cs].* arXiv: 2009.13888 [cs]. http://arxiv.org/abs/2009.13888 (2020) (Sept. 2020).

Mahoney, M. J. Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System. *Cognitive therapy and research* **1,** 161–175 (1977).

Bender, E. M. The #BenderRule: On Naming the Languages We Study and Why It Matters. en. Sept. 2019. https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/ (2020).

Wang, J., Veugelers, R. & Stephan, P. Bias against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators. *Research Policy* **46,** 1416–1436. https://www.sciencedirect.com/science/article/pii/S0048733317301038 (2017).