

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.

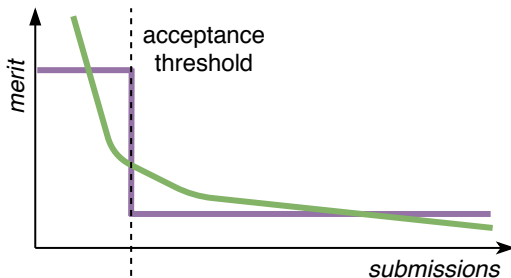
Goals of peer review

- 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

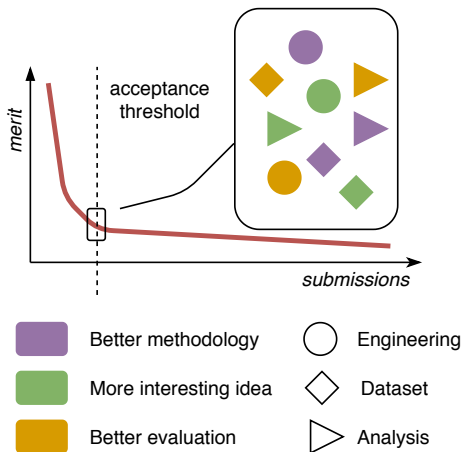


Paper merit distribution, with which
peer review could be reliable



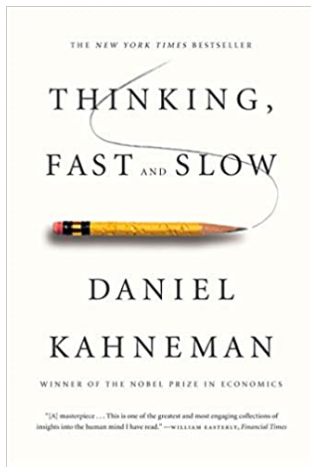
Realistic paper merit distribution,
adapted from Anderson (2009)

Why peer review is a difficult task



How do people reason in high-uncertainty situations?

Biases to the
rescue!



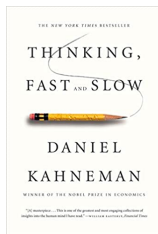
Implicit bias

"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"



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.”



Language heuristic: definition

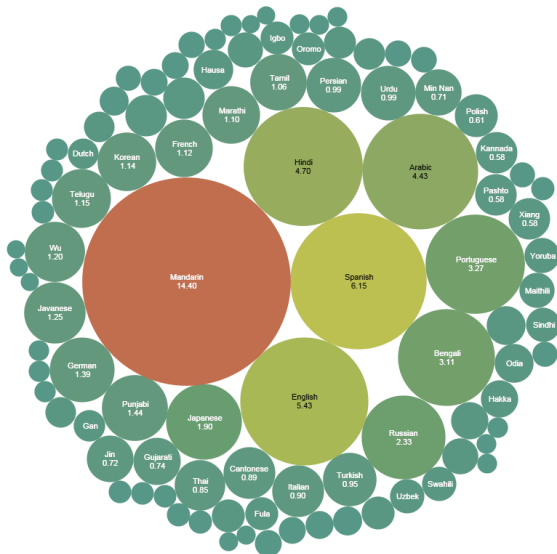
Difficult question:

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 is 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>.

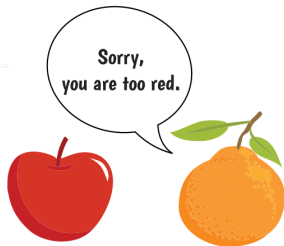
“Preferred methodology” heuristic: definition

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?

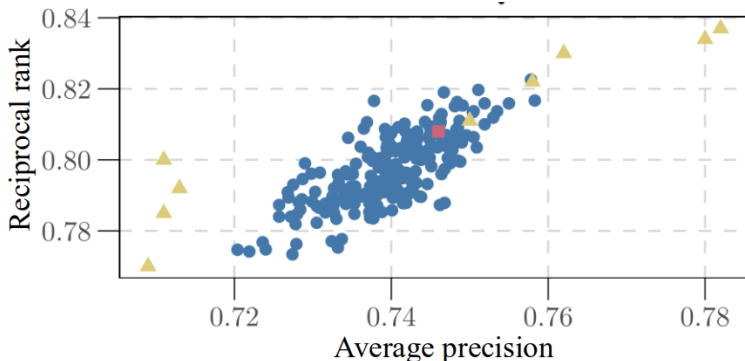


Easy question:

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
work on smth similar and
might be interested?

“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 ...

☆ 97 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://thegradients.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 \neq 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.

“Not-excited” heuristic: definition

Difficult question:
Is this paper good?



Easy question:
Is it interesting for me
personally?

“Not-excited” heuristic: nothing wrong, but...

1. ACL Findings:

To continue the success of Findings at EMNLP 2020, the ACL-IJCNLP 2021 reviewing committee has selected papers that are not accepted for publication in the main conference, but nonetheless have been assessed by the Program Committee as **solid work with sufficient substance, quality and novelty** to Findings. Therefore, the ACL decisions will have the following three basic types : Accept-to-Main-Conference, Accept-to-Findings, and Reject. Some accepted papers are conditional. Please refer to Sections 2 and 3 below for more information.

[https://2021.aclweb.org/blog/acceptance-decision/
#1-acl-findings](https://2021.aclweb.org/blog/acceptance-decision/#1-acl-findings)

“Not-excited” heuristic: issues

- exacerbates perceived randomness in peer review;
- encourages the authors to resubmit multiple times more strain on the system more inexperienced reviewers more randomness

If you do not find a given topic the most exciting:

- do you agree that it overall moves the field forward?
- do you know that there are people who would find this relevant & build on it?

If so, try to evaluate the methodology & results without subconsciously lowering the scores.

“Not-surprising” heuristic: definition

Difficult question:
Is this paper good?



Easy question:
Has it significantly
changed what I already
believed?

“Not-surprising” heuristic: is it worthless if we saw it coming?



Timothy O'Leary
@Timothy0Leary



Dear Peer Reviewers

When did 'surprising' become a requisite for publishing a new and important finding?

Was the discovery of the Higgs Boson surprising? Or the crystal structure of the potassium channel? Or the latest data dump from a neuropixels probe?

Stick to your job.

10:23 AM · Jan 16, 2020 · Twitter for iPhone

“Not-surprising” heuristic: issues

- most things are obvious in hindsight;
- the better the paper is written, the more obvious the conclusion will seem, so the authors are *penalized* for good writing;
- disadvantages much-needed replication & generalization studies.

Most research papers are about confirming, solidifying, and building on existing knowledge.

“Too risqué” heuristic: definition

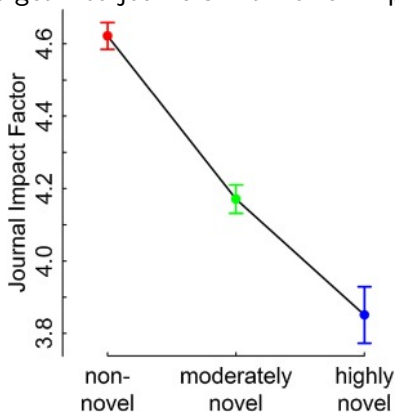
Difficult question:
Is this paper good?



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.

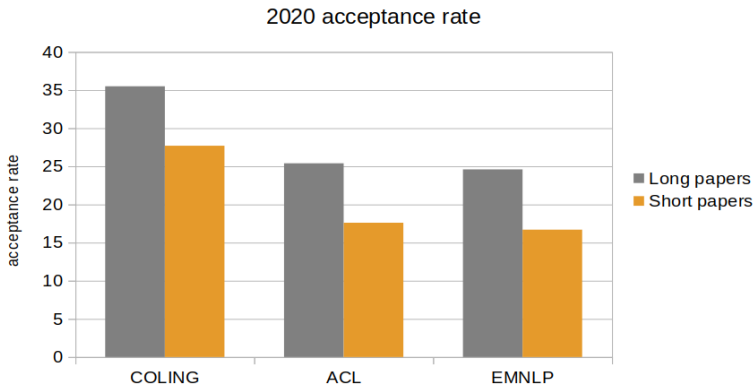
“Could something be added?” heuristic: definition

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.

Discussion break

- Any other biases?
- Any other ideas for addressing them?



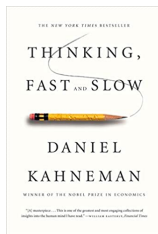
Anonymity in Peer Review



Anna Rogers, University of Copenhagen

The heuristic mechanism

"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."



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?

Thought experiment: imagine a “professor”

Forbes

Mar 27, 2019, 07:34am EDT | 2,372 views

The 3 'Godfathers' Of AI Have Won The Prestigious \$1M Turing Prize



Sam Shoad Former Staff

AI & Big Data

I cover tech in Europe.

 This article is more than 2 years old.



Computer scientist Yoshua Bengio. YOUTUBE/LUCIDWORKS

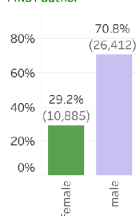
Gender in authorship (ACL anthology)

Kevin B.
Cohen, Karën
Fort, Margot
Mieskes,
Aurélië
Névéol, Anna
Rogers

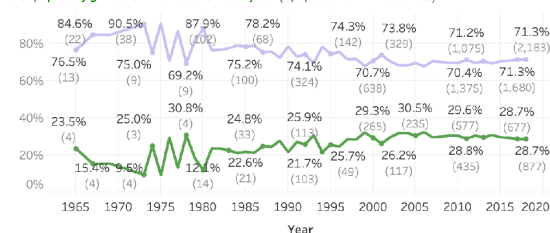
Biases in
Review

Anonymity

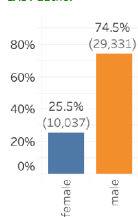
% papers by gender of
FIRST author



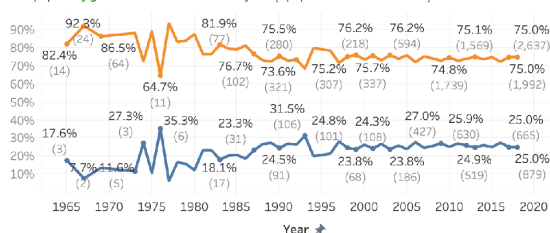
% of papers by gender of FIRST author each year (#papers shown in brackets)



% papers by gender of
LAST author



% of papers by gender of LAST author each year (#papers shown in brackets)



*

*Mohammad, S. M. 2020. Gender Gap in Natural Language Processing Research: Disparities in Authorship and Citations. <https://www.aclweb.org/anthology/2020.acl-main.702>.

The effects of social bias

Reviewers may fall prey to biases:

- in favor of established researchers
- in favor of established labs/institutions
- in favor of the wealthy nations (where the established labs/institutions/researchers are concentrated)
- against marginalized communities (gender, race, LGBTQ...)

Evidence: reputation bias

12 papers were re-submitted* to the same prestigious psychology journals that *already published* those papers.

- the names and institutions of the authors were changed from well-known to unknown;
- only three journals detected the resubmission;
- 16/18 reviewers recommended rejection, often for “serious methodological flaws”!

*Peters, D. P. & Ceci, S. J. 1982. The Fate of Published Articles, Submitted Again. *Behavioral and Brain Sciences*.

<https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/fate-of-published-articles-submitted-again/17914E617A57CDF8ABF3D95F9F28E7FE>.

Evidence: institution and national bias

- bias against smaller institutions*;
- bias in favor of authors from the US and English-speaking countries;†;
- “national publication bias”: European medical journals favoring the authors from their home countries‡;

*Murray, D. L., Morris, D., Lavoie, C., Leavitt, P. R., MacIsaac, H., Masson, M. E. J. & Villard, M.-A. 2016. Bias in Research Grant Evaluation Has Dire Consequences for Small Universities. *PLOS ONE*. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0155876>.

†Ross, J. S., Gross, C. P., Desai, M. M., Hong, Y., Grant, A. O., Daniels, S. R., Hachinski, V. C., Gibbons, R. J., Gardner, T. J. & Krumholz, H. M. 2006. Effect of Blinded Peer Review on Abstract Acceptance. <https://doi.org/10.1001/jama.295.14.1675>.

‡Ernst, E. & Kienbacher, T. 1991. Chauvinism. <https://www.nature.com/articles/352560b0>.

Evidence: social biases

- Even women rate articles higher with a “male” author name on them*!
- in the US context, evidence for racial bias in peer review for grant applications†;
- strong self-reported preference for anonymous review by Chinese early-career researchers‡.

The presence and effect of biases is debated§, but there’s also evidence of bias in citations.

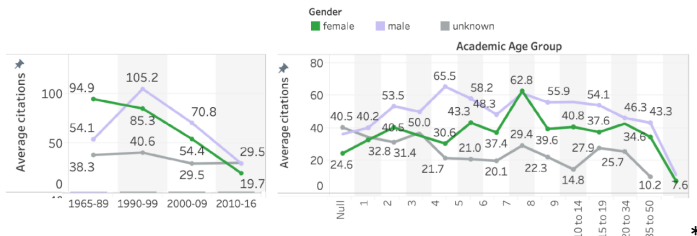
*Goldberg, P. 1968. Are Women Prejudiced against Women?

†Ginther, D. K., Schaffer, W. T., Schnell, J., Masimore, B., Liu, F., Haak, L. L. & Kington, R. 2011. Race, Ethnicity, and NIH Research Awards. <https://science.sciencemag.org/content/333/6045/1015>.

‡Xu, J., Chen, D., He, C., Zeng, Y., Nicholas, D. & Wang, Z. 2020. How Are the New Wave of Chinese Researchers Shaping up in Scholarly Communication Terms? *Malaysian Journal of Library & Information Science*. <https://mjlis.um.edu.my/article/view/27823>.

§Squazzoni, F., Bravo, G., Farjam, M., Marusic, A., Mehmani, B., Willis, M., Birukou, A., Dondio, P. & Grimaldo, F. 2021. Peer Review and Gender Bias: A Study on 145 Scholarly Journals. *Science Advances*. <https://advances.sciencemag.org/content/7/2/eabd0299>; Yang, J., Vannier, M. W., Wang, F., Deng, Y., Ou, F., Bennett, J., Liu, Y. & Wang, G. 2013. A Bibliometric Analysis of Academic Publication and NIH Funding. *Journal of Informetrics*. <https://www.sciencedirect.com/science/article/pii/S175115771200096X>.

Gender gap in citations (ACL anthology)



*Mohammad, S. M. 2020. Gender Gap in Natural Language Processing Research: Disparities in Authorship and Citations. <https://www.aclweb.org/anthology/2020.acl-main.702>.

Slideslive views of ACL 2020 papers



(((yoav' ()J()J)))
@yoavgo

...

I was told that people are too lazy to read code, so here's English: papers whose author names have the bigrams qi, zh, ao, yi, xi, xu and a bunch of others account for half of the papers at the conference but only 42% of paper views, and the median views in this group is lower.



(((yoav' ()J()J))) @yoavgo · Jul 13, 2020

I initially thought I shouldn't go there but then I decided heck, why not. This is a 5 minutes hack and in no means rigorous in any way, but I think it does provide a signal. [twitter.com/yoavgo/status/...](https://twitter.com/yoavgo/status/1281111111111111111)

```
er().strip().split("\t")
|liu |lin ", authors):

(all)*100} % of all papers)")
(all)*100} % of all views)")

), "median of all:", numpy.median
```

275633958103
049488757994
.0 median of
command to

Bottom line

Reputation bias is undisputed, and there is enough evidence of social biases to warrant concerns. Training helps to some extent*, but generally humans struggle to consciously control their biases. So anonymous reviews are the best tool we have.

*Régner, I. *et al.* Committees with Implicit Biases Promote Fewer Women When They Do Not Believe Gender Bias Exists. *Nature human behaviour* 3, 1171–1179. <http://affectfinance.org/wp-content/uploads/2019/09/s41562-019-0686-3.pdf> (2019)

Reviewing models





Open Review Papers, Reviews and Identities are visible to all parties

Single-Blind Author is known to Reviewer – Reviewer remains unknown to Author

Double-Blind Author is unknown to Reviewer and Reviewer is unknown to Author




Pros and cons

fully open:

-  Reviewer is accountable
-  Authors do not need to withhold scientific content to preserve anonymity
-  Reviewer may withhold criticism or face retaliation
-  Vulnerable to all biases





Pros and cons

single-blind:

-  Allows the reviewer to be fully honest
-  Authors do not need to withhold scientific content to preserve anonymity
-  Vulnerable to all biases mentioned above

Pros and cons

double-blind:

-  Avoids various biases
-  Allows the reviewer to be fully honest
-  Reviewer is not accountable
-  Authors need to withhold scientific content to preserve anonymity

Characteristics of Various Venues

(with changes over the years)

Venue	Anonymity / level	Area Chair	Response Period
*ACL	Double-blind	identified	sometimes
EMNLP	Double-blind	identified	yes
IJCAI	Double-blind	internal	yes
AAAI	Double-blind	internal	yes
LREC	Single-blind	no	no
Swisstext	Double-blind	internal	no
TALN	Double-blind	internal	no
AMIA	Single-blind	internal	no

And then there's this thing called arXiv...

The logo for arXiv.org, featuring the text "arXiv.org" in a white, sans-serif font centered on a solid red rectangular background.

arXiv.org

The official ACL anonymity policy:

Anonymity period: from 1 month until submission deadline and until notification or withdrawal

- no posting the preprint of the work under review;
- no publicizing of existing preprints on social media;
- no updates to existing preprints, except for the purpose of correcting names.

https://www.aclweb.org/adminwiki/index.php?title=ACL_Policies_for_Submission,_Review_and_Citation

Issues with the anonymity period

- delays circulation of results, but does not fully solve the problem: the reviewers can still look up or accidentally discover the identities of authors of the preprinted papers;
- shifts the deadline: if possible, the authors now publish preprints a month ahead.

Other issues with anonymity

- some papers get so well-known that anonymous reviewing is impossible;
- anonymity breaches through personal communication, talks etc.;
- it is often possible to tell by the paper itself whether the authors are native speakers from a well-resourced lab;
- in a narrow subfield, an expert might still even guess the specific authors by the topic and the focus of the study.

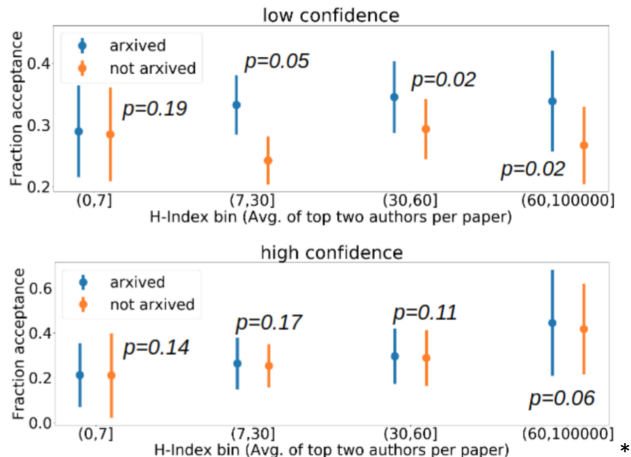
Solutions

- Take/promote training* in biases in peer review;
- Do the best you can to maintain the anonymous process.

- ✓ Do not actively try to discover the authors.
- ✓ If you know the paper, ask to reassign it.
- ✓ Do not search for related work until you have read the paper and formed your opinion.
- ✓ If it is an *ACL conference, report any breaches of anonymity period that you may observe.

*Régner, I. *et al.* Committees with Implicit Biases Promote Fewer Women When They Do Not Believe Gender Bias Exists. *Nature human behaviour* 3, 1171–1179. <http://affectfinance.org/wp-content/uploads/2019/09/s41562-019-0686-3.pdf> (2019)

Especially if your confidence is low!



*Bharadhwaj, H., Turpin, D., Garg, A. & Anderson, A. 2020. De-Anonymization of Authors through arXiv Submissions during Double-Blind Review. *arXiv:2007.00177 [cs]*. <http://arxiv.org/abs/2007.00177>.

Anonymity during rebuttals

- *This is John Smith, I am the reviewer 1...*

! The conference should provide explicit instructions about the expected level of anonymity during rebuttal. If the process is anonymous, do NOT reveal your names to the fellow reviewers and/or area chairs.

Biases to watch out for during rebuttals:

- *Confirmation bias*: you have already formed an opinion of the paper, and humans don't like to change their mind. Seriously consider that you may be wrong, and give the authors a fair chance to persuade you.
- *Emotional reaction*: neither authors nor reviewers are consistently polite. If the response made you angry, still try to focus on the content and not the tone.
- *Groupthink*: after you know the opinion of other reviewers, you may regress to the mean*.
- In a non-anonymous discussion, you may regress to the opinion of the high-reputation senior reviewers.

*Gao, Y., Eger, S., Kuznetsov, I., Gurevych, I. & Miyao, Y. 2019. Does My Rebuttal Matter? Insights from a Major NLP Conference. <https://www.aclweb.org/anthology/N19-1129>.

Holding the reviewers accountable

- *ACL conferences now ask area chairs to run quality checks before rebuttals. Submit early enough for that to be possible, and make corrections as needed.
- If using secondary reviewers - declare them.
- Accountability mechanisms vary by venue and are changing. Stay up-to-date about the policies of your venues.



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).



Mohammad, S. M. Gender Gap in Natural Language Processing Research: Disparities in Authorship and Citations. in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Online, July 2020), 7860–7870.
<https://www.aclweb.org/anthology/2020.acl-main.702> (2020).



Peters, D. P. & Ceci, S. J. The Fate of Published Articles, Submitted Again. en. *Behavioral and Brain Sciences* **5**, 199–199. ISSN: 1469-1825, 0140-525X. <https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/fate-of-published-articles-submitted-again/17914E617A57CDF8ABF3D95F9F28E7FE> (2020) (June 1982).



Murray, D. L. *et al.* Bias in Research Grant Evaluation Has Dire Consequences for Small Universities. en. *PLOS ONE* **11**, e0155876. ISSN: 1932-6203. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0155876> (2021) (June 2016).



Ross, J. S. *et al.* Effect of Blinded Peer Review on Abstract Acceptance. **295**, 1675–1680. ISSN: 0098-7484. <https://doi.org/10.1001/jama.295.14.1675> (2021) (Apr. 2006).



Ernst, E. & Kienbacher, T. Chauvinism. en. **352**, 560–560. ISSN: 1476-4687. <https://www.nature.com/articles/352560b0> (2021) (Aug. 1991).



Goldberg, P. Are Women Prejudiced against Women? **5**, 28–30 (1968).



Ginther, D. K. *et al.* Race, Ethnicity, and NIH Research Awards. en. **333**, 1015–1019. ISSN: 0036-8075, 1095-9203. <https://science.sciencemag.org/content/333/6045/1015> (2021) (Aug. 2011).



Xu, J. *et al.* How Are the New Wave of Chinese Researchers Shaping up in Scholarly Communication Terms? en. *Malaysian Journal of Library & Information Science* **25**, 49–70. ISSN: 1394-6234. <https://mjlis.um.edu.my/article/view/27823> (2021) (Dec. 2020).



Squazzoni, F. *et al.* Peer Review and Gender Bias: A Study on 145 Scholarly Journals. en. *Science Advances* **7**, eabd0299. ISSN: 2375-2548. <https://advances.sciencemag.org/content/7/2/eabd0299> (2021) (Jan. 2021).



Yang, J. *et al.* A Bibliometric Analysis of Academic Publication and NIH Funding. en. *Journal of Informetrics* **7**, 318–324. ISSN: 1751-1577. <https://www.sciencedirect.com/science/article/pii/S175115771200096X> (2021) (Apr. 2013).



Régner, I., Thinus-Blanc, C., Netter, A., Schmader, T. & Huguet, P. Committees with Implicit Biases Promote Fewer Women When They Do Not Believe Gender Bias Exists. *Nature human behaviour* **3**, 1171–1179. <http://affectfinance.org/wp-content/uploads/2019/09/s41562-019-0686-3.pdf> (2019).



Bharadhwaj, H., Turpin, D., Garg, A. & Anderson, A. De-Anonymization of Authors through arXiv Submissions during Double-Blind Review. *arXiv:2007.00177 [cs]*. arXiv: 2007.00177 [cs]. <http://arxiv.org/abs/2007.00177> (2020) (June 2020).



Gao, Y., Eger, S., Kuznetsov, I., Gurevych, I. & Miyao, Y. Does My Rebuttal Matter? Insights from a Major NLP Conference. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Association for Computational Linguistics, Minneapolis, Minnesota, June 2019), 1274–1290. <https://www.aclweb.org/anthology/N19-1129> (2020).