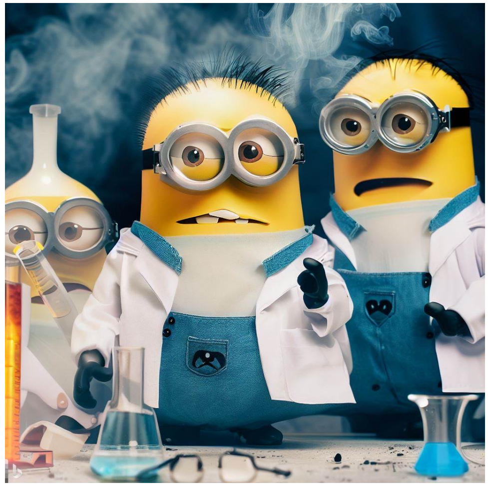
What to do about NeurIPS Reviewer #2? Unearthing Peer Review's Secrets through Scientific Experiments

Nihar B. Shah

Machine Learning and Computer Science Departments

Carnegie Mellon University





Preliminaries

- Overview article: bit.ly/PeerReviewOverview
- Slides available online (see NeurIPS abstract page)
- Multi-disciplinary research on peer review
  - Many studies conducted outside of computer science
  - Pictorial examples tailored to machine learning for illustration
  - Studies in computer science are accompanied by conference names

Objectives of Peer Review

Ensure rigor of published research

Filter to select more interesting or better research

Additionally: feedback to authors, improve the research, learning experience for reviewers

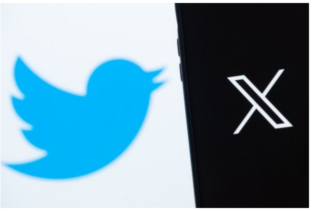
[Benos et al. 07, Wing et al. 11, Jefferson et al. 02, Smith 97]

Problems in peer review…

- Hamper scientific progress [Travis et al. 1991]
- Hurt careers (rich gets richer) [Triggle et al. 2007, Merton 1968]
- Negatively affect wellbeing [Allen et al. 2020, Han et al. 2019, Evans et al. 2011]
- In medical research can harm patients [Poutoglidou et al. 2022]
- Wasteful allocation of up to billions of dollars in annual grants [Fang et al. 2016]
- Degrade public perception of science [Wing et al. 2011, Jamieson 2018, Kharasch et al. 2021]

Objectives of this tutorial

- Make the community **congnizant of systemic** problems in peer review
- Promote **discourse based on scientific principles**



We should just do [blah] and the problem will be solved!

- Inform reviewers about (subconcious) **reviewing pitfalls**
- Catalyze evidence-based **policies**
- Highlight **technical open problems**

Outline

- Peer-review policies
- Seen such a review?
- Reviewer incentives
- Objectives of peer review
- Epilogue

Outline

- **Peer-review policies**
- **Seen such a review?**
- **Reviewer incentives**
- **Objectives of peer review**
- **Epilogue**

Peer review policies

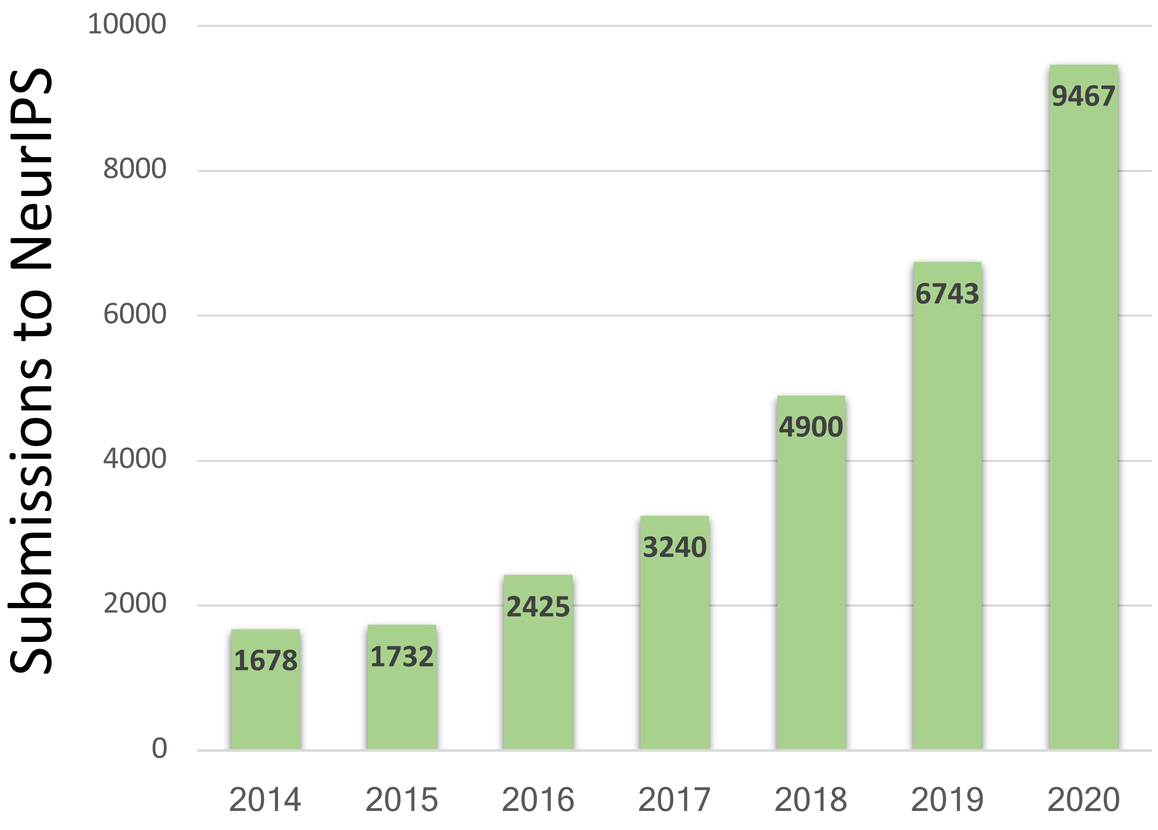
Reviewer pool creation — Paper submission — Reviews — Authors' rebuttal — Discussions — Decisions

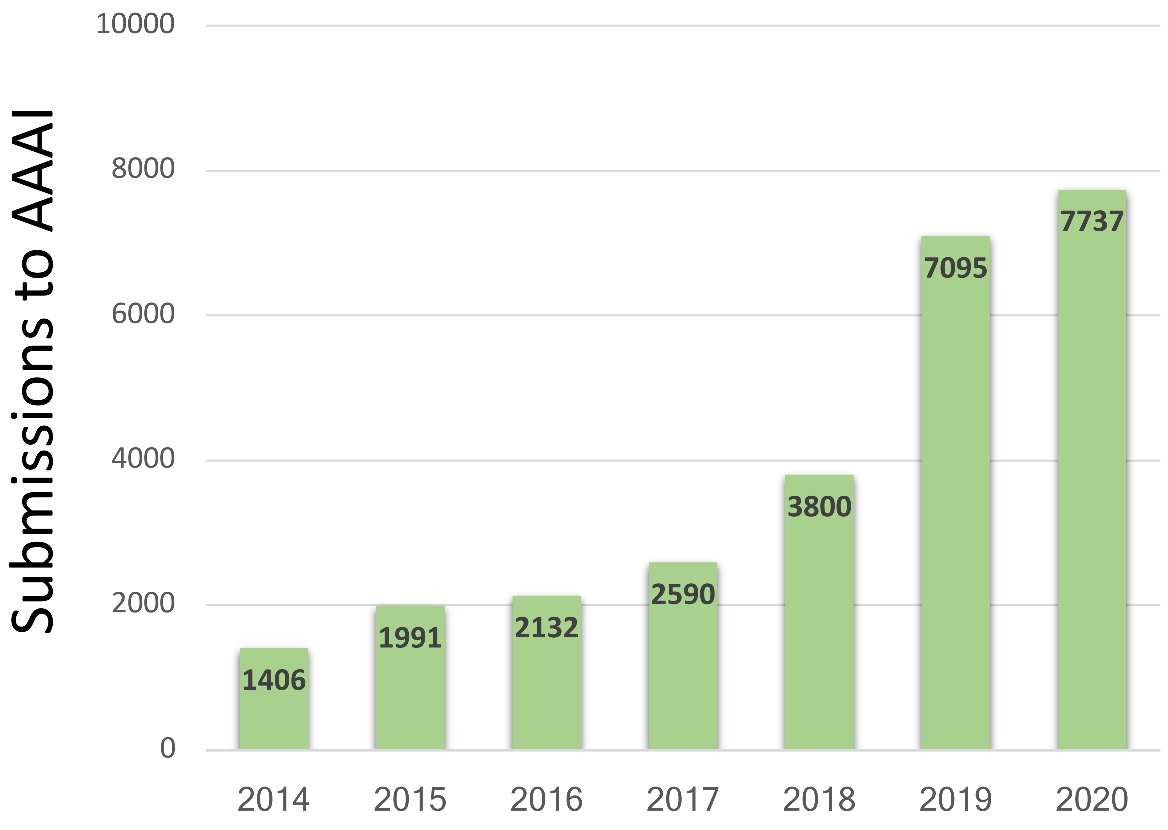
Alright, so here's what everyone must do…

Peer review policies

Reviewer pool creation — Paper submission — Reviews — Authors' rebuttal — Discussions — Decisions

Alright, so here's what everyone must do…





In other disciplines: "Submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint" [McCook 2006]

Reviewer training and mentoring

- **ICML 2020:** Junior reviewer **selection and mentoring** [Stelmakh et al. 2021]

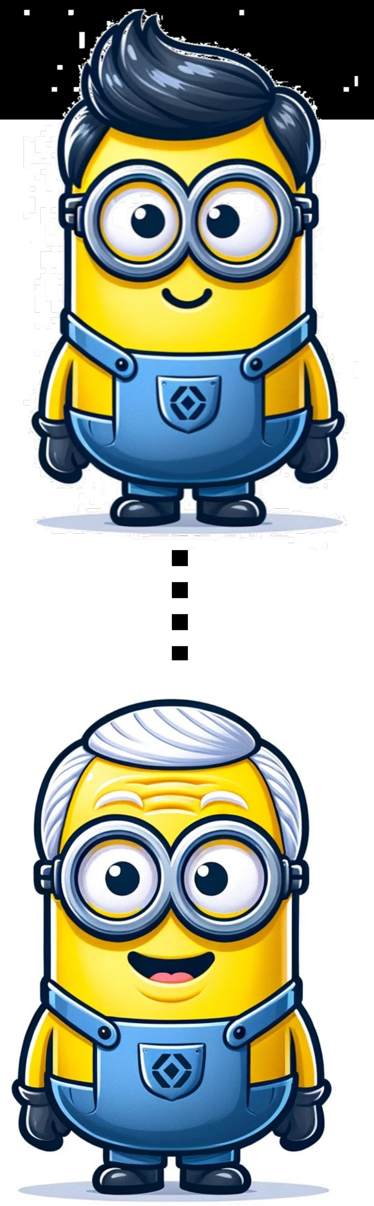
	Junior	Regular	P-value
Positive bids	34.6	27.4	0.43
Fraction of timely review submission	0.92	0.81	0.41
Review length (characters)	4759	2858	**<0.001**
Fraction review updated after rebuttal	0.61	0.43	**<0.001**
Fraction active in discussion	0.68	0.58	0.33
Meta-reviewer’s evaluation of review quality	2.26	2.08	**<0.001**

- **Grant proposal review:** For both novice and experienced reviewers, a **training** video increased the inter-reviewer agreement, improved alignment with rubrics, reviewers spent more time to read the review criteria [Sattler et al. 2015]

Reviewer training and mentoring

Longitudinal studies: Quality of an individual's review falls over time, at a slow but steady rate [Callaham et al. 2011; Joyner et al. 2020]

Randomized controlled trial: Reviewer performance can initially be better by training them, but the quality of trained and untrained reviewers becomes indistinguishable six months after the training [Schroter et al. 2004]



Peer review policies

Reviewer pool creation — **Paper submission** — Reviews — Authors' rebuttal — Discussions — Decisions

Alright, so here's what everyone must do…

- Previous rejections in ICLR and other venues publicly available online

- Many conferences ask authors to declare previous rejections of submitted paper

IJCAI-PRICAI YOKOHAMA 2020

"The cover letter should be inserted at the beginning of the submitted PDF, along with the previous reviews and previous anonymized rejected submission, before the 6+1 pages of the paper

Do reviewers get biased when they know that the paper they are reviewing was previously rejected from a similar venue?

Randomized controlled trial

- Associated to ICML 2020
- 134 junior reviewers each reviewing 1 paper
- Randomly divided into:

A SUPER* Algorithm to Optimize Paper Bidding in Peer Review

Author checklist:
- If applicable, will you make the code and data publicly available upon acceptance?
  Answer: Yes

Abstract

A number of applications involve the sequential arrival of users, and require showing each user a set of items. It is well known that the order in which the items are presented to a user can have a

In typical peer review process, when the bidding phase begins, reviewers enter the system in an arbitrary sequential order. Upon entering, a list of papers is shown and the reviewer places bids on papers they would prefer to review.

It is known that the order of papers presented to reviewers

Control condition

A SUPER* Algorithm to Optimize Paper Bidding in Peer Review

Author checklist:
- If applicable, will you make the code and data publicly available upon acceptance?
  Answer: Yes
- Was this paper submitted to NeurIPS'19?
  Answer: **Yes, the paper was rejected from NeurIPS**

Abstract

A number of applications involve the sequential arrival of users, and require showing each user a set of items. It is well known that the order in which the items are presented to a user can have a

In typical peer review process, when the bidding phase begins, reviewers enter the system in an arbitrary sequential order. Upon entering, a list of papers is shown and the reviewer places bids on papers they would prefer to review.

It is known that the order of papers presented to reviewers

Treatment condition

[Stelmakh et al. 2021]

Reviewers biased against resubmissions

	Score difference (10-pt scale)	P-value
Overall score	-0.78	**0.036**
Quality	-0.46	**0.005**
Clarity	-0.44	**0.022**
Significance	-0.36	**0.037**
Originality	-0.21	0.105
Confidence	-0.01	0.902

[Stelmakh et al. 2021]

Peer review policies

Reviewer pool creation — Paper submission — **Reviews** — Authors' rebuttal — Discussions — Decisions

Alright, so here's what everyone must do…

REVIEW #2

You should find some male researchers to work with

True story

Review in PLOS ONE, 2015

Authors: Fiona Ingleby, Megan Head

https://www.science.org/content/article/plos-one-ousts-reviewer-editor-after-sexist-peer-review-storm

Single blind versus double blind

A Principled Interpretation of Minion Speak

S. Overkill and F. Gru
Cartoony Minion University

In this paper we present a new understanding of…

A Principled Interpretation of Minion Speak

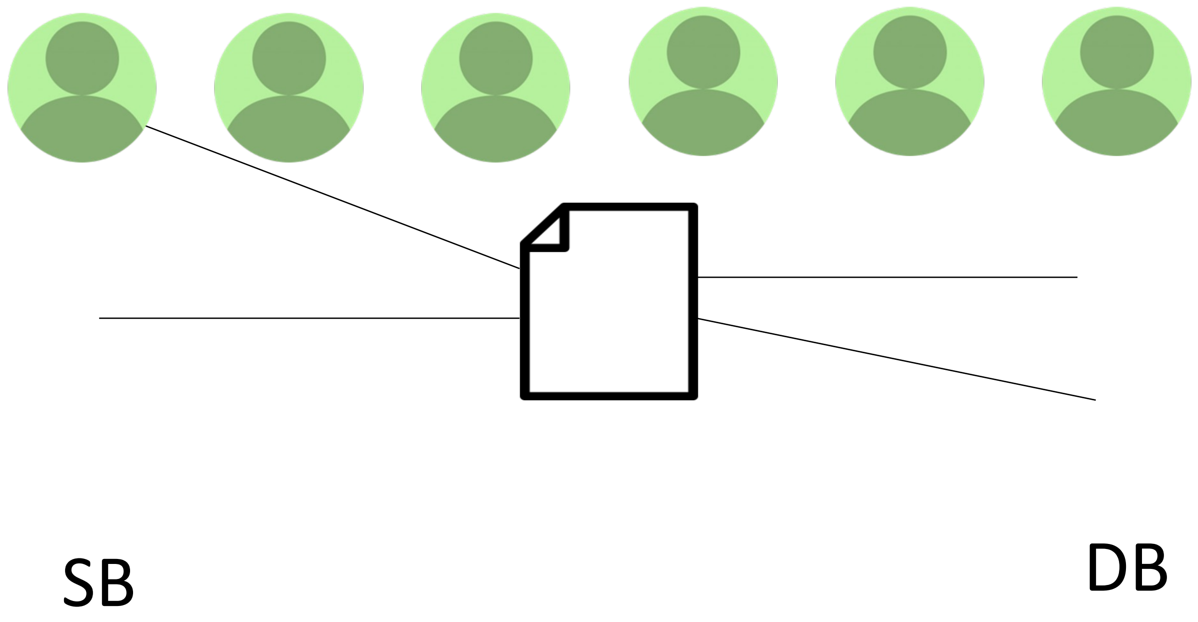
Anonymous Authors
Anonymous Affiliation

In this paper we present a new understanding of…

"Single blind leads to biases with respect to fame/gender/race/…"

"Author identities may be useful…Where is the evidence of bias in my research community?"

WSDM'17 experiment



- Reviewers randomly split into single blind (SB) and double blind (DB) conditions
- Each paper assigned 2 SB reviewers and 2 DB reviewers

[Tomkins et al. 2018]

WSDM'17 experiment

- Famous author
- Top university
- Top company

Significant bias

- At least one woman author

Not statistically significant; high effect size
Meta analysis is statistically significant

- From USA vs. not
- Academia vs. industry
- Reviewer same country as author

No evidence of bias

WSDM moved to double blind from the following year

[Tomkins et al. 2018]

Some issues with experimental methods [Stelmakh et al., 2019]

Many other studies

- Biases in review text [Manzoor et al. 2021]
  - Uses ICLR's switch from single to double blind
  - Evidence of affiliation bias; no evidence of gender bias

- Studies on single-blind bias in other fields [Section 7.2 of bit.ly/PeerReviewOverview]
  - Affiliation bias found consistently
  - Mixed evidence for gender and other biases

- *"Author identities may sometimes be useful."* ITCS 2023 experiment:
  - Reviewers are allowed to use author identities, after giving initial unbiased evaluations? [Shah 2023]
  - 7% overall scores changed; uncorrelated with author affiliations

Ban arXiv and social media

No restrictions

- **ICML 2021 and EC 2021:** 36% and 42% reviewers (anonymously) self-reported actively searching online for the paper they were reviewing [Rastogi et al. 2022]

- **PLDI, OOPSLA, ASE:** Reviewers provided guesses of author identities with 70%-86% reviews. Among these, 72%-85% guessed at least one author correctly [Le Goues et al. 2018]

- Paper's content can reveal authors; algorithms can identify authors to a moderate degree [Hill et al. 2003; Caragea et al. 2019; Matsubara et al. 2020]

- "Embargo periods" debated in NLP/Vision communities: ICML 2021 and EC 2021 experiments find no difference in preprint posting and visibility during versus outside embargo periods [Rastogi et al. 2022]

Peer review policies

Reviewer pool creation — Paper submission — Reviews — **Authors' rebuttal** — Discussions — Decisions

Alright, so here's what everyone must do…

Authors' rebuttal

NAACL 2015, NeurIPS 2016, ACL 2017: *Only 10-20% review scores changed after rebuttal*

Why?

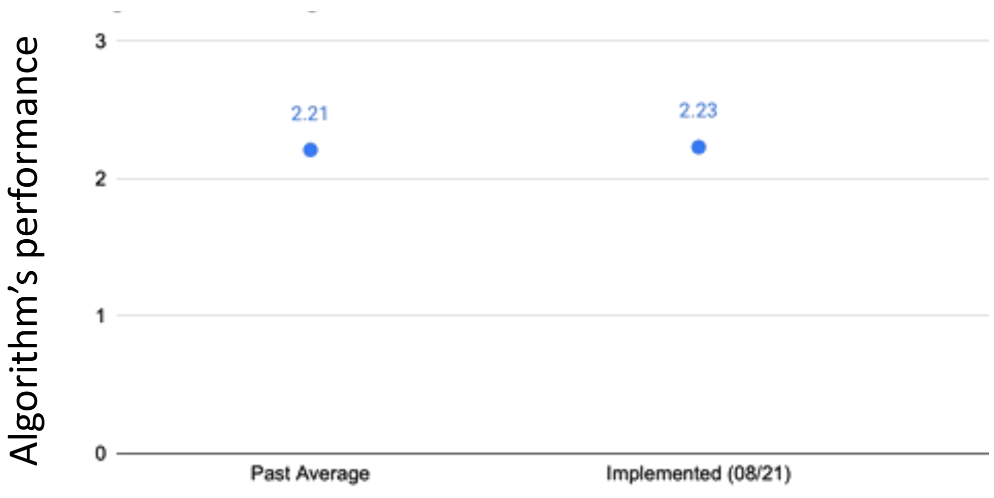
- Many reviewers don't show up for rebuttal/discussions
- Even if they show up, they don't change their opinion

Is it due to "anchoring bias"? [Liu et al. 2023]

- People make an estimate by starting from an initial value (pre-rebuttal score) and then adjust it to yield their answer, but this adjustment is insufficiently small [Tversky & Kahneman 1974]

Are reviewers anchored to their initial low score?

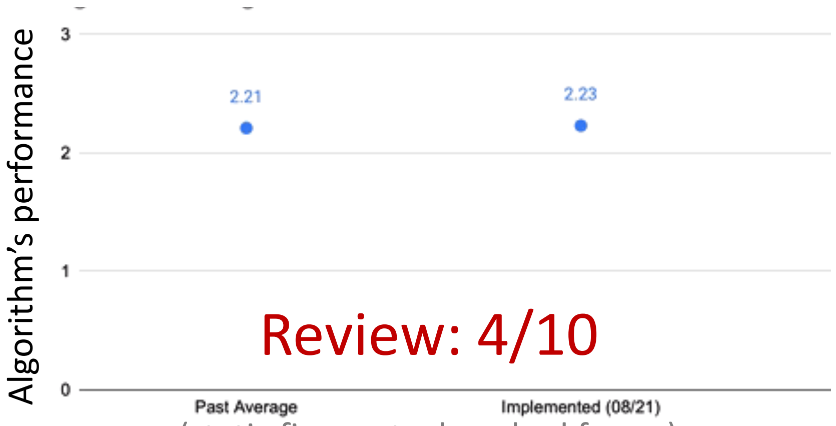
Randomized controlled trial



(animated figure)

Review: 8/10

vs.

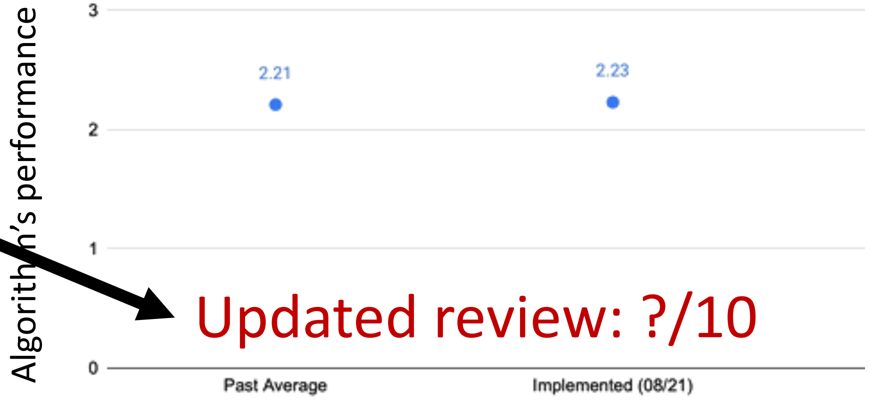


Review: 4/10

(static figure stuck on bad frame)

Some issue with your browser. Hit "refresh".

No Difference



Updated review: ?/10

(animated figure)

[Liu et al. 2023]

Peer review policies

Reviewer pool creation — Paper submission — Reviews — Authors' rebuttal — **Discussions** — Decisions

Alright, so here's what everyone must do…

Inter-reviewer discussions

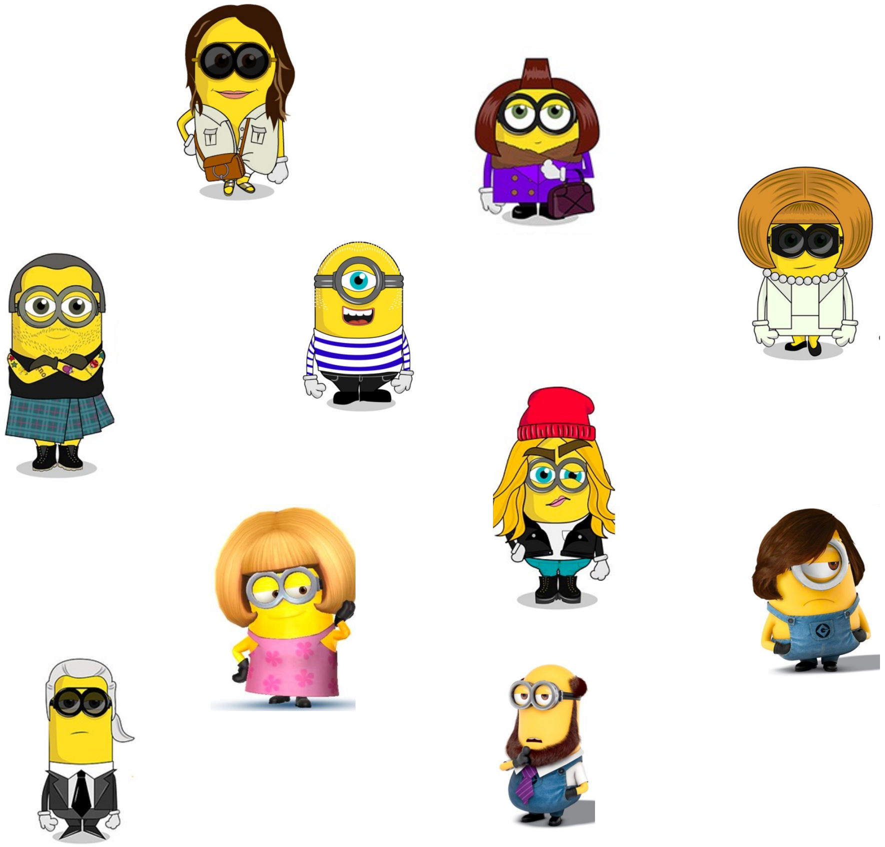
Reviewer #3: This paper is a real peach!

Reviewer #1: You reminded me of the peach fruit!

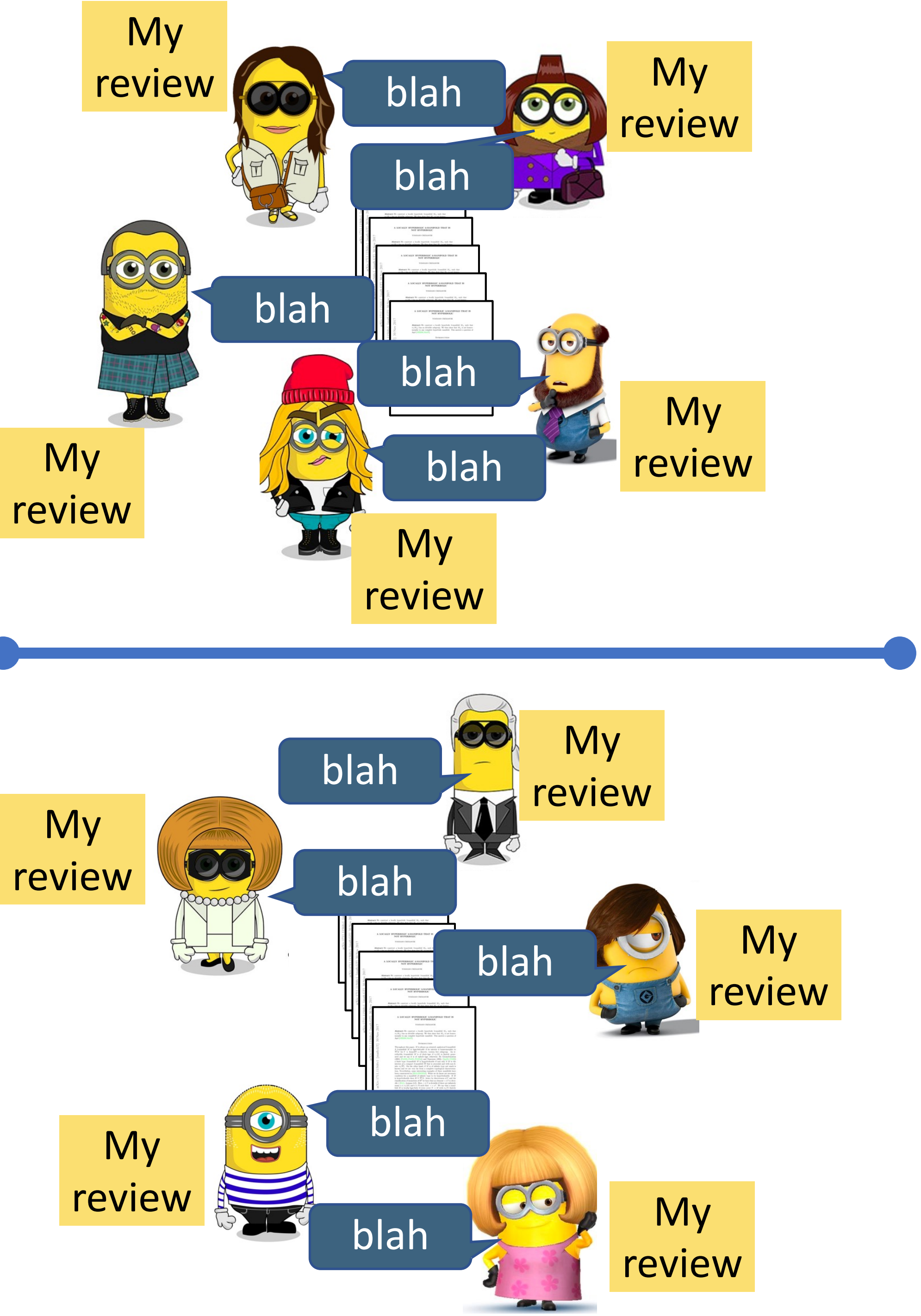
Reviewer #2: Peaches taste yuck. Reject.

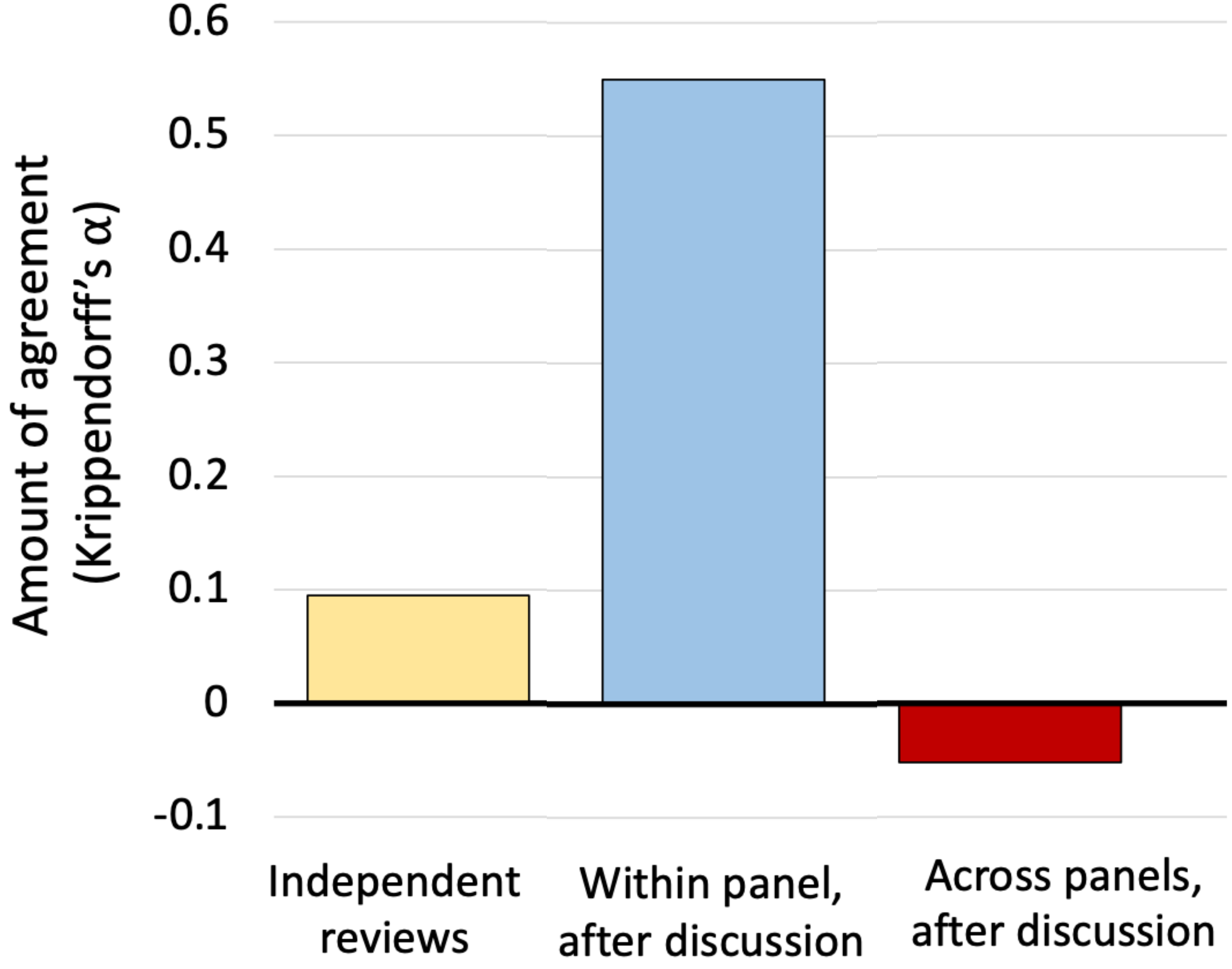
- (In)consistency
- Herding
- Superfluous influence
- Anonymity

(In)consistency of outcomes



[Obrecht et al. 2007; Fogelholm et al. 2012; Pier et al. 2017]





[Obrecht et al. 2007; Fogelholm et al. 2012; Pier et al. 2017]

Herding

- In many other settings: decision of a group **biased towards the opinion of the group member who initiates the discussion**. [Asch 1951, McGuire et al. 1987; Dubrovsky et al. 1991; Weisband 1992; Banerjee 1992]

- In our review processes, no specified policy on who initiates the discussion.

If herding exists in peer-review discussions, then problematic:
Final decisions depends on who initiated discussion

Randomized controlled trial at ICML 2020

- 1500 papers, 2000 reviewers
- Split papers uniformly at random into two groups

First ask most positive reviewer to start the discussion, then later ask the most negative reviewer to contribute to the discussion.

First ask the most negative reviewer to start the discussion, then later ask the most positive reviewer to contribute to the discussion.

If herding, acceptance rate in left condition > right condition

Result: No difference in outcome (i.e., no evidence of herding)

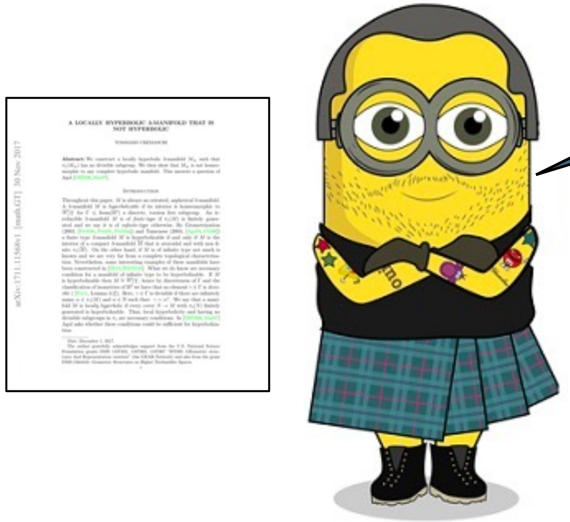
[Stelmakh et al. 2020]

Superfluous influence

“Other reviewers gave this paper scores of 2 to 5. You may update your score if you see fit.”

Good paper. 8/10.

I’ll update my score to 6/10.



- Large fraction of reviewers updated their scores
- P(reduced updated score | high initial score) >> P(increased updated score | low initial score)
- In first study: women changed much more often than men, highly cited researchers changed less often

[Teplitskiy et al. 2019, Lane et al. 2022]

Anonymity Discussions can compromise anonymity of reviewers to authors

1. Timing of discussion posts

Based on analysis of major conference [Goldberg et al. 2023]

9.00 am Dec 11, 2023
Scarlett Overkill (Reviewer #1) commented on paper 44 that you are also reviewing: …

9.02 am Dec 11, 2023
Anonymous Reviewer #2 commented on paper 63 that you have authored: *"Bad paper. Reject."*

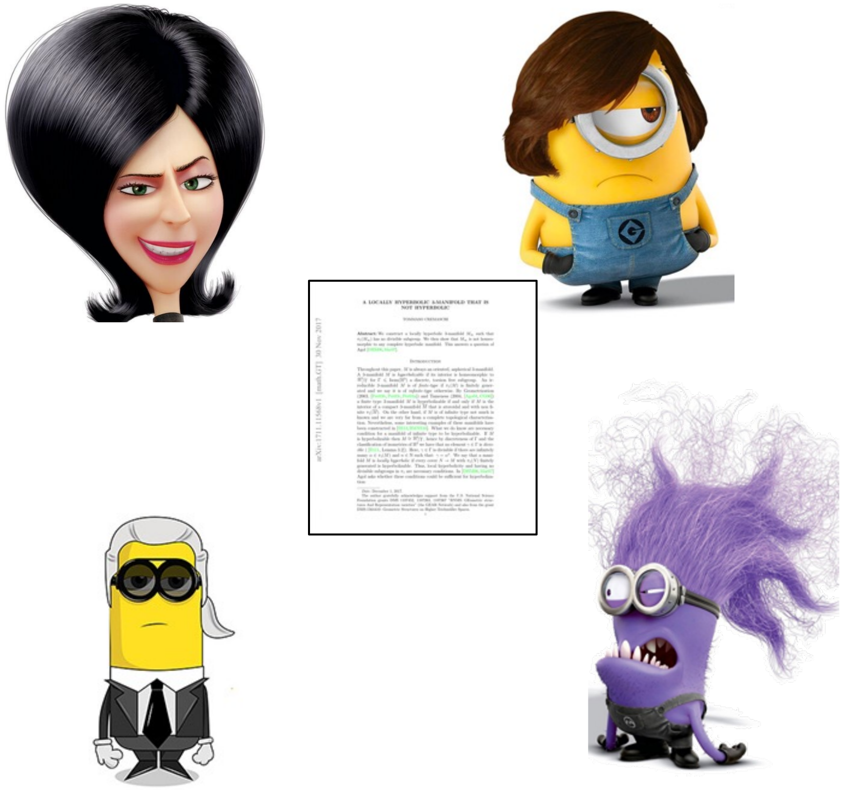
Anonymity

2. Mole in review panel

Author

Psst.. Scarlett Overkill is Reviewer #1

Reviewer discussion



Anonymity

2. Mole in review panel

Author

Reviewer discussion

- Anecdotal evidence [Lauer 2020]
- **UAI 2022 experiment** [Rastogi et al. 2023]
  - ~7% reported experiencing such an issue either in UAI or another conference
  - Solution: Anonymize reviewers to each other; also reduces biases

Peer review policies

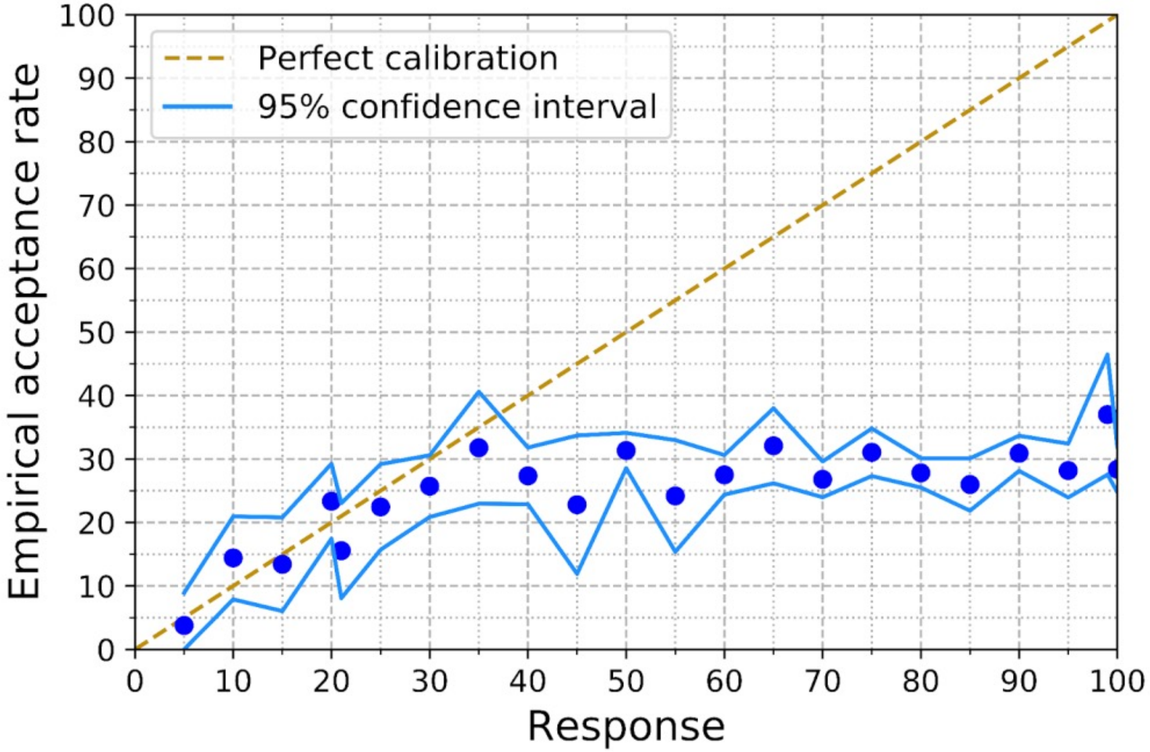
Reviewer pool creation — Paper submission — Reviews — Authors' rebuttal — Discussions — **Decisions**

Alright, so here's what everyone must do…

2021 Experiment on Author Perceptions

[During submission] "What is your best estimate of the probability (as a percentage) that this submission will be accepted? (Acceptance rate of previous 4 years = 21%)"

Mean prediction = 67%



[Rastogi et al. 2023]

Outline

- Peer-review policies
- **Seen such a review?**
- Reviewer incentives
- Objectives of peer review
- Epilogue

Seen such a review?

- Dr. Fox effect
- Surprisingness bias
- Confirmation bias
- Positive-outcome bias
- Citation bias
- Commensuration bias
- Miscalibration

Dr. Fox effect

Too simple. Reject.

We add up the two values…

Denote the two values as α and β and compute $(\alpha + \beta)$…

What an advanced paper. Accept

[Armstrong 1980]

Dr. Fox effect

Complex presentation can influence reviewers positively

"If you can't convince them, confuse them"

https://www.youtube.com/watch?v=RcxW6nrWwtc
"The Dr. Fox Lecture"

[Naftulin et al. 1972]

[Armstrong 1980]

Surprisingness bias

“Does RLHF for safety reduce accuracy of model?” Yes/No

Hindsight group

Result: [chosen at random from the two possibilities]

How surprising is this result?

Foresight group

~~**Result: …**~~

How surprising would it be if:

- *it reduces accuracy?*
- *it does not reduce accuracy?*

[Slovic et al. 1977]

Surprisingness bias

Too obvious. Reject!

The answer is not obvious to me.

- Less surprising when reviewer had read the results (hindsight group).
- Difference between hindsight and foresight reduces if the hindsight group is additionally asked a counterfactual question *“How surprised would you have been if the result was the opposite?”*
- **When writing manuscripts, stress the unpredictability of the results and make the reader think about the counterfactual.**

[Slovic et al. 77]

Confirmation bias

✌️ Vision transformers

Review: Wonderful paper with rigorous methods. Accept!

Question: Vision Transformers vs. Convolutional Neural Networks

Methods:...

Conclusion: Vision Transformers beat Convolutional Neural Networks

❤️ Convolutional neural networks

Review: Poor paper with fatally flawed methods. Reject!

[Mahoney 1977, Travis et al. 1991, Ernst et al. 1994]

Confirmation bias

Reviewers are favorable to those manuscripts whose results agree with the reviewer's own views.

Papers that agreed with reviewer's views:

- rated as methodologically better
- as having better data presentation
- making a higher overall scientific contribution

[Mahoney 1977, Travis et al. 1991, Ernst et al. 1994]

Positive-outcome bias

Can GPT-4 win a gold medal in the International Mathematical Olympiad?

Accept!

Introduction…
Methods…

Result: Yes

Introduction…
Methods…

Result: No

Reject!

[Emerson et al. 2010]

Positive-outcome bias

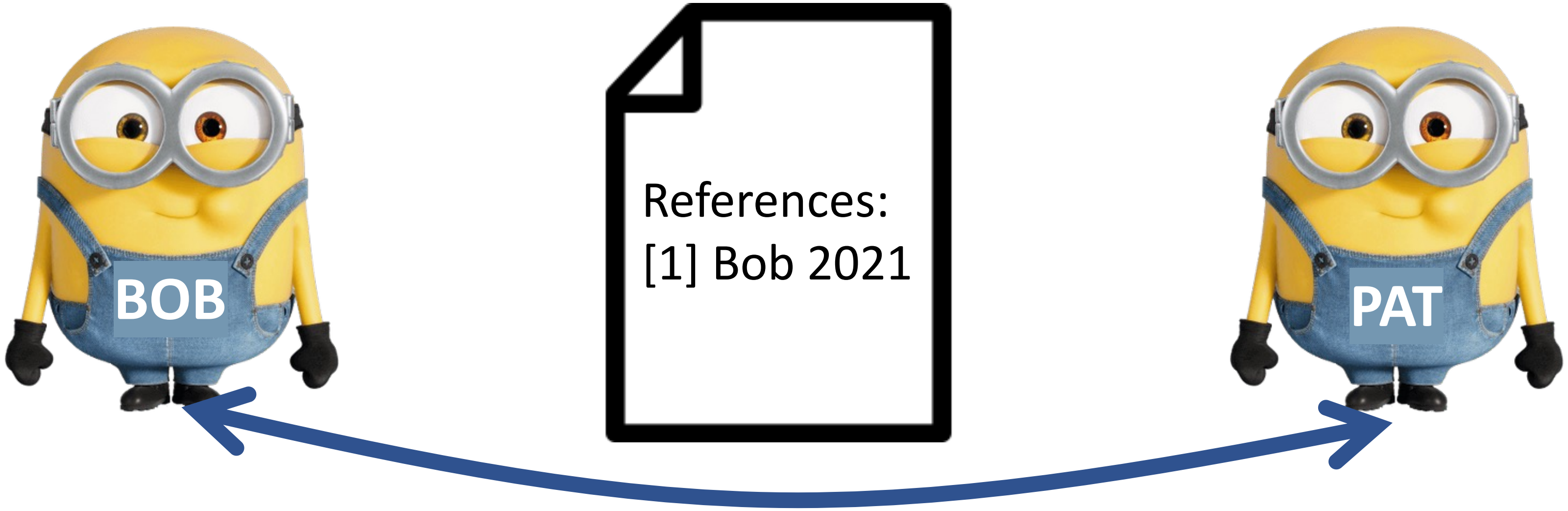
Reviewers also detected roughly twice as many (deliberately inserted) errors in the negative-outcome version. [Emerson et al. 2010]

Solutions:

- Submission for review only contains intro and methods, but no results [Smulders 2013]
- Bias incentivizes authors to get "positive results" (p-hacking, HARKing). Solution: preregister experiments [Nosek et al. 2018]

Citation bias

Experiments at EC 2021 and ICML 2020



Reviewers identical in other ways: bids, paper-reviewer similarity, self-reported expertise, reviewer seniority, paper-dependent factors, and no genuinely missing citations.

Conference	Score difference	P-value
ICML 2020	0.16 (6-point scale)	**0.004**
EC 2021	0.23 (5-point scale)	**0.009**

Cited reviewers more positive

[Rastogi et al. 2022]

Commensuration bias

Reviewers have differing opinions about relative importance of different criteria

Theory: 10
Experiments: 0
Clarity: 8
Overall score: 2 (Reject)

Theory: 10
Experiments: 0
Clarity: 8
Overall score: 9 (Accept)

“Commensuration bias”

Reviewers have different mappings from criteria scores to overall scores

Leads to arbitrariness/unfairness in the review process

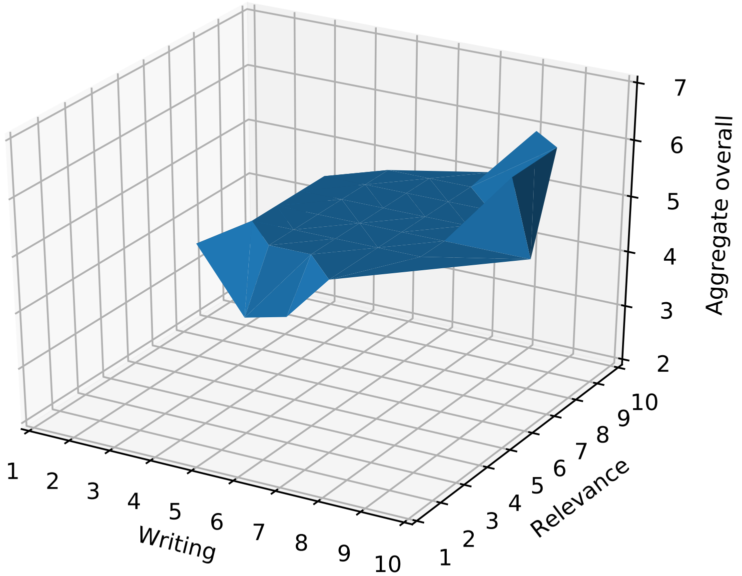
[Kerr et al. 1977, Bakanic et al. 1987, Hojat et al. 2003, Lamont 2009, **Lee 2015**]

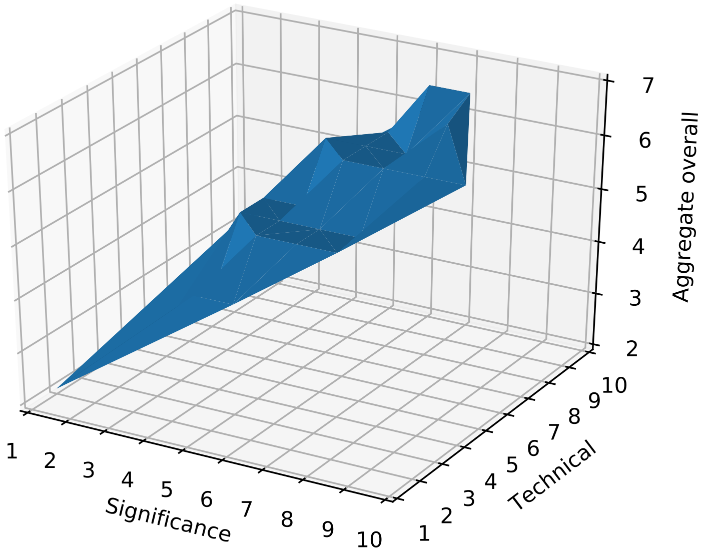
Solution: "Learn a mapping"

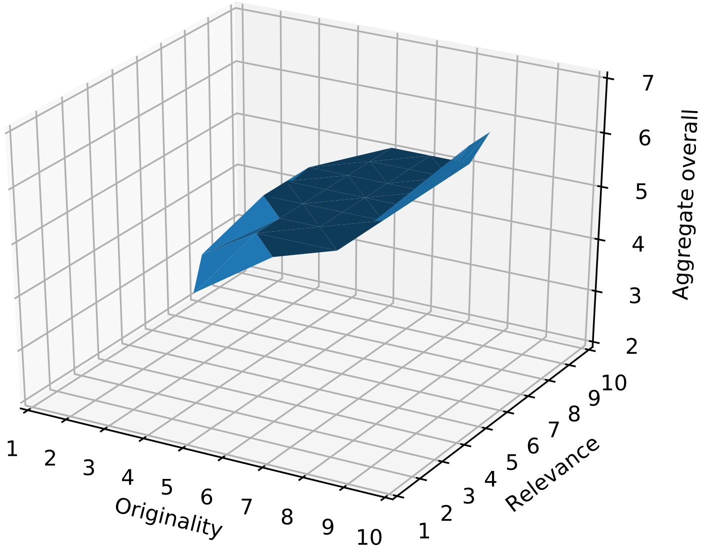
- Obtain (criteria scores, overall score) for every review
- Learn a mapping from criteria scores to overall scores
  - *Social choice theory: Use L(1,1) loss*
- For every review, apply learnt mapping to criteria scores to obtain a new overall score

[Noothigattu et al. 2018]

IJCAI 2017







- Writing and Relevance: Really bad - significant downside, really good - appreciated, in between - irrelevant.
- Technical quality and Significance: high influence; the influence is approximately linear.
- Originality: moderate influence.

[Noothigattu et al. 2018]

Miscalibration

Miscalibration in ratings

This is a moderately decent paper.
8/10

This is a moderately decent paper.
4/10.

"the existence of disparate categories of reviewers creates the potential for **unfair treatment of authors**. Those whose papers are sent by chance to assassins/demoters are at an unfair disadvantage, while zealots/pushovers give authors an unfair advantage."

[Siegelman 1991]

Editor's Page

Stanley S. Siegelman, MD

Assassins and Zealots: Variations in Peer Review

[Mitliagkas et al. 2011, Ammar et al. 2012, Freund et al. 2003, and many others]

Two approaches in the literature

1 **Assume simplified (affine) models for calibration**

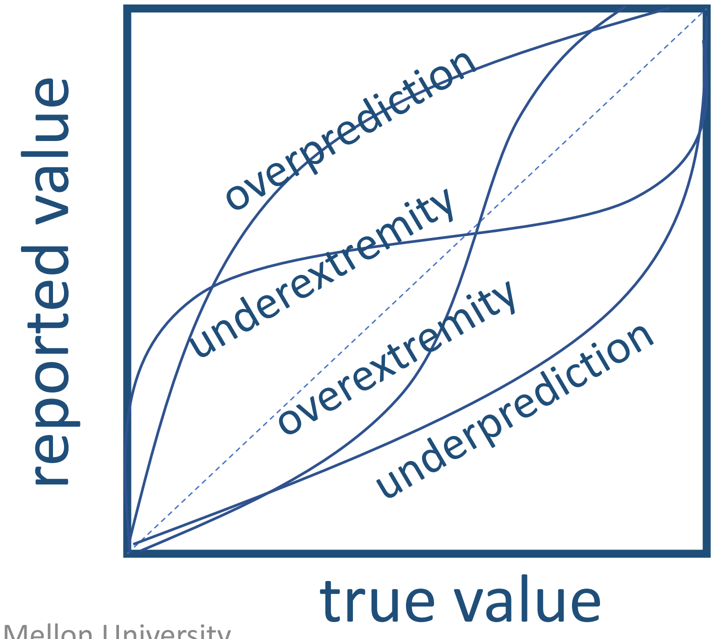
[Paul 1981, Flach et al. 2010, Roos et al. 2011, Baba et al. 2013, Ge et al. 2013, Mackay et al. 2017]

- Did not work well [NeurIPS 2016 program chairs; personal communication]
- *"We experimented with reviewer normalization and generally found it significantly harmful."* [Langford (ICML 2012 program co-chair)]

Miscalibration is quite complex:

[Brenner et al. 2005]

Small sample sizes per reviewer.



Two approaches in the literature

2 **Use rankings** [Rokeach 1968, Freund et al. 2003, Harzing et al. 2009, Mitliagkas et al. 2011, Ammar et al. 2012, Negahban et al. 2012]

- Use rankings induced by ratings or directly collect rankings
- Downside: lose useful rating information [Wang et al. 2018]
- Use rankings and ratings together [Shah et al. 2018, Pearce et al. 2023, Liu et al. 2023]

Outline

- Peer-review policies
- Seen such a review?
- **Reviewer incentives**
- Objectives of peer review
- Epilogue

Benign

Malicious

Benign

- **Quantity of reviews**
- **Quality of reviews**

Malicious

Freerider problem:

Researchers submitting papers but not contributing to reviewing

publons

- Verified record of researchers' reviewing
- Can be included in CVs
- Reviewers doing most reviews also incentivized via badges and awards
- Concerns: reviewers chase points by delivering superficial or poor reviews [Silva et al. 2017]
- Study [Pomponi et al. 2019]
    - Top-tier researchers scarcely seen on leaderboards
    - Top 250 reviewers carried out an average of over 180 reviews annually, but hardly wrote papers themselves

ML/AI venues: Authors must also review

We may incentivize number of reviews

I’m busy counting these bananas.
I can spend only a few minutes on the review.

How to incentivize high-quality reviews?

How to incentivize high-quality reviews?

Theory

Xiao et al. 2014
Xiao et al. 2018
Kong et al. 2018
Srinivasan et al. 2021
Ugarov 2023
Lee 2023

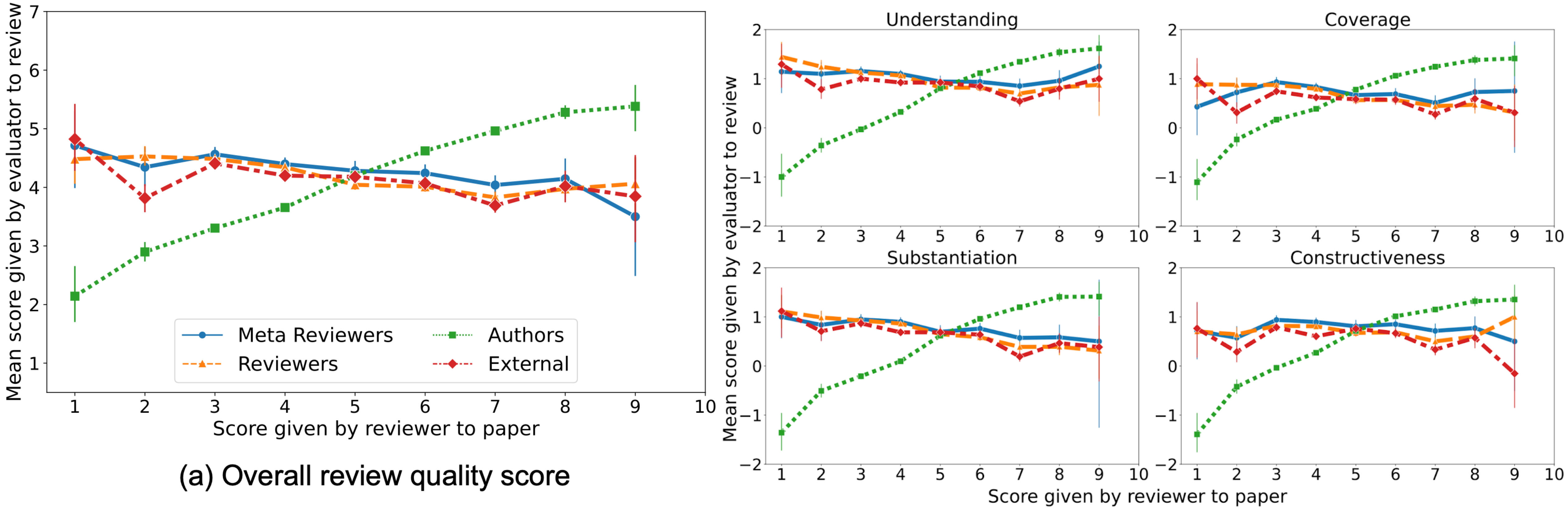
Practice

Reviewer awards
Blacklists

Rely on evaluation of the quality of each review

Can one reliably evaluate the quality of reviews?

NEURAL INFORMATION PROCESSING SYSTEMS 2022 Experiment on Reviewing Reviews



(a) Overall review quality score

(b) Criteria scores

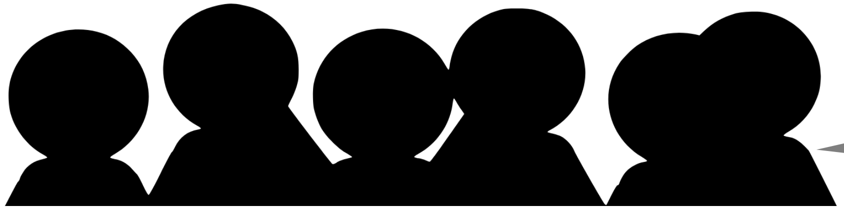
Mann-Whitney U test, controlling for various factors (P<0.0001)

Authors are biased by positivity of the reviews

[See Weber et al., 2002; Van Rooyen et al. 1999; Papagiannaki, 2007; Khosla, 2013; Kerzendorf et al. 2020 for more evidence; Wang et al. 2021 for some work on debiasing]

NEURAL INFORMATION PROCESSING SYSTEMS

2022 Experiment on Reviewing Reviews



Authors know their papers best, so ask authors to evaluate reviews

Or ask other reviewers or meta-reviewers or other experts

NEURAL INFORMATION PROCESSING SYSTEMS **Best Reviewers**

Randomized Controlled Trial: Original review …

Summary: [freeform text]

Strengths And Weaknesses: [freeform text]

Questions for authors: [freeform text]

Ethics Flag: No

Soundness: 2 Fair

Presentation: 4 Excellent

Contribution: 3 Good

Rating: 7: Accept: Technically solid paper, with high impact on at least one sub-area, or…

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but…

…made longer without useful information

Summary: [freeform text]

<Replicate abstract>

Strengths And Weaknesses: [freeform text]

<Replicate>

Questions for authors: [freeform text]

Let me briefly summarize the paper and its contributions. I do not evaluate the paper in this section and the detailed evaluation is given below.

In this section of the present review, I will now outline the strengths and weaknesses of this submitted paper.

Here are some questions I have for authors. I would like to see the response to these questions in the rebuttal.

Overall, in my opinion, <replicate everything from dropdown options>

Ethics Flag: No

Soundness: 2

Presentation: 4 Excellent

Contribution: 3 Good

Rating: 7: Accept: Technically solid paper, with high impact on at least one sub-area, or…

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but…

RCT: Review length bias

Original review

Uselessly elongated review

Mean score: 3.73 4.29

Criteria	**P-value (Mann-Whitney U test)**	**Difference in mean scores**
Overall score	**< 0.0001**	0.56 (7-pt scale)
Understanding	**0.04**	0.25 (5-pt scale)
Coverage	**<0.0001**	0.83 (5-pt scale)
Substantiation	**0.001**	0.31 (5-pt scale)
Constructiveness	**0.001**	0.37 (5-pt scale)

NEURAL INFORMATION PROCESSING SYSTEMS 2022 Experiment on Reviewing Reviews

- **Amount of inter-evaluator inconsistency, miscalibration, subjectivity at least as high as in reviews of papers**

- Reviewing reviews has similar issues as reviewing papers

- How to incentivize quality reviews?

Benign

Malicious

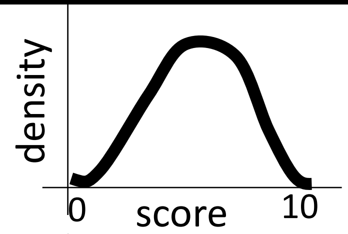
- **Lone wolf**
- **Collusions**

Lone wolf

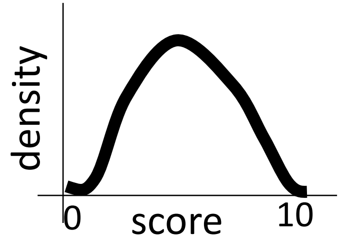
Rejecting competing papers will increase chances of my own paper getting accepted! Ha ha ha ha!

Randomized control trial

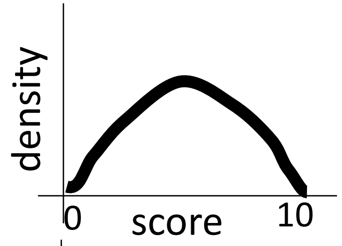
Score ≥ 7 → Accept; Review different conference from your submission



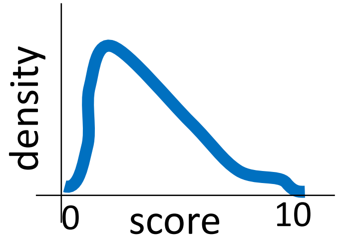
Score ≥ 7 → Accept; Review same conference as your submission



Accept top 20%; Review different conference from your submission



Accept top 20%; Review same conference as your submission



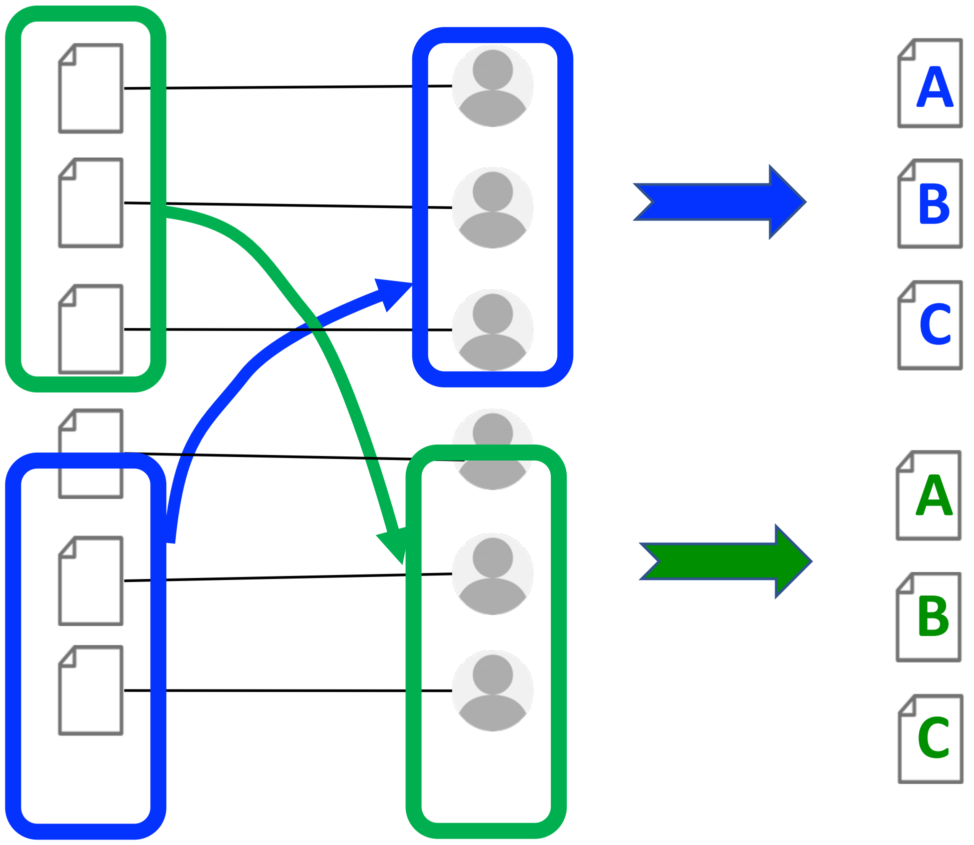
"substantial amount of gaming of the review system is taking place…
competition incentivizes reviewers to behave strategically…
the number of [strategic] reviews increases over time"

[Balietti et al., 2016]   Also see [Anderson et al. 2007, Langford 2008 (blog), Akst 2010, Thurner and Hanel 2011]

How to ensure that no reviewer can influence decision of their own paper?

Partitioning method

Authorship graph:



[Alon et al. 2011]

Work in progress

- Homogeneous expertise such as peer grading: [Alon et al. 2011, Holzman et al. 2013, Bousquet et al. 2014, Fischer et al. 2015, Kurokawa et al. 2015, Kahng et al. 2017; see also Aziz et al. 2019, Mattei et al. 2020]
- Heterogeneous expertise as in peer review: [Xu et al. 2018, Dhull et al. 2022]
- Statistical test to detect such behavior [Stelmakh et al. 2021]
- Dataset from a controlled experiment [Stelmakh et al. 2021]

Collusions

Why don’t you try to get assigned my paper and give it a positive review. In return, I’ll accept your grant proposal.

Sounds like a plan!

T. N. Vijaykumar May 12, 2020 · 5 min read

Potential Organized Fraud in ACM/IEEE Computer Architecture Conferences

"investigators found that a group of PC members and authors colluded to bid and push for each other's papers. They give high scores to the papers. Our process is not set up to combat such collusion."

"There is a chat group of a few dozen authors who in subsets work on common topics and carefully ensure not to co-author any papers with each other so as to keep out of each other's conflict lists (to the extent that even if there is collaboration they voluntarily give up authorship on one paper to prevent conflicts on many future papers)."

Such collusions also uncovered in conferences in other research areas and in grant reviews
[Lauer 2020, Littman 2021]

[https://medium.com/@tnvijayk/potential-organized-fraud-in-acm-ieee-computer-architecture-conferences-ccd61169370d]

Defense	Attack that breaks defense
1. Conflicts of interest 	• Colluders may not be collaborators/colleagues • Colluders skirt conflicts-of-interest detectors [Vijaykumar 2020] 
2. Detect or Remove Rings [Guo et al. 2018, Boehmer et al. 2021, Leyton-Brown et al. 2022]	• A reviewer may target an author's paper, and author may offer quid pro quo elsewhere
3. Bidding is easily gameable [Jecmen et al. 2020, Wu et al. 2021] So disable outlier bids [Wu et al. 2021] Dataset from controlled experiment [Jecmen et al. 2022]	• Bids of honest reviewers hardly influence the papers assigned to them [Jecmen et al. 2022]. Can't correct errors in text similarities; disincentivizes bidding altogether. • Attacks on text-matching [Markwood et al. 2017; Tran and Jaiswal 2019; Eisenhofer et al. 2023] • Other aspects of automated assignment systems, like subject area choices can be gamed [Ailamaki et al. 2019] • Colluding reviewers may already have expertise for that paper, and may be assigned even without bids [Vijaykumar 2020]
4. Geographic restrictions [Leyton-Brown et al. 2022]	• May collude across geographies (or if a colluder moves countries)
5. Randomized assignment [Jecmen et al. 2020]	• Establish collusions *after* papers are assigned

Quantification of tradeooffs: Jecmen et al. 2022

Outline

- Peer-review policies
- Seen such a review?
- Reviewer incentives
- **Objectives of peer review**
- Epilogue

Objectives of Peer Review

Ensure rigor of published research

Filter to select more interesting or better research

Additional objectives: feedback to authors, improve the research, learning experience for reviewers

Objectives of Peer Review

Ensure rigor of published research

Filter to select more interesting or better research

In journals outside computer science

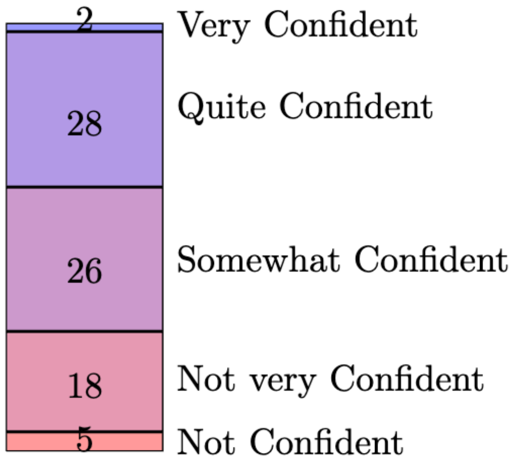
- Papers with major errors deliberately inserted
- Can reviewers spot these errors?

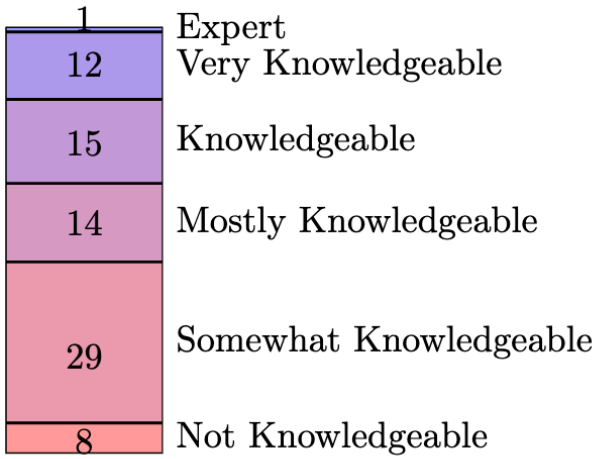
Study	#Errors inserted	#Reviews	%Errors detected on average
Baxt et al. 1998	10	203	34%
Godlee et al. 1998	8	221	25%
Schroter et al. 2004	9	1380	31%
Schroter et al. 2008*	9	1390	31%
Emerson et al. 2010	5	210	8% and 17%

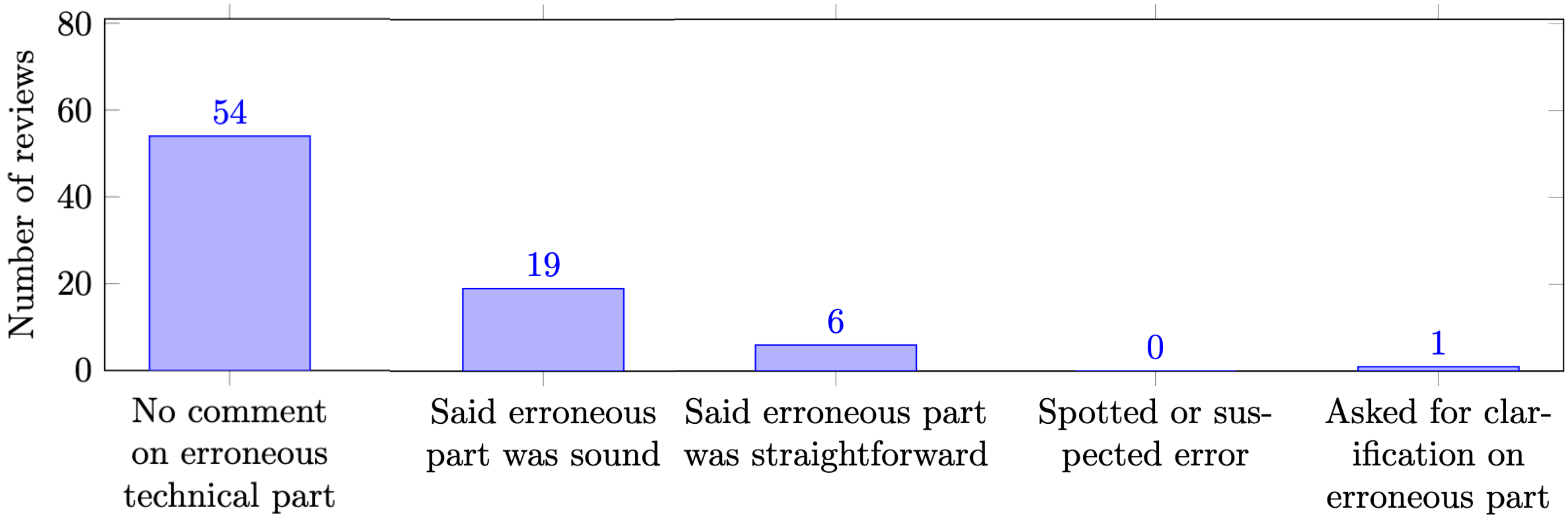
*Further analysis: >90% reviewers caught at least one error [see Shah 2023]

Experiment in a major AI/ML conference

- Three variants of a paper: Each variant had one major error in a claimed key contribution *(convexity of estimator; statistical identifiability; choosing hyperparameters on test data)*
- Error in main text
- 79 reviews
- Caveat: Generalizability



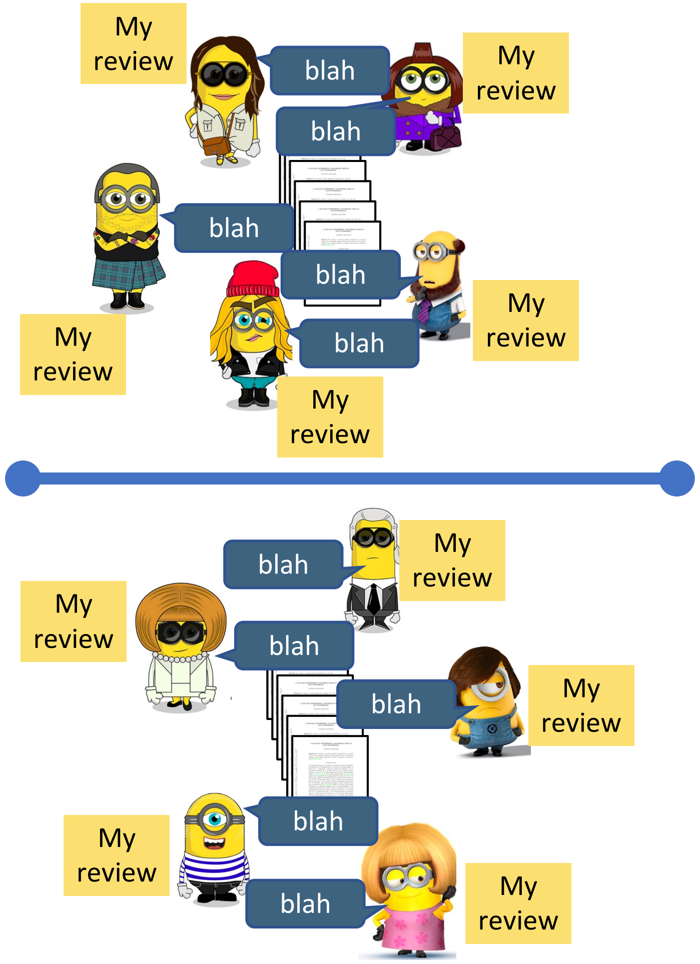


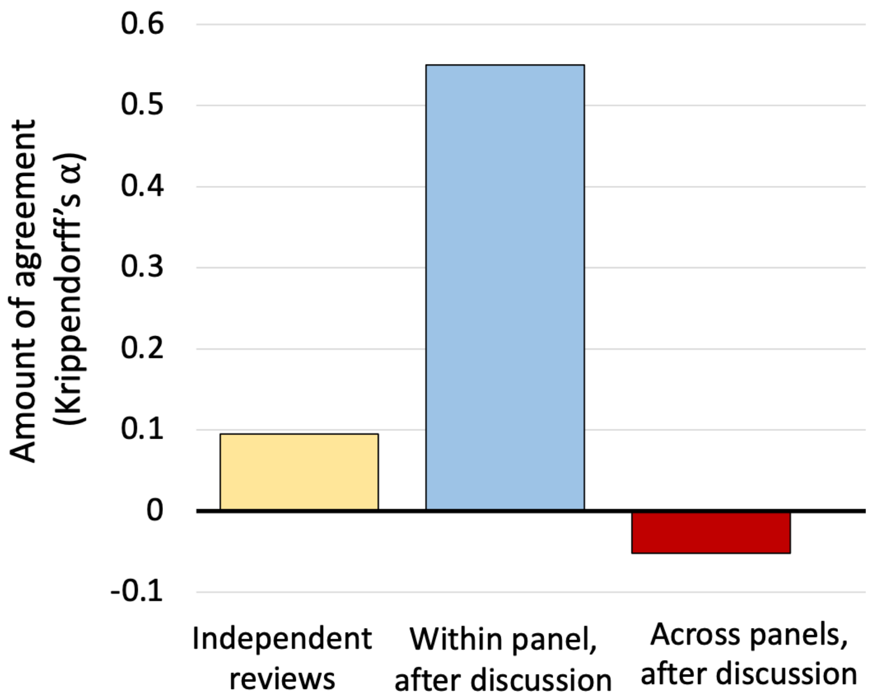


Objectives of Peer Review

Ensure rigor of published research

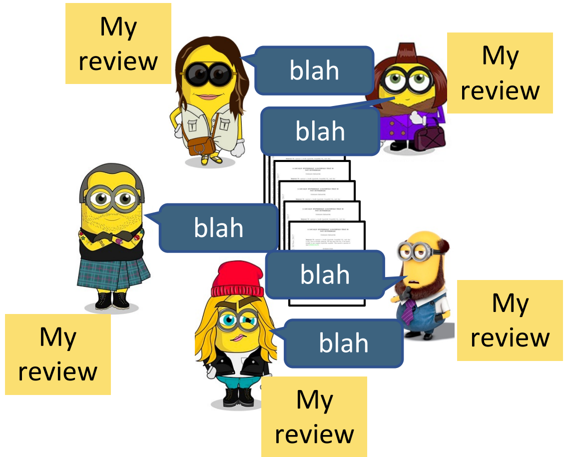
Filter to select more interesting or better research





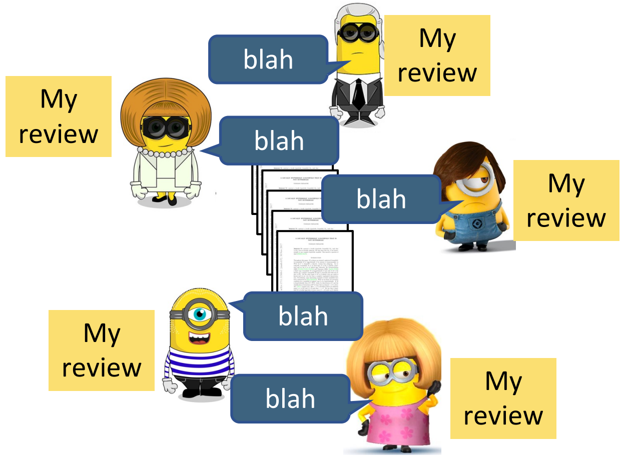
[Obrecht et al. 2007; Fogelholm et al. 2012; Pier et al. 2017]

NEURAL INFORMATION PROCESSING SYSTEMS 2014 Consistency Experiment [Cortes et al. 2014]



- 26% papers had inconsistent outcomes
- *Another interpretation:* 57% papers accepted by one committee were rejected by the other (perfect would be 0%, random 77%) [Price 2014]

NEURAL INFORMATION PROCESSING SYSTEMS 2021 Consistency Experiment [Beygelzimer et al. 2021]



- 23% papers had inconsistent outcomes (perfect would be 0%, random 35%)
- More than half of all spotlights recommended by one committee were rejected by the other.

Peer review vs. citations

- Reviewer scores uncorrelated with citations or downloads for accepted papers [Ragone et al. 2013, Connolly et al. 2014]
- NeurIPS 2014: no correlation between accepted papers' citations and scores; weak correlation for rejected papers [Cortes and Lawrence 2021]
- Review scores of perceived impact uncorrelated with citations, but correlated with social media impressions [Eysenbach 2022]
- When asked to forecast future citations, evaluators unsuccessful [Schroter et al. 2022]

NEURAL INFORMATION PROCESSING SYSTEMS

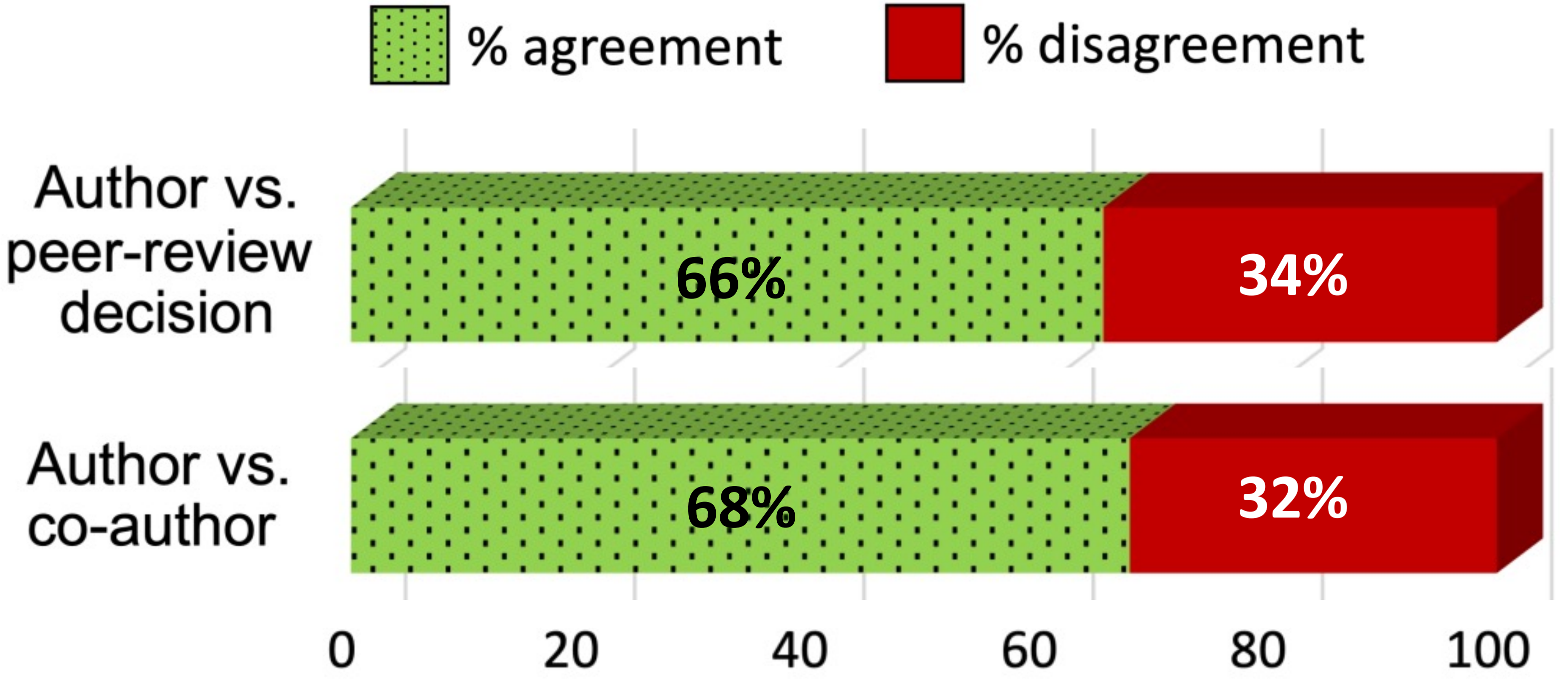
2021 Experiment on Author Perceptions

- Highly-selective venues aim to select the "best" papers
- Is there an "objective" ranking of papers?
- Disagreements between reviewers: but reviewers may be lazy etc.
- Maybe authors know "objective" ranking of their own papers
  - Independently, [Su 2022] proposed asking authors to give a ranking of their papers ("you are the best reviewer of your own papers") which will determine their review outcomes
- If there is such an objective ranking, co-authors should generally agree on it…

[Rastogi et al. 2023]

2021 Experiment on Author Perceptions

Rank your submissions in terms of your own perception of their scientific contributions to the NeurIPS community, if published in their current form.



[Rastogi et al. 2023]

Outline

- Peer-review policies
- Seen such a review?
- Reviewer incentives
- Objectives of peer review
- **Epilogue: My opinion**


BarbieLand
This Way


My paper got accepted to NeurIPS!!

Do you guys ever think about review quality?

Truly remarkable achievement to get into the most prestigious, highly selective ML conference!

It means your paper was rigorously reviewed and found to be technically sound and one of the most impactful!

Wow, you will now get better grad school admits, better jobs,...

You have to go to the real world. You can go back to your regular resubmissions-rebuttals, and forget any of this ever happened. Or you can know the truth about these conferences.

Credits: “Barbie” movie

Credits: “Piled Higher and Deeper” by Jorge Cham

We should radically rethink NeurIPS reviewing

- **Focus on evaluating (only) correctness** [PLOS ONE, TMLR]
  + Less stress ☺; emphasis on rigor
  - Publication of high volume of incremental work
  ? Space constraints for conference presentations?

- **Signed reviews: Reviewers' names revealed** [f1000research, eLife, JSys, Goodlee et al. 1998, van Rooyen et al. 1999, 2010, Walsh et al. 2000]
  + Incentives for quality review; mitigate collusions
  - Possible author retaliation; junior reviewers more hesitant to review
  ? May be ok if focus is on correctness?

- **Post-publication review:** Publish everything with/without reviews; market forces of online commenting and citations take over [pubpeer.com, openreview.net, Kriegeskorte 2012, Bordignon 2020]
  + Less burden on peer review
  - No author anonymity ⇒ potential biases

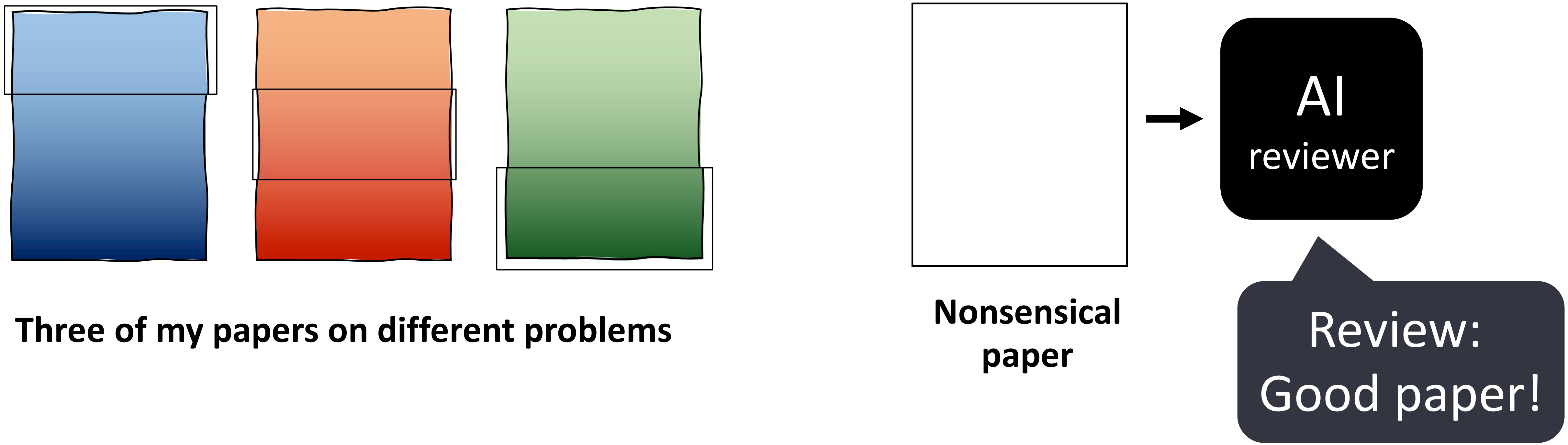
What about fully automating reviewing of manuscripts?

- Many **computational tools** already used
  - Most prominently for reviewer assignment
  - And others (bidding, subjectivity, dishonesty, etc.) [see survey bit.ly/SurveyPeerReview]

- **"AI Reviewer"**
  - Pre-ChatGPT [Huang 2018, Wang et al. 2020, Yuan et al. 2021]
  - Post-ChatGPT [Liang et al. 2023 and other unpublished work]
- **Evaluation**
  - Subjective (human) evaluation: biases in evaluating review quality
  - Objective evaluation?

Focus on peer review's objectives

A basic "Chimera" test



Three of my papers on different problems

Nonsensical paper

Focus on peer review's objectives

Dataset of carefully constructed short papers

- **Correctness objective**
  - Deliberately inserted errors
  - GPT-4 detects inserted errors in 50% constructed papers, including in conceptual arguments and mathematical proofs
  - Bard and open source models exhibit poor performance
  - Don't prompt "write a review," but instead be specific "find errors"
- **Selecting "better" papers objective**
  - Pairs of abstracts such that one objectively superior to the other
  - Slightly tweaked some of them in terms of language etc.
  - Performance is poor, fooled/gamed easily

[Liu et al. May 2023]

Conclusions

- **Scientific reviewing from a scientific lens**
  - *Designing peer review systems:* Think about objectives, evidence-based policies
  - *Discussions of reviews:* Does the review exhibit an established problem?
  - *Ideas to improve peer review:* Literature may shed some light on it
- **Many computational opportunities and challenges**

bit.ly/PeerReviewOverview