

Peer review should constrain evaluative authority

Hanrui Wang^{1*} · Timo Spinde^{1,2} · Isabella Haberer^{2,3}
Chun-Shien Lu⁴ · Isao Echizen^{1,2}

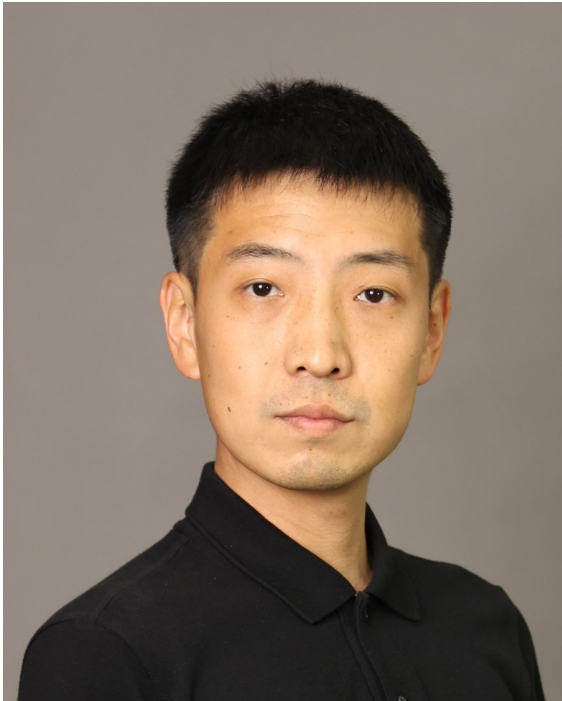
Speaker: Hanrui Wang

¹ National Institute of Informatics, Japan ² Media Bias Group, Japan

³ Georg-August-University Göttingen, Germany ⁴ Academia Sinica, Taiwan



Hanrui Wang



Assistant Professor

Echizen Laboratory

National Institute of
Informatics (NII)

Tokyo, Japan

RESEARCH HIGHLIGHTS · AI SECURITY & PRIVACY

01 Security & Privacy of Emerging AI Systems

AI Agents · Computer-Use Systems · World Models · 4D Models

Security and privacy risks in autonomous interaction
and environmental understanding.

02 Adversarial Machine Learning

**Adversarial Robustness · Attack Transferability
Robustness Evaluation · Model Inversion**

Understanding, evaluating and exposing security and privacy risks.

03 AI Evaluation & Governance ← **This project**

Scientific Peer Review · AI-Assisted Evaluation · AI Governance

Reliable and accountable AI in consequential evaluation and decisions.

Ph.D., Computer Science · Monash University | Former IT industry director

Five questions for peer-review reform

A

Why is peer review under pressure?

Not just more submissions.

B

What can't we fix?

No ground truth for a “good” or “right” review.

C

Why aren't existing reforms enough?

Seeking “good” or “right” reviews without ground truth.

D

What do we propose to change?

No evaluator should have unchecked authority.

E

How far are we from the goal?

A position paper is only a starting point.

A

Peer review has changed.

The decisions are still consequential.
The pressure is on a different scale.



Peer review shapes science and academic opportunities

Peer review



Decides whether a submission is accepted at the venue.

Publication



Builds a visible record of research contributions.

Academic opportunities



Recognition, careers, funding and academic qualifications.

Peer review is under growing pressure

Unmatched scale



Submissions are increasing, but there are not enough qualified reviewers.

Weak reviewing incentives



Publishing advances careers. Careful reviewing earns much less recognition.

Competing interests



Researchers judge papers while competing for publication, funding, and career opportunities.

B

“Peer review: a flawed process
at the heart of science
and journals”

“the practice of peer review is
**based on faith in its effects,
rather than on facts**”

— Richard Smith, 2006



Richard Smith

Editor-in-Chief, BMJ,
1991–2004

Chief Executive
BMJ Publishing Group,
1991–2004

Many review criteria lack reliable ground truth

Review criterion	Relatively reliable Missing details can be checked directly	Limited Evidence allows only partial checks	Unavailable No reliable way to tell which judgment is right
Reporting completeness	●		
Technical correctness		●	
Methods and experimental evidence		●	
Coverage of prior work		●	
Reproducibility		●	
Ethics		●	
Novelty and originality			●
Importance and expected impact			●
Fit with the venue			●
Clarity and presentation			●

The same paper meets different standards

“Strong enough” depends partly on who is reviewing.

Sources of reviewer-side subjectivity



Review criterion	Topic familiarity	Technical expertise	Scholarly pursuit	Personal preferences	Reviewer effort	Verification resources
Reporting completeness					●	
Technical correctness	●	●			●	●
Methods and evidence	●	●	●	●		
Coverage of prior work	●	●	●	●	●	
Reproducibility	●	●	●		●	●
Ethics	●	●	●	●		
Novelty and originality	●	●	●	●		
Importance and impact	●	●	●	●		
Fit with the venue	●	●	●	●		
Clarity and presentation	●		●	●		

● Illustrative influences on how each criterion is assessed

Disagreement can persist even when both reviewers act in good faith.

We see the review. We may not see the motive.

The same criticism can reflect honest judgment, low effort, or self-interest.

How incentives can shape reviews

Incentive-coupled evaluation	Scarce rewards	Conflicts of interest	Weak incentives for careful review
Strategic distortion Inflate or deflate judgments	●	●	
Low-effort / AI-substituted review Limited independent scrutiny	●		●
Cautious non-commitment Hedge or stay borderline		●	

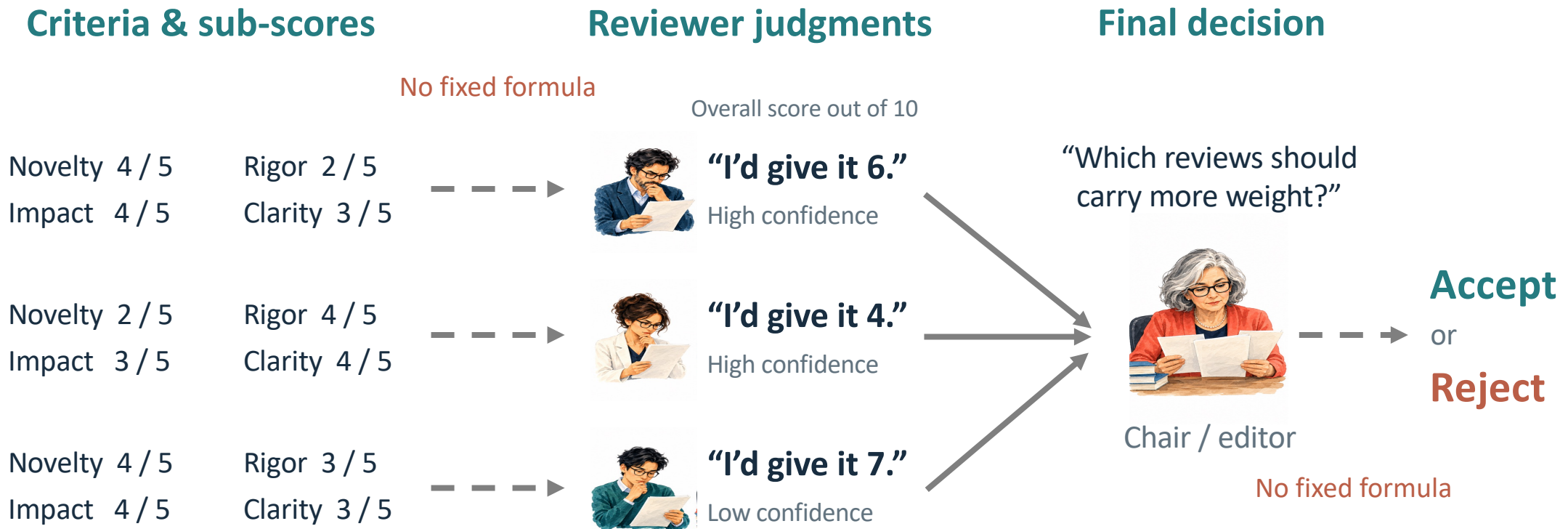
● Illustrative links, not evidence of any individual reviewer’s motive



Peer review relies on trust we cannot fully verify.

No fixed formula links criteria to final decisions

Criterion scores do not uniquely determine a reviewer's recommendation.

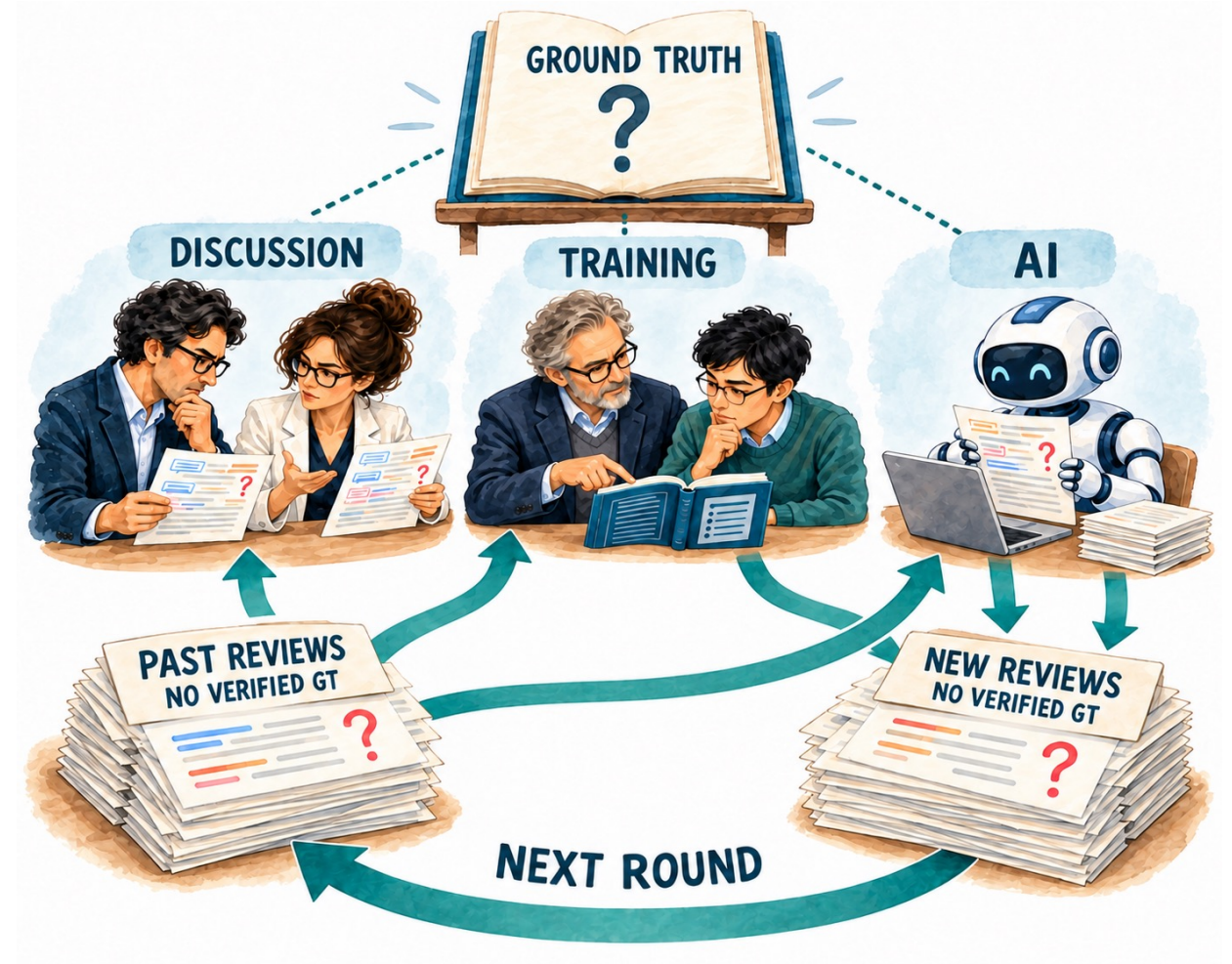


Reviewer recommendations inform final decisions, but neither is determined by a fixed formula.

C

No ground truth.
No guarantee of
“good” or “right”.

Past reviews are not verified ground truth.
Learning from them does not create it.



More reviews do not ensure timely correction

Today's standard system combines multiple reviews, author rebuttal, reviewer discussion, and chair/editor judgment.

Multiple reviews + scores

Author rebuttal

Reviewer discussion

Chair / editor decision



Influence can begin



Rebuttal begins



The same uncertainty remains

More reviewers still face missing ground truth, differing standards, and competing incentives.

One review can steer the outcome

Early judgments can anchor deliberation. Under fierce competition, one negative review may outweigh several positive ones.

More reviews do not eliminate disagreement

No study establishes how many reviews are enough to prevent a minority of judgments from exerting consequential influence.

A review may shape the decision before its claims are challenged.

Training cannot eliminate subjectivity or ensure compliance

What can improve



Clearer standards and training



Transparency, accountability, and incentives



Better matching and a larger reviewer pool

What remains

Standards remain subjective

Training cannot eliminate differences in expertise, priorities, and preferences.

Compliance remains uncertain

Knowing the rules does not ensure following them when interests, time, and effort conflict.

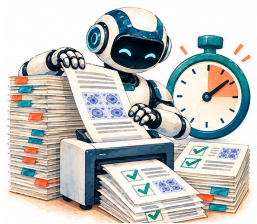
Expertise cannot scale on demand

Developing capable reviewers takes longer than submissions take to multiply.

Even a highly capable AI needs a reliable learning target we do not have.

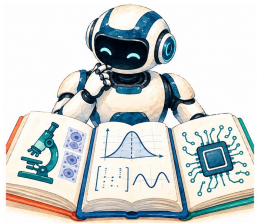
Suppose AI makes no mistakes.

What teaches AI to be a “good” reviewer?



Less workload

Fast, scalable reviewing



Broader expertise

Knowledge across research domains



More consistent treatment

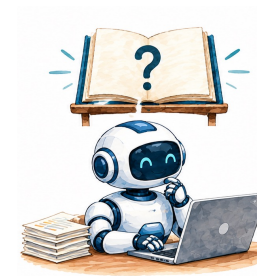
Shared standards,
no personal COI

Past reviews



No verified
ground truth

Training & validation



No reliable target for
“good” or “right”

New review data



More reviews,
harder to verify

Next training round

Inconsistent, low-effort, or strategic reviews can further contaminate the learning signal.

D

Peer review should
constrain
evaluative authority.

Consequential judgments must face scrutiny
before they gain influence over decisions.



Authority-constrained peer review

Before a review can shape the decision, its consequential claims must withstand independent scrutiny.

01

Separate critique from authority

Writing a critique does not grant unilateral power to turn it into a consequential score.

02

Require an independent check

Inspect support and interpretation before consequential influence.

03

Allow challenge and correction

Clarify, narrow, downgrade or reverse an unsupported judgment.

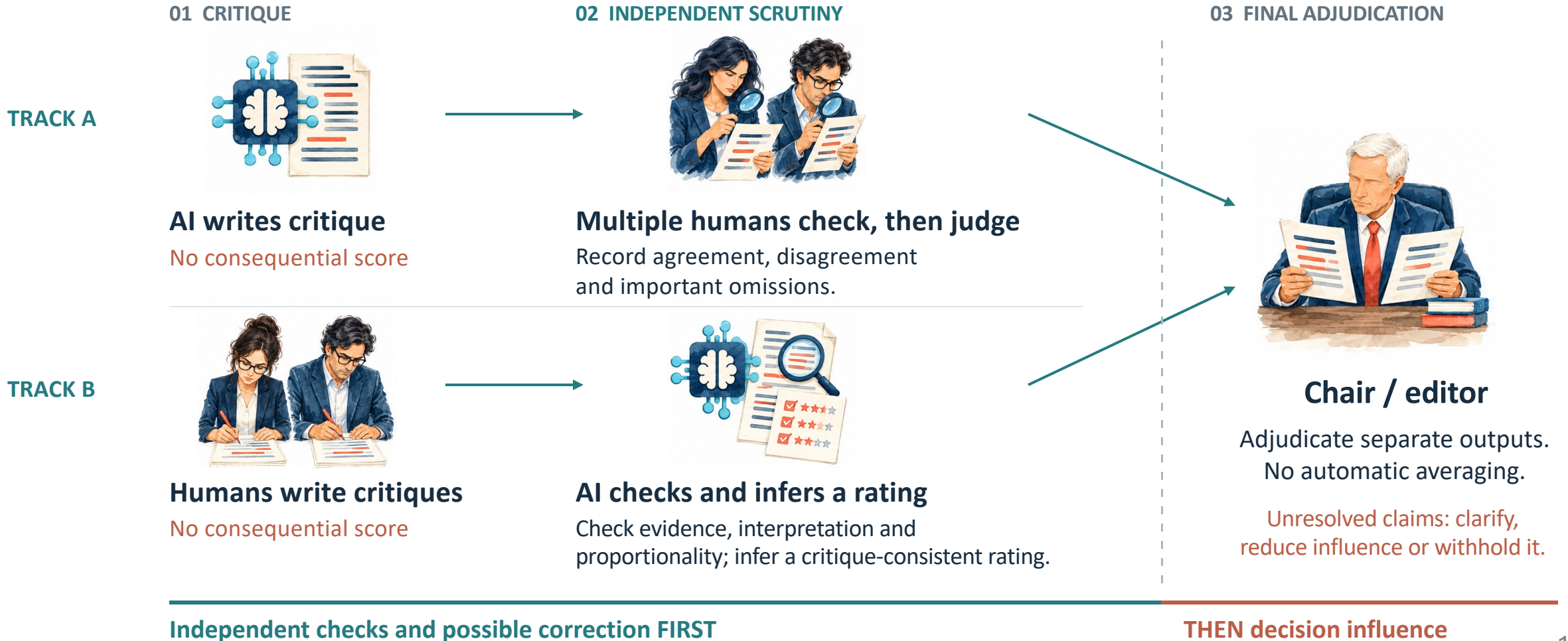
04

Keep an accountable record

Preserve consequential claims and corrections in the record.

Human–AI mutual supervision

Neither humans nor AI may unilaterally turn their own critique into a consequential score.



The workflow produced usable counter-signals

One paper · four new human supervisors · four existing reviews · one AI · one chair

Track A: AI critique → Human scoring

Human supervisor	#1 (Researcher in AI Safety)	#2 (Researcher in AI Safety)	#3 (Researcher in Machine Learning)	#4 (Ph.D. Student in LLM Security)
Resulting rating	5/10	5/10	5/10	6/10

Track B: Human critiques → AI scoring and scrutiny

Human review	Original rating	AI-inferred rating	AI scrutiny flags
#5	3/10	6/10	Superficial · inaccurate · misguided
#6	6/10	8/10	Superficial
#7	6/10	6/10	Inaccurate · superficial
#8	8/10	8/10	Inaccurate (minor)

Final decision: Reject → Reject.

Supervision felt easier—but one case is not validation

Reported experience	#1	#2	#3	#4	Chair
Difficulty vs usual role	Easier	Easier	Easier	Easier	Neither
Perceived output quality	Improve	Reduce	Improve	Improve	Neither
Consulted manuscript	No	Yes	Yes	No	No

4 / 4

Supervisors found it easier

3 / 4

Perceived better output

Chair workload remains uncertain

More signals to assess, but unsupported or conflicting judgments may be easier to identify.

E

We are still far from the goal.

The next step is to turn a workable idea into a tested institutional practice.



Technical objectives for authority-constrained review

Five objectives open to technical collaboration.

Objective	Technical task	Research area
I · Operationalize and validate structured scrutiny	Develop concrete checks for evidence, interpretation and proportionality; test their reliability and limits.	Evaluation methodology
II · Build datasets for evaluating scrutiny	Include verifiable review failures while preserving disagreement, uncertainty and provenance.	Datasets & benchmarks
III · Develop AI supervision methods	Infer critique-consistent ratings and flag potential review failures using both reviews and submissions.	NLP LLM evaluation
IV · Improve AI critique reliability and study its influence	Improve factual grounding, reduce specific biases, and test how critique presentation shapes human judgments.	LLM alignment HCI
V · Secure AI critique and scrutiny	Evaluate and defend against prompt injection, adversarial edits, data poisoning and implicit manipulation.	AI security Adversarial ML

Preparing a real-venue pilot

Start with AI scrutiny of existing human reviews, then evaluate the full dual-track design.

PREPARATION

01	02	03	04
Build and maintain the project website	Connect research teams	Propose workshops	Contact venue organisers
Share the proposal, resources and ways to get involved.	Engage peer-review, ethics and Social AI researchers.	Bring researchers and venue organisers together to shape a pilot.	Find a conference or journal partner and agree the pilot scope.

WHO WE HOPE TO WORK WITH

Venue organisers	Research collaborators
Co-design the workflow, review policies and a feasible pilot in a real venue.	Develop and evaluate scrutiny methods, human–AI interaction and safeguards.

We may not be able to guarantee the right review,
but we can decide how much unchecked authority any review gets.

Hanrui Wang

hanrui_wang@nii.ac.jp

POSTER LOCATION

Sydney Poster Session 1, Hall 1-4
Tue, Dec 8, 2026 • 10:00 AM – 1:00 PM AEDT

Project page



Personal website

