A WORKING PAPER

The Measure of Good AI Is Better Humans, Not Fewer Humans

A Human-Centered Framework for AI Governance

Priscilla Osaro
Founder, Inside the Bid Room
Bid & Proposal Strategist | Africa Bid Community

Priscilla Osaro. (2026) The Measure of Good AI Is Better Humans, Not Fewer Humans: A Human-Centered Framework for AI Governance. Working Paper.

Abstract

Organizations are increasingly using artificial intelligence to review proposals, grant applications, procurement submissions, hiring decisions, and other high volume evaluations. These systems are usually judged by how much time they save, how many cases they process, and how efficiently they reduce workload. This paper argues that those measures overlook a more important question. After prolonged use, are the people making the final decisions becoming better or worse at exercising the judgment their roles require?

Drawing on evidence from procurement, hiring, healthcare, cognitive psychology, and AI governance, the paper examines how AI-assisted evaluation can gradually shift responsibility away from human judgment and toward automated triage. It explores recurring patterns such as the gap between compliant and compelling proposals, the structural bias against organizations with limited but high quality past performance, and the gradual decline of evaluative capability when efficiency becomes the primary measure of success.

This paper introduces the **Better Humans Standard**, a human centered framework for evaluating AI systems by their long term effect on the people who use them. Instead of asking whether AI makes evaluation faster or more efficient, the framework asks if sustained use helps people become better at exercising judgment, handling uncertainty, and making accountable decisions.

The framework is built around four design principles. It encourages AI systems to surface uncertainty instead of hiding it. It preserves friction where friction strengthens judgment. It makes evaluator reasoning visible and open to review. It also measures if people become better decision makers over time rather than simply measuring how quickly work is completed.

The paper concludes by considering what this framework means for AI governance, public policy, and institutional practice. Its central argument is straightforward. AI should not be judged only by what it automates. It should also be judged by what it does to human judgment. The measure of good AI is not how many people it replaces, but how many people it makes better at deciding what matters.

I. The Evaluator Who Stopped Reading

She has forty one proposals to score by Friday. She has eleven minutes for each.

The scoring rubric has not changed in decades. What arrives on her desk has. What was once a stack of paper submissions is now hundreds of digital files. She is now expected to use AI to review them. An AI-assisted triage system checks for compliance, flags missing certifications, and ranks proposals according to how closely they match the solicitation.

It works. Mostly.

It catches obvious disqualifiers. It saves her the first two rounds of reading. More importantly, it begins to shape how she works. Proposal after proposal, it teaches her to trust the shortlist.

By the third week, she notices something has changed. She has stopped reading the proposals the system ranks near the bottom. Not because she has decided they are weak. She has not read them closely enough to make that judgment. She skips them because the system has already suggested they are unlikely to matter.

Somewhere in the proposal ranked fourteenth out of forty-one is a small organization applying for its first federal contract. Its staffing plan is unconventional but well designed. Its past performance section is thin, not because the work was poor but because the organization is new. The system reads "thin." She reads the score.

This is not a story about a flawed tool or a careless evaluator. The tool is doing exactly what it was designed to do. It reduces noise, saves time, and standardizes the first review. The evaluator is experienced and under real pressure to process a growing volume of work. Nothing about this situation reflects negligence or incompetence.

Something quieter is happening.

Judgment is slowly being replaced by triage in a process whose value depends on someone taking the time to look closely.

This paper examines that shift. It asks what happens when AI changes not only how decisions are made, but also how people learn to make them. More importantly, it asks what it would take to design AI systems that strengthen human judgment instead of gradually replacing it.

Contributions of this Paper

This paper makes four contributions.

• It introduces the Better Humans Standard as a framework for evaluating AI by its long-term effect on human judgment rather than automation alone.

• It argues that evaluator capability is an overlooked dimension of AI governance.

• It proposes four design principles for designing systems that strengthen rather than diminish professional judgment.

• It suggests evaluator calibration as an outcome that organizations should measure alongside efficiency

II. What Gets Displaced When Judgment Becomes Triage

A. The Compliant and Compelling Gap

The clearest sign of what triage displaces is not a dramatic failure. It is a pattern so common that experienced evaluators stop noticing it. The difference between a proposal that is **compliant** and one that is **compelling**.

A compliant proposal satisfies the checklist. The required certifications are included. The format follows the solicitation. Every evaluation criterion is addressed. An experienced evaluator can usually identify these within minutes.

A compelling proposal does something a checklist cannot measure. It demonstrates that the organization genuinely understands the problem it has been asked to solve. Its strength lies not in ticking every box, but in the quality of its reasoning, judgment, and understanding.

These are not the same thing.

Under pressure to review large volumes of submissions, automated screening is naturally better at identifying compliance than recognizing genuine quality. Compliance follows patterns. Strong reasoning does not.

Recognizing insight requires close reading. It requires an evaluator to distinguish real understanding from language that simply sounds convincing. No screening tool can replace that kind of judgment because it depends on interpretation rather than pattern recognition.

A similar pattern has already been observed in academic peer review, which shares many characteristics with proposal evaluation. A 2025 study presented at the International Conference on Learning Representations found that AI-assisted reviews increased paper scores and acceptance rates, not because the research itself improved, but because the review process changed (Russo et al., 2025). The quality of the work remained the same. The evaluation did not.

That distinction lies at the heart of this paper.

A procurement example

Consider a USAID-funded health systems strengthening procurement evaluated in 2024. Forty percent of the technical score was allocated to organizational capacity, measured largely through previous USAID contracts and contract value.

Two organizations submitted proposals.

Bidder A was a large international consultancy with $15 million in previous USAID contracts, an extensive management structure, and global operations. It received a compliance score of 95 out of 100.

Bidder B was an African-led consortium with $2.8 million in successful USAID contracts, no disputes, an unconventional staffing model, and an innovative approach to community health worker training. It received a compliance score of 78 and was flagged as having limited organizational capacity.

The assessment was technically correct. Bidder B had a smaller track record.

The problem was the assumption that smaller meant weaker.

The real question was if the consortium could deliver this particular programme.

One organization had scale. The other had precision.

The scoring framework could not distinguish between them because it treated organizational size as a proxy for organizational capability.

An evaluator who reads closely may recognise that difference.

A process optimized for ranking submissions is much less likely to.

B. What Happens to Judgment Over Time

The deeper cost is not what happens to a single evaluation.

It is what happens to the evaluator.

Judgment is a skill. Like every professional skill, it improves through repeated practice. Evaluators who regularly read proposals closely become better at separating genuine capability from polished language. That ability develops gradually through experience.

Evidence suggests the opposite can also happen.

A multicentre randomized study found that physicians who routinely relied on AI support during colonoscopy identified significantly fewer adenomas once that support was removed. Detection rates fell from 28.4 percent to 22.4 percent (Bhatt et al., 2026). The technology had improved performance while it was present, but it had not strengthened the underlying skill. Once the technology disappeared, the decline became visible.

A broader review reached a similar conclusion across multiple medical specialties. Heavy reliance on decision support was associated with measurable declines in professional judgment, particularly among less experienced practitioners (Artificial Intelligence Review, 2025).

The same risk exists wherever evaluation becomes increasingly dependent on automated screening.

Experienced proposal professionals often describe a subtle change in how they read. Over time, they become less willing to spend time on submissions that appear weak at first glance. A proposal with an unusual structure or a missing section is no longer explored for underlying merit. It is simply passed over.

That habit matters.

Some of the strongest proposals are unconventional. Some of the weakest are perfectly formatted.

The time saved by screening is easy to measure.

The gradual loss of evaluative skill is not.

It becomes visible only when an unusual proposal depends on someone who still knows how to recognise quality beneath an imperfect presentation.

Research also suggests that human oversight is often less effective than organisations assume. A 2025 study at the University of Washington found that reviewers frequently accepted biased recommendations unless the error was obvious (Wilson et al., 2025).

Simply keeping a person in the process does not guarantee meaningful oversight.

A person who has learned to trust the ranking without questioning it is no longer exercising independent judgment. They are confirming a recommendation that has already shaped the decision.

C. Structural Bias in Past Performance Evaluation

Another problem is built into the evaluation process itself.

Most scoring rubrics reward organizations with larger contracts, bigger portfolios, and longer track records because those factors are easy to verify and score consistently.

The result is a structural bias against organizations that have delivered high-quality work but have not yet had the opportunity to build large contract histories.

This pattern is especially familiar in federal procurement. Many women-owned, disadvantaged, and service-disabled veteran-owned businesses enter the market with strong technical capability but relatively limited past performance. The challenge is well documented. The U.S. Government Accountability Office has repeatedly found that firms without established contract histories face a persistent disadvantage when trying to compete for larger awards (GAO-12-102R, 2011). Procurement practitioners often describe this as the "chicken-and-egg problem." Organizations need past performance to win contracts, but they need contracts to build past performance (Koprince, 2021).

Technology does not create this bias.

It inherits it.

When a screening model is trained to treat contract value, portfolio size, or organizational scale as indicators of capability, it simply learns the assumptions already built into the evaluation framework.

The result is consistency.

The same organizations are disadvantaged in the same way, proposal after proposal, at a scale no human evaluation team could reproduce on its own.

That is why the problem is larger than algorithmic bias.

The deeper issue is that technology can make an imperfect evaluation framework appear objective while applying its assumptions with greater speed and consistency than ever before.

III. The Measurement Problem

If the argument so far is correct, another problem becomes impossible to ignore.

Most organizations are measuring the wrong thing.

Procurement offices track time to award. Foundations measure applications processed per program officer. Donor-funded organizations report shorter review cycles and higher processing capacity.

These are useful operational measures.

They are not measures of judgment.

Most organizations adopting these technologies care deeply about making good decisions. The problem is not a lack of commitment. The problem is that decision quality is much harder to measure than operational efficiency.

Time saved can be reported immediately.

Better judgment cannot.

Sometimes the quality of a decision is only visible months or years later. A proposal rejected today may prove to have been the strongest solution. A highly ranked proposal may ultimately fail during implementation.

Because those outcomes are difficult to observe, institutions naturally optimize the measures they can see.

This is Goodhart's Law in practice.

When efficiency becomes the primary measure of success, efficiency becomes the primary objective.

Judgment quietly becomes secondary.

The challenge is not that judgment cannot be measured.

It usually is not.

Kahneman, Sibony, and Sunstein describe this problem through the concept of *noise*. In one study, judges reviewing identical asylum cases reached dramatically different conclusions simply because different judges happened to hear them. Approval rates ranged from 5 percent to 88 percent (Kahneman, Sibony, & Sunstein, 2021).

That variation is measurable.

Many organizations simply choose not to measure it.

This paper argues for a different approach.

Instead of evaluating technology only by how much work it automates, we should also evaluate what it does to the people who rely on it.

Imagine two evaluators using the same screening tool for a year.

One becomes better at recognizing uncertainty because the system encourages independent reasoning and reflection.

The other gradually stops questioning the recommendations placed in front of them because every difficult decision has already been simplified into a ranking.

Both evaluators become faster.

Only one becomes better.

That is the difference between measuring productivity and measuring judgment.

It is also the foundation of the Better Humans Standard.

IV. The Better Humans Standard

A Framework for Strengthening Human Judgment

The previous sections argue that most approaches to evaluating AI focus on efficiency while overlooking what matters most. They measure how quickly work is completed but rarely ask what repeated use does to the people making the decisions.

The Better Humans Standard begins with a different question.

Does this technology help people become better at exercising judgment over time?

That question shifts the focus from automation to capability. Speed and efficiency still matter, but they are no longer the primary measure of success. A system should also be judged by whether it helps people reason through uncertainty, recognise quality, and make decisions they can explain and defend.

If the measure of good AI is better humans, not fewer humans, then that principle has to appear in the design itself.

Table 1. Rethinking How We Measure Good AI

Conventional Measures	The Better Humans Standard
Time saved	Better judgment over time
Throughput	Better Calibration
Cost Reduction	Better Accountability
Automation	Stronger Human Capability
Faster Decisions	Better Decisions
Task completion	Stronger Professional Reasoning

The Better Humans Standard is built around four design principles.

Table 2: *The Better Humans Standard*

Principle	Core Question
Surface Uncertainty	Does the system show where it is unsure instead of hiding uncertainty?
Preserve Friction	Does the design strengthen judgment instead of removing every pause for reflection?
Make Reasoning Visible	Can people explain and review why a decision was made?
Design for Calibration	Does the system help evaluators become better over time?

Principle 1

Surface Uncertainty Instead of Hiding It

One of the quickest ways to weaken judgment is to present uncertainty as certainty.

A ranked list suggests that every proposal has been placed exactly where it belongs. In reality, many evaluations involve borderline cases, incomplete information, and competing strengths that cannot be reduced to a single score.

A better approach is to make uncertainty visible.

Instead of quietly assigning a ranking, the system identifies the submissions where confidence is low and returns those cases to the evaluator for closer review.

This changes where attention goes.

Instead of spending time confirming obvious decisions, evaluators spend time on the cases where their judgment is most valuable.

Research on human interaction with algorithms shows why this matters. People do not consistently trust automated recommendations, nor do they consistently reject them. They often move between overreliance and excessive scepticism depending on previous experience (Dietvorst, Simmons, & Massey, 2015).

Making uncertainty visible gives people a reason to think rather than simply react.

Example

Imagine a foundation reviewing 500 grant applications.

Rather than dividing every application into "recommended" and "not recommended," the review process could create three groups.

- Clear candidates that confidently meet the required standard.
- Ambiguous applications where evidence is mixed or confidence is low.
- Clear declines that do not meet essential requirements.

The evaluator's role changes.

Instead of reviewing every application equally, they focus their attention where judgment adds the greatest value.

A useful prompt might simply ask,

"What makes this application uncertain, and is that uncertainty caused by the proposal or by the evaluation framework itself?"

That question encourages reflection instead of automatic acceptance.

Principle 2
Preserve Friction Where It Protects Judgment

Not all friction is a problem.

Some forms of friction are how professional judgment develops.

An evaluator who pauses to explain a score is not wasting time. They are practising the reasoning their role requires. Removing that step may make the process faster, but it also removes an opportunity to strengthen judgment.

Research supports this idea.

Studies on cognitive forcing functions show that requiring people to form an independent view before seeing a recommendation reduces overreliance on automated advice (Buçinca, Malaya, & Gajos, 2021). Earlier work by Robert and Elizabeth Bjork reached a similar conclusion. Learning that feels easy often produces weaker long-term understanding, while learning that requires effort produces stronger and more durable skill (Bjork & Bjork, 2011).

The goal is not to make work harder.

It is to preserve the moments where thinking matters.

Good design removes unnecessary friction.

It protects the friction that helps people become better decision makers.

Example:

A government procurement office could implement this principle by requiring:

Instead of: Evaluator clicks "accept score" next to AI recommendation

Implement: Before seeing the AI score, evaluator answers:

- "What is the primary strength of this proposal's technical approach?"
- "What is one significant gap or concern?"
- Then the evaluator sees the AI-suggested score and can accept, modify, or justify disagreement

This friction forces the evaluator to form an independent judgment before the AI recommendation anchors their thinking. The friction trains the evaluator to read carefully; removing it removes the training. The evaluator who writes these justifications 100 times per cycle becomes better at spotting genuine capability versus polished language.

Principle 3
Make Human Reasoning Visible

Good decisions are not valuable simply because they are correct.

They are valuable because the reasoning behind them can be understood, questioned, and improved.

Most evaluation systems preserve the recommendation but not the thinking that produced the final decision. The score is recorded. The evaluator's reasoning often disappears.

The Better Humans Standard reverses that priority.

Instead of asking only *"What did the system recommend?"* it also asks *"Why did the evaluator agree or disagree?"*

Making reasoning visible serves two purposes.

First, it creates accountability. Decisions can be reviewed, challenged, and understood after they are made.

Second, it strengthens judgment itself.

People become better decision makers by explaining their reasoning, reflecting on it, and learning from it over time. If that reasoning is never recorded, it cannot be examined or improved.

Research in explainable AI has shown the importance of making automated decisions interpretable rather than opaque (Adadi & Berrada, 2018). European regulation has moved in the same direction by requiring meaningful explanations for significant automated decisions.

The same principle should apply to people.

If institutions believe evaluators are exercising independent judgment, they should also value the reasoning behind that judgment.

Example

Imagine a donor organization evaluating implementing partners.

Instead of recording only a score, the evaluation record also captures a short explanation.

> Why did this proposal receive 85 instead of 75?

Over time, those explanations become a valuable source of learning.

They may reveal that some evaluators consistently reward organizational size while others place greater weight on demonstrated capability. They may show that certain assumptions appear repeatedly across different reviewers.

The purpose is not to identify who was right or wrong.

It is to understand how decisions are being made and whether those patterns improve the quality of evaluation.

Principle 4
Design for Better Judgment Over Time

Most organizations measure whether technology improves productivity.

Very few measure whether it improves the people using it.

That is the final principle of the Better Humans Standard.

Instead of asking whether evaluations become faster, organizations should also ask whether evaluators become better at making them.

Judgment improves through feedback.

An experienced evaluator becomes more reliable because they learn which observations predicted success, which assumptions proved incorrect, and where their own reasoning consistently drifted.

Most organizations never collect that information.

Kahneman, Sibony, and Sunstein describe this as a failure to measure *noise*. Institutions often assume consistency where none exists because they never examine how judgments change over time (Kahneman, Sibony, & Sunstein, 2021).

Technology makes it possible to close that gap.

Rather than producing another efficiency dashboard, organizations could measure how evaluators develop.

Do their judgments become more consistent?

Do their assessments better predict future outcomes?

Do they become more accurate when dealing with ambiguous cases?

These questions tell us far more about the quality of an evaluation system than processing speed alone.

Example

A procurement office could review one year's worth of evaluations and compare them with actual contract performance.

An evaluator who consistently identifies successful contractors is developing good calibration.

Another whose highest-scoring contractors regularly underperform has an opportunity to improve.

Neither conclusion exists to replace professional judgment.

They exist to strengthen it.

The goal is not to create perfect evaluators.

The goal is to create evaluators who continue learning.

V. Anticipating Common Objections

No design standard is without trade offs. The Better Humans Standard asks organizations to value judgment alongside efficiency, and that inevitably raises questions about speed, cost, and practicality. The strongest objections deserve careful consideration.

"Won't this make evaluation slower"

It may.

Some of the practices proposed in this paper require evaluators to pause, explain their reasoning, or spend more time with ambiguous cases. Those activities take time.

The more important question is whether every minute removed from an evaluation process is genuinely waste.

Not all delays reduce performance. Some improve it.

An evaluator who spends an extra minute explaining why a proposal received a particular score is also strengthening the reasoning behind future decisions. A review process that pauses on uncertain cases may take slightly longer, but it is also more likely to distinguish genuine quality from superficial compliance.

The Better Humans Standard does not argue for slower evaluation.

It argues for spending human attention where human judgment matters most.

"Won't this increase costs"

It might.

Organizations that adopt these principles may invest more in evaluator training, oversight, or system design.

Those costs should be acknowledged rather than ignored.

The difficulty is that organizations rarely calculate the cost of declining judgment.

A procurement office can estimate the cost of new software or additional training. It is much harder to estimate the cost of repeatedly overlooking innovative suppliers, rewarding superficial compliance, or gradually weakening evaluator expertise.

Those costs are real even if they rarely appear in a budget.

The relevant comparison is not between spending money and spending nothing.

It is between two different kinds of investment.

One strengthens institutional judgment.

The other assumes it will always take care of itself.

"Our current approach works."

Perhaps it does.

Many organizations using automated evaluation achieve acceptable outcomes. That does not mean the underlying process cannot improve.

The central claim of this paper is not that current evaluation systems are failing.

It is that most organizations are measuring success too narrowly.

An organization may process applications more quickly while its evaluators become less confident handling unusual or ambiguous cases.

Both outcomes can exist at the same time.

The only way to know is to measure judgment directly.

That is why calibration matters.

If evaluators become more consistent, more accurate, and better aligned with long-term outcomes, the system is working.

If those capabilities decline while efficiency improves, organizations should ask whether they have optimized the right objective

VI. From Principle to Practice

The Better Humans Standard is intended as a design framework rather than a fixed implementation model. Different organizations will apply these principles in different ways depending on their sector, regulatory environment, and evaluation processes.

The purpose of this framework is not to prescribe a single method of implementation. It is to encourage organizations to ask better questions about the technologies they adopt and the capabilities they choose to strengthen.

Regardless of context, four questions remain central.

- Does the system make uncertainty visible or hide it?
- Does it strengthen professional judgment or gradually replace it?
- Does it preserve the reasoning behind important decisions?
- Does it help evaluators become better over time?

The answers will differ across procurement, grant making, hiring, healthcare, and other domains. The principles should not.

The Better Humans Standard is not intended to replace existing governance frameworks. It is intended to complement them by introducing a question that current evaluation models rarely ask.

What effect does this technology have on the people who rely on it?

The practical implementation of these principles will vary across organizations, but the underlying objective remains constant. Technology should strengthen the people making important decisions, not quietly diminish the judgment those decisions depend on.

VII. Implications for Policy and Practice

The Better Humans Standard is more than a design framework. It is also a different way of thinking about AI governance.

Much of today's governance focuses on model performance, safety, transparency, and accountability. Those questions remain essential. This paper argues that another question deserves equal attention.

What effect does a system have on the people who rely on it?

If AI changes how professionals develop judgment, evaluate evidence, and make decisions over time, then those effects should become part of how we assess responsible AI.

AI Governance Research

Research on AI evaluation has largely focused on system performance, fairness, robustness, and explainability. Much less attention has been given to the long-term effect of these systems on professional judgment.

That gap presents an important research opportunity.

Longitudinal studies could examine whether evaluators become more accurate, more consistent, and better calibrated after working alongside these technologies over extended periods. They could also explore whether some design choices strengthen judgment while others gradually weaken it.

Without that evidence, organizations may optimize systems without understanding what they are optimizing people to become.

Policy and Standards

Governments and standards bodies increasingly require organizations to evaluate the risks associated with high-impact AI.

Those assessments should extend beyond technical performance.

Organizations should also be encouraged to explain how their systems support human judgment. Do they make uncertainty visible? Do they preserve meaningful human reasoning? Do they help professionals improve over time?

These questions complement existing governance frameworks rather than replacing them. They recognize that responsible AI depends not only on trustworthy technology but also on capable people.

Professional Practice and Education

Professional organizations, universities, and training providers play an equally important role.

Future evaluators should learn not only how to use intelligent systems but also how to question them, challenge them, and understand where human judgment remains indispensable.

The Better Humans Standard offers one way of teaching those skills.

If professional education treats judgment as a capability that can either strengthen or decline through repeated interaction with technology, future practitioners will be better prepared to work alongside increasingly capable systems.

The implications of this paper therefore extend beyond procurement or proposal evaluation. They apply wherever technology increasingly participates in decisions that require human judgment.

VIII. Better Humans, Not Fewer Humans

She still has forty-one proposals to score by Friday.

She still has eleven minutes for each.

Nothing about the volume has changed, and nothing in this paper suggests that it will.

What has changed is what those eleven minutes are for.

They are no longer spent confirming a ranking she never questioned or inheriting a decision someone else has already made.

They are spent exercising judgment where judgment matters most.

That is the standard this paper has argued for.

The question is not whether technology can help organizations process more work.

It can.

The more important question is what repeated use does to the people making the decisions.

Does it help them become more thoughtful?

More discerning?

More confident in handling uncertainty?

Or does it slowly reduce judgment to the act of confirming recommendations that already feel inevitable?

Institutions already ask what these systems save.

They should also ask what they cultivate.

The future of AI will not be determined only by faster models, larger datasets, or more capable algorithms.

It will also be shaped by the kinds of professionals those systems help create.

Technology should leave people more capable than it found them.

More curious.

More accountable.

More willing to question.

More able to recognize quality when it cannot be reduced to a score.

That is the promise of human-centered AI.

And that is the standard this paper proposes.

The measure of good AI is not how many people it replaces. It is how many people it makes better at deciding what actually matters.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160.

Artificial Intelligence Review. (2025). AI-induced deskilling in medicine: A mixed-method review and research agenda for healthcare and beyond. *Springer Nature*.

Bhatt, A., et al. (2026). Artificial intelligence in medicine: A scoping review of the risk of deskilling and loss of expertise among physicians. *ESMO Real World Data and Digital Oncology*.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*. Worth Publishers.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5, 1–21.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126.

Goodhart, C. (1975). Problems of monetary management: The U.K. experience. *Papers in Monetary Economics*, Reserve Bank of Australia.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A Flaw in Human Judgment*. Little, Brown Spark.

Koprince, S. (2021). The past performance chicken-and-egg problem for new government contractors. *Small GovCon*.

NIST. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology, U.S. Department of Commerce.

Russo, G., Horta Ribeiro, M., Davidson, T. R., Veselovsky, V., & West, R. (2025). The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates. *Proceedings of the ACM on Human-Computer Interaction*, 9(7), 1–18.

U.S. Government Accountability Office. (2011). *Prior Experience and Past Performance as Evaluation Criteria in the Award of Federal Construction Contracts (GAO-12-102R)*.

Wilson, K., et al. (2025). People mirror AI systems' hiring biases. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '25)*

This working paper is shared to encourage discussion and feedback. Future versions may incorporate additional evidence, peer feedback, and empirical validation.