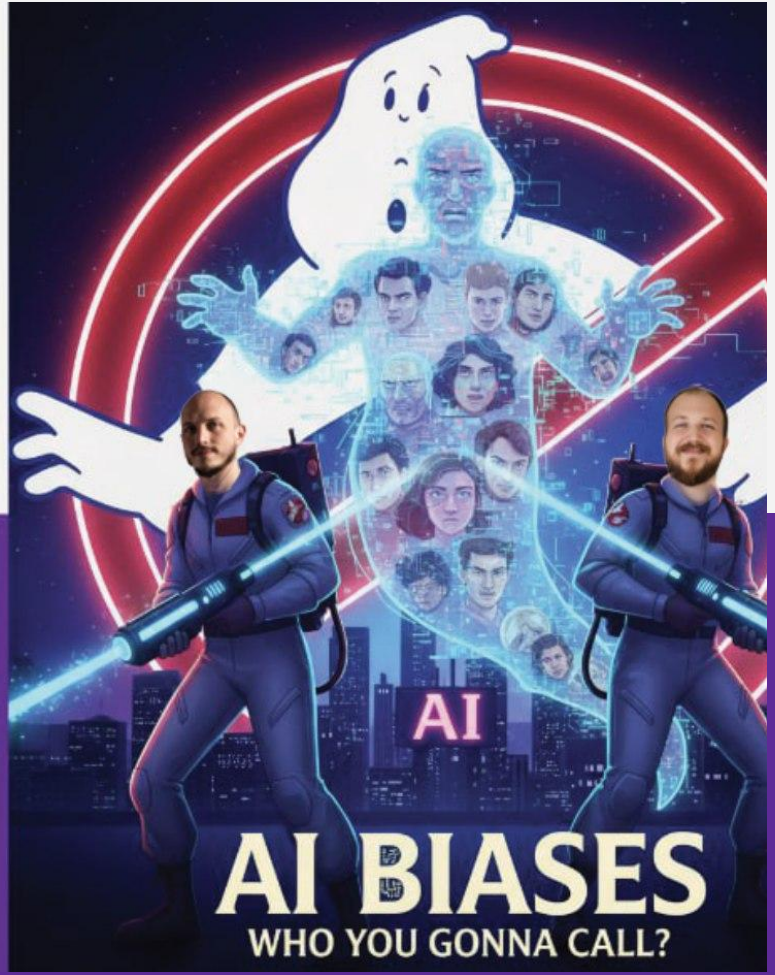
AI Biases

Who You Gonna Call?



October 15, 2025.

Djordje Babić
&
Mitrović Luka

Agenda

- Why bias shows up in human–AI work
- Four common traps to watch
- Four guardrails that work in Agile
- TRIZ mini-exercise
- Your 10-day experiment + Q&A

INTRODUCTION

Why This Matters to you?

Key reasons bias impacts human–AI decision-making

- Bias degrades decision quality
- Creates compliance risks
- Naïve trust in AI and reflexive skepticism destroy value
- Simple guardrails
- Practical mitigation strategies
- Bias Awareness

Recruiting/HR tech: Amazon scrapped an internal **AI recruiter** that favored men—training data reflected historical male dominance in tech roles.
So what: historical data ≠ neutral; audits need to look for disparate impact even after obvious feature removals.

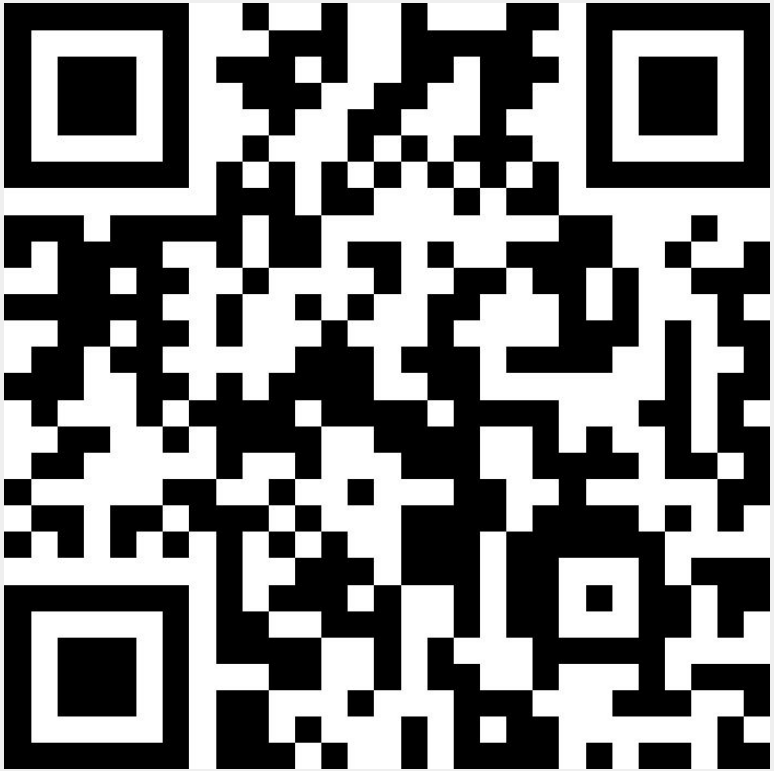
Serbia AI Adoption low but rising

So what: 34% companies using AI in Jan but only 20% of them have a plan for it.

Serbia has a plan for AI adoption and usage 2025-2030

INTRODUCTION

Which AI do you use?



slido.com

2788053

BIAS

What We Mean by Bias

Bias is a systematic error in judgment influenced by both human and AI factors

- Bias = systematic error in judgment
- Human cognitive shortcuts
- AI systems are biased
- Reinforcing feedback loop
- Address both

Anchoring: first AI estimate frames effort/scope and drags Jira sizing toward it.

Automation bias: we accept LLM summaries or labels without independent checks.

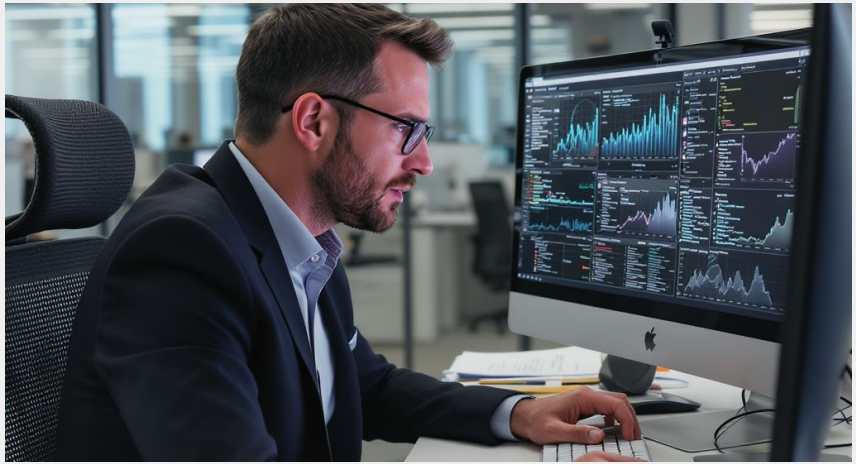
Historical/selection bias: training data over-represents past winners, undercuts novel briefs.

Feedback-loop bias: promoted outputs get more clicks → retraining sees more of them → model doubles down.

TRAPS

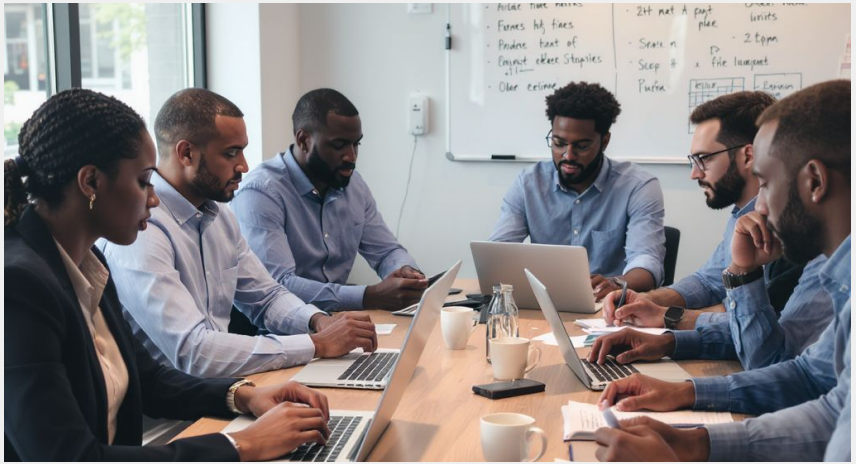
Traps: Automation Bias & Anchoring

Automation Bias



Over-trusting AI suggestions

Anchoring



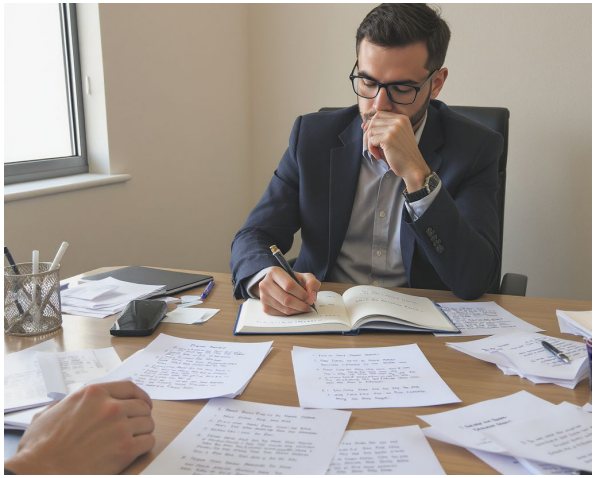
The first number or idea pulls estimates toward it

Over-trusting AI under pressure and anchoring on initial AI outputs can distort decisions; independent checks and the AI-last rule help mitigate these biases.

TRAPS

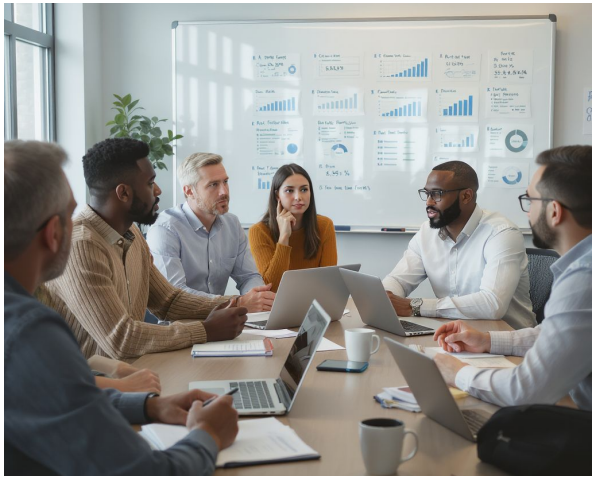
Traps: Framing Effect & Overconfidence

Prompt wording can steer AI outputs and decisions, while fluent AI answers can create overconfidence and deskilling—both require active mitigation through broad perspectives and verification.



Framing Effect

Prompt wording steers outputs and decisions



Overconfidence & Deskilling

Fluent answers are not always correct + damage skills

FEEDBACK LOOP

Human–AI Feedback Loop

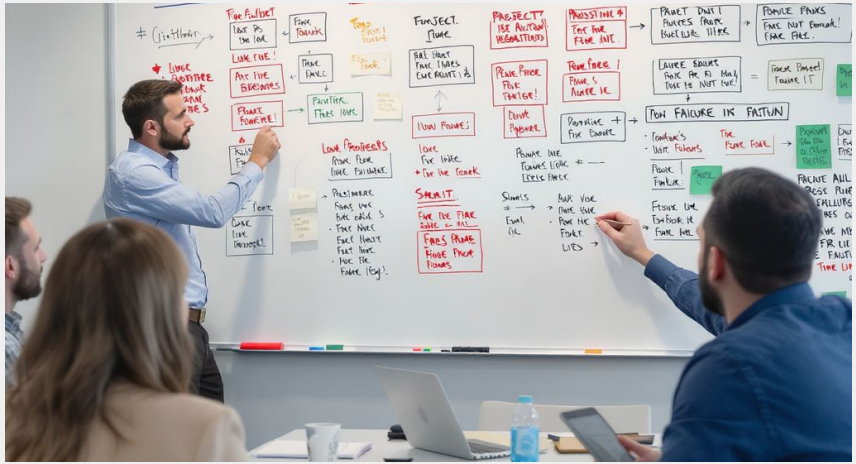
Human biases in prompt design and AI's reliance on data and framing create a reinforcing feedback loop that can skew decisions. Breaking this cycle requires deliberate prompts, encouraging dissenting views, and implementing review gates to ensure balanced outcomes.

How human biases & AI outputs reinforce beliefs and how to break the loop

- Human biases → biased prompts
- AI reflects training data
- Reinforcing feedback loop
- Overconfidence loop
- Breaking the loop (how can this fail in the real world?)
- Encourage human dissenting views
- Implement (AI) review gates

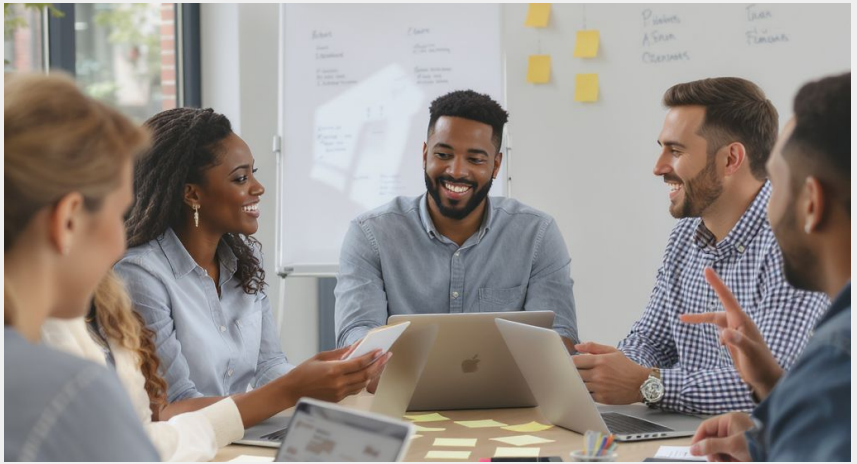
Guardrails Part 1: Pre-mortem & Planning Poker

Pre-mortem



The project failed - why?

Planning Poker (AI-last)

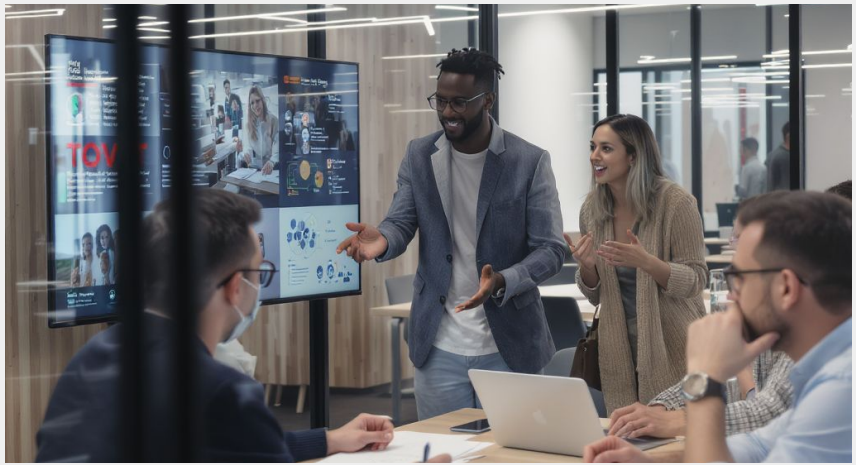


Gather human estimates first

Using pre-mortem and Planning Poker with the AI-last rule helps teams identify hidden risks and avoid bias from early AI input, fostering more accurate and balanced decisions.

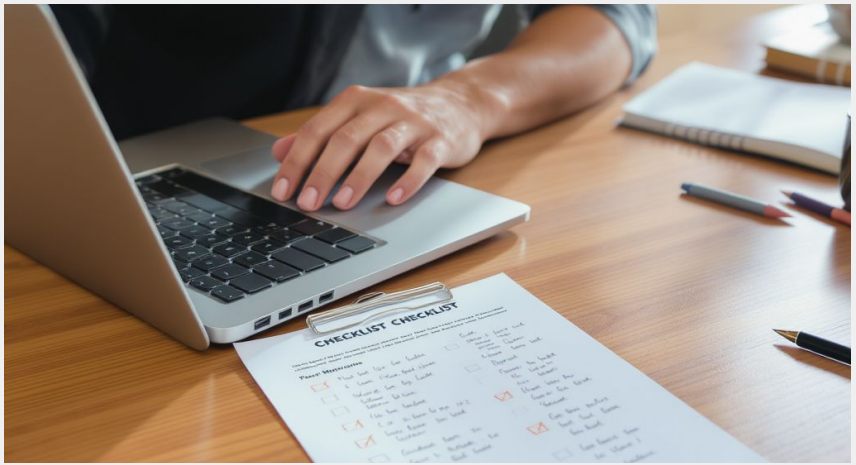
Guardrails Part 2: Red/Blue Teaming & Bias Checklist

Red/Blue Teaming



Rotate challenge roles

Bias Checklist (fast DoR pause)



Quick checklist pause

Implementing Red/Blue Teaming and a Bias Checklist encourages constructive dissent and thorough evaluation, ensuring AI outputs are robust and decisions are well-calibrated.

TOOLKIT

Your Toolkit for Ceremonies

Essential tools to integrate bias guardrails into Agile ceremonies

- DoR Bias Checklist (Backlog refinement)
- Pre-mortem prompt (Sprint planning)
- Red/Blue script (Retrospective/Review)
- AI-last rule card (Estimation)

Using targeted tools like the DoR Bias Checklist, Pre-mortem prompts, Red/Blue scripts, and AI-last rule cards helps teams embed bias guardrails directly into Agile ceremonies, enhancing decision quality and team accountability.

ACCOUNTABILITY

Delegation & Accountability

Key practices for defining roles and ensuring accountability in human-AI teams

- Who makes the final call?
- “Moral crumple zone” - explicitly log AI inputs
- Make accountability explicit
- Keep an auditable trail of significant AI enhanced decisions
- Transparent documentation practices
- Regularly review

Clear boundaries between AI advisory roles and human decision-making responsibilities, combined with transparent logging and accountability practices, prevent ethical pitfalls and ensure traceability in decision processes.

EXERCISE

TRIZ Mini-Exercise

Maximize AI-Induced Rework: Identify and Stop Behaviors

- If we wanted to maximize AI bias in our product, what would we do?
- What are we currently doing?
- What can we do different starting Monday?

Using a reverse-thinking exercise helps teams identify and stop behaviors that maximize AI-induced rework, improving decision quality and efficiency.

EXPERIMENT

Your 10 Minute Experiment - PCR

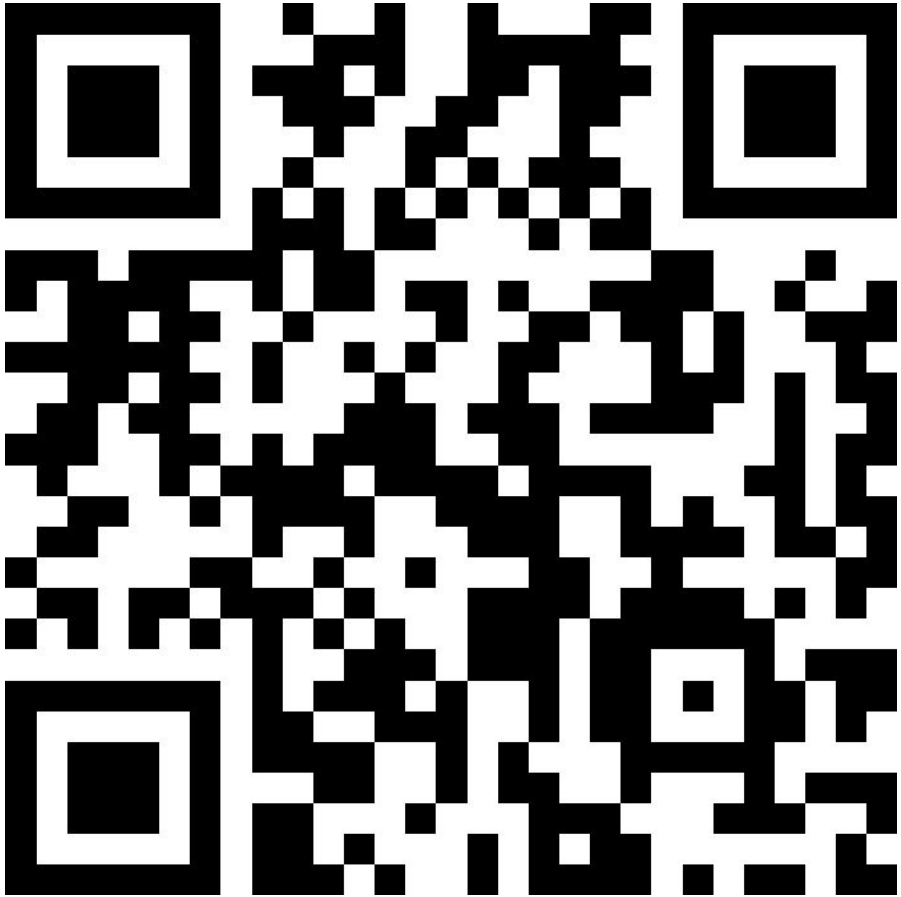
Implement a focused bias-proofing experiment to build resilient decision-making in your team

Pre-Commit & Replay

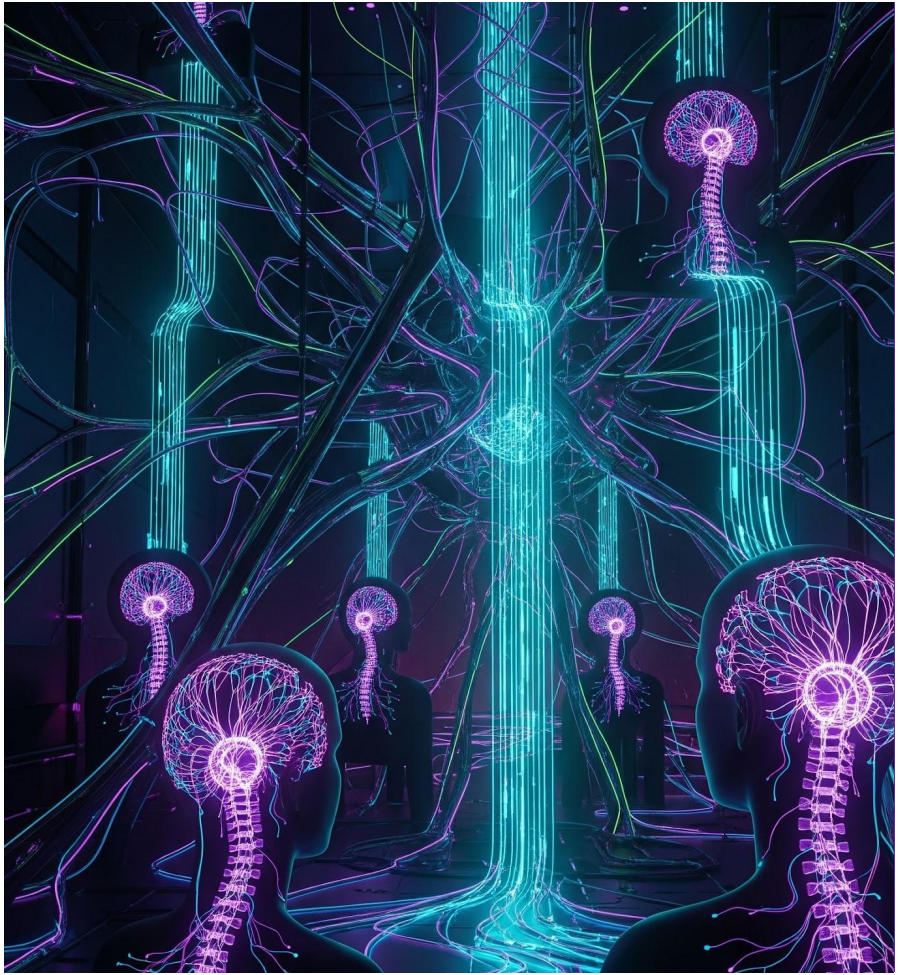
- Write what you want from your AI interaction (desired outcome)
- Three assumptions (what we believe)
- Constraint behavior (2 separate viewpoints)
- How will we know the result is good? (give examples)

Try on another LLM and compare

Choosing one bias and one guardrail to apply in your Agile ceremonies over a 10-day period enables practical learning and measurable improvements. Sharing results in retrospectives fosters continuous team growth and bias mitigation.



Q&A



Thank you!

djordjebabic@hivemind.rs

luka.b.mitrovic@gmail.com