



***Designing Meaningful Oversight of Agentic AI Systems:
Experimental Effects of Overseer Coupling Granularity on
Appropriate Reliance***

Sarah Starr (Presenting) & Dr. Zoe Szajnfarber

AI4SE & SE4AI Workshop
September 22, 2026

Engineering Management and Systems Engineering
The George Washington University



Agenda



Introduction and motivations

Study Design

Results

Discussion

Conclusion and future work

Human approval is being redesigned without making tradeoffs explicit

MOTIVATION

- 1 Agentic AI is entering operational workflows faster than the methods for overseeing it.**
Stanford HAI 2026; Tinkoff et al. (McKinsey) 2026
- 2 Industry guidance is moving oversight from per-action confirmation to plan-level review.**
Anthropic 2026a, 2026b; Passi 2025
- 3 Granularity is treated as a friction dial. Its price in oversight has been assumed, not measured.**
NIST 2023; Shavit et al. 2023; Gabriel et al. 2024 · Bansal et al. 2021; Lai et al. 2023; Buçinca et al. 2021

Study instrumented to address two related research questions:

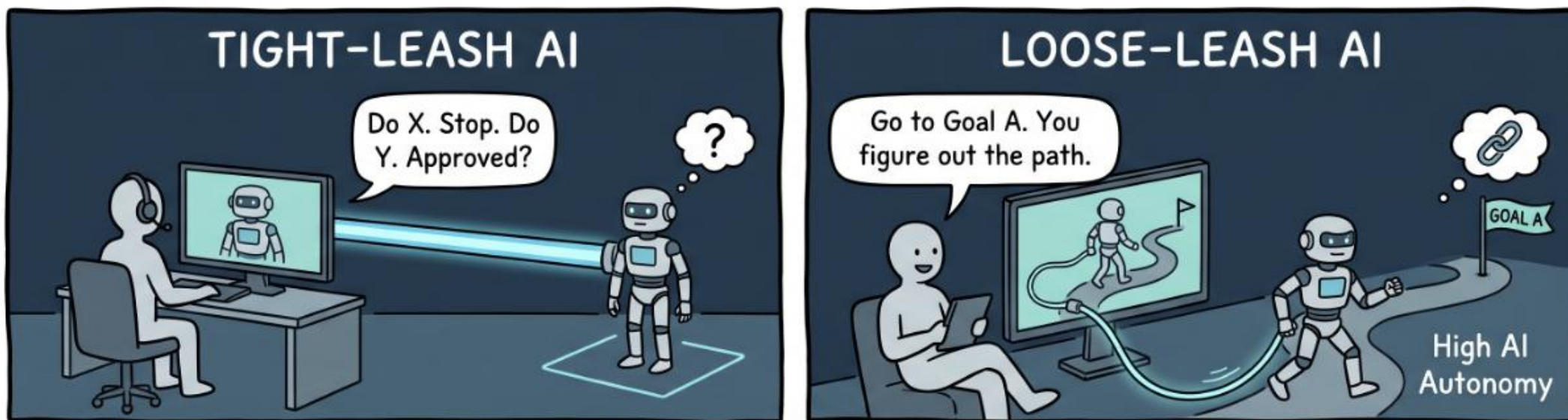
RESEARCH QUESTIONS

RQ
#1

How does human-AI coupling shape functional reliance on AI output?

RQ
#2

How does coupling shape the alignment between self-reported trust (attitude) and functional reliance behavior?



Study instrumented to address two related research questions:

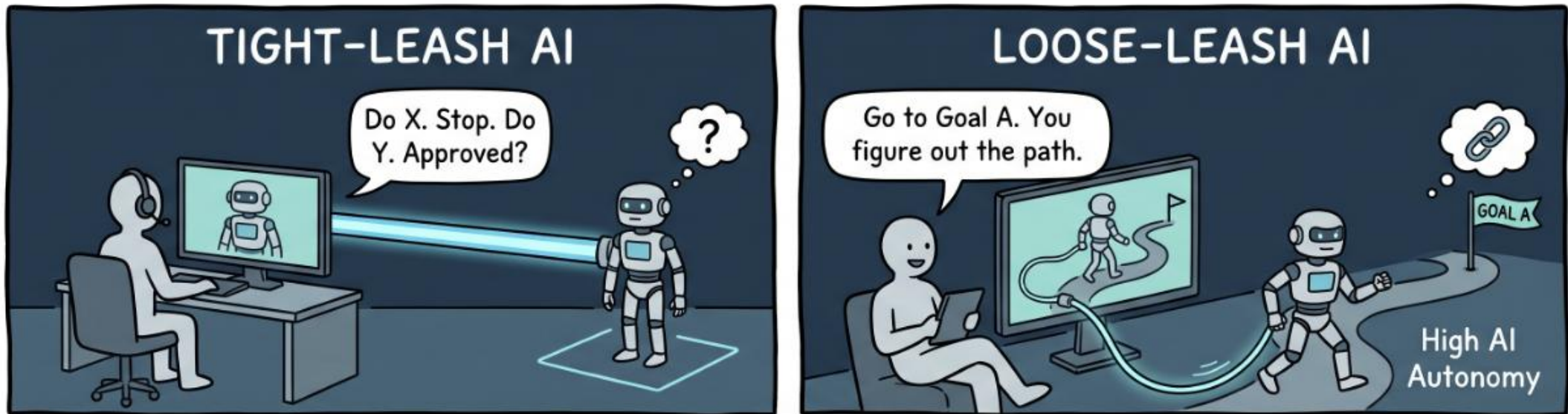
RESEARCH QUESTIONS

RQ
#1

How does human-AI coupling shape functional reliance on AI output?

RQ
#2

How does coupling shape the alignment between self-reported trust (attitude) and functional reliance behavior?



Agenda

Introduction and motivations

Study Design

Results

Discussion

Conclusion and future work

Isolating where approval attaches to agent action as a variable

STUDY DESIGN

- 1

Single factor, two conditions
oversight coupling: TIGHT per-action vs LOOSE plan-level
- 2

Randomized, between-subjects
each participant sees one condition
- 3

12 participants, observed in depth
6 per condition, 3 rounds per participant

MANIPULATION: WHERE APPROVAL ATTACHES

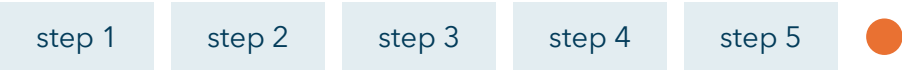
TIGHT: per-action coupling

the agent pauses after every step and waits for approval



LOOSE: plan-level coupling

the agent runs the whole plan; one review at the end



MODEL WORLD: COURSE REGISTRATION

An AI advising agent builds and submits a course registration for a fictional student persona while the participant oversees it.



Same agent, personas, materials, and 8 seeded errors in both conditions

Bounded but non-trivial option space a full course catalog

Constraints that must hold jointly term · times · prerequisites · credit cap

A terminal commitment submitting the registration

*Same shape as other workflows of interest such as **software development** and **intel BTD**.*

Interface overview: what participants saw and could use

Both conditions used the same agent, personas, and reference materials, and every open, switch, and toggle was logged as a verification action.

TIGHT · approve each step

PERSONA 1 OF 3

Course Registration: Marcus Bennett

The AI agent will propose a sequence of registration actions for the persona. Approve each action, or tell the agent how to refine it.

SEARCHED THE COURSE CATALOG

SHOW REASONING

THE AI PROPOSES THESE COURSES:

- APSC 2114 - Engineering Analysis I (3cr)
- APSC 2115 - Engineering Analysis II (3cr)
- CE 2220 - Introduction to the Mechanics of Solids (3cr)
- CE 3054 - Physical Hydrology (3cr)
- CE 3250 - Structural Analysis (3cr)

REVIEW PROPOSED COURSES

CONTINUE TO NEXT STEP

Tell the agent how to change this step...

Once you review a step you can approve it if no changes are needed

Agent proposes one step; Continue approves it

LOOSE · approve the plan once

PERSONA 1 OF 3

Course Registration: Marcus Bennett

The AI agent will propose a sequence of registration actions for the persona. Approve each action, or tell the agent how to refine it.

SEARCHED THE COURSE CATALOG

SHOW REASONING

THE AI PROPOSES THESE COURSES:

- APSC 2114 - Engineering Analysis I (3cr)
- APSC 2115 - Engineering Analysis II (3cr)
- CE 2220 - Introduction to the Mechanics of Solids (3cr)
- CE 3054 - Physical Hydrology (3cr)
- CE 3250 - Structural Analysis (3cr)

REVIEW PROPOSED COURSES

CONTINUE TO NEXT STEP

Tell the agent how to change this step...

Once you review the plan you can approve it if no changes are needed

Agent runs all five steps; one approval window

Edit instruction

PERSONA 1 OF 3

Course Registration: Marcus Bennett

The AI agent will propose a sequence of registration actions for the persona. Approve each action, or tell the agent how to refine it.

SEARCHED THE COURSE CATALOG

SHOW REASONING

THE AI PROPOSES THESE COURSES:

- APSC 2114 - Engineering Analysis I (3cr)
- APSC 2115 - Engineering Analysis II (3cr)
- CE 2220 - Introduction to the Mechanics of Solids (3cr)
- CE 3054 - Physical Hydrology (3cr)
- CE 3250 - Structural Analysis (3cr)

REVIEW PROPOSED COURSES

CONTINUE TO NEXT STEP

Tell the agent how to change this step...

If changes are needed you can descriptive edits to the AI Agent

Free-text instruction; the agent re-plans in response

Plan changes

PERSONA 1 OF 3

Course Registration: Marcus Bennett

The AI agent will propose a sequence of registration actions for the persona. Approve each action, or tell the agent how to refine it.

SEARCHED THE COURSE CATALOG

SHOW REASONING

THE AI PROPOSES THESE COURSES:

- APSC 2114 - Engineering Analysis I (3cr)
- APSC 2115 - Engineering Analysis II (3cr)
- CE 2220 - Introduction to the Mechanics of Solids (3cr)
- CE 3054 - Physical Hydrology (3cr)
- CE 3250 - Structural Analysis (3cr)

REVIEW PROPOSED COURSES

CONTINUE TO NEXT STEP

Tell the agent how to change this step...

Close plan changes here

Before/after diff of what an edit added, removed, kept

Agent reasoning

PERSONA 1 OF 3

Course Registration: Marcus Bennett

The AI agent will propose a sequence of registration actions for the persona. Approve each action, or tell the agent how to refine it.

SEARCHED THE COURSE CATALOG

SHOW REASONING

THE AI PROPOSES THESE COURSES:

- APSC 2114 - Engineering Analysis I (3cr)
- APSC 2115 - Engineering Analysis II (3cr)
- CE 2220 - Introduction to the Mechanics of Solids (3cr)
- CE 3054 - Physical Hydrology (3cr)
- CE 3250 - Structural Analysis (3cr)

REVIEW PROPOSED COURSES

CONTINUE TO NEXT STEP

Tell the agent how to change this step...

Show reasoning

"Show reasoning" opens the agent's rationale for the current step

Degree Map

PERSONA 1 OF 3

Course Registration: Marcus Bennett

The AI agent will propose a sequence of registration actions for the persona. Approve each action, or tell the agent how to refine it.

SEARCHED THE COURSE CATALOG

SHOW REASONING

THE AI PROPOSES THESE COURSES:

- APSC 2114 - Engineering Analysis I (3cr)
- APSC 2115 - Engineering Analysis II (3cr)
- CE 2220 - Introduction to the Mechanics of Solids (3cr)
- CE 3054 - Physical Hydrology (3cr)
- CE 3250 - Structural Analysis (3cr)

REVIEW PROPOSED COURSES

CONTINUE TO NEXT STEP

Tell the agent how to change this step...

Degree Map tab contains the student's records and which classes they have taken

Persona's completed and remaining requirements

Course catalog

PERSONA 1 OF 3

Course Registration: Marcus Bennett

The AI agent will propose a sequence of registration actions for the persona. Approve each action, or tell the agent how to refine it.

SEARCHED THE COURSE CATALOG

SHOW REASONING

THE AI PROPOSES THESE COURSES:

- APSC 2114 - Engineering Analysis I (3cr)
- APSC 2115 - Engineering Analysis II (3cr)
- CE 2220 - Introduction to the Mechanics of Solids (3cr)
- CE 3054 - Physical Hydrology (3cr)
- CE 3250 - Structural Analysis (3cr)

REVIEW PROPOSED COURSES

CONTINUE TO NEXT STEP

Tell the agent how to change this step...

Resources tab contains course catalog details for each course in greater detail

Full catalog under Resources: terms, times, prereqs

Weekly calendar

PERSONA 1 OF 3

Course Registration: Marcus Bennett

The AI agent will propose a sequence of registration actions for the persona. Approve each action, or tell the agent how to refine it.

SEARCHED THE COURSE CATALOG

SHOW REASONING

THE AI PROPOSES THESE COURSES:

- APSC 2114 - Engineering Analysis I (3cr)
- APSC 2115 - Engineering Analysis II (3cr)
- CE 2220 - Introduction to the Mechanics of Solids (3cr)
- CE 3054 - Physical Hydrology (3cr)
- CE 3250 - Structural Analysis (3cr)

REVIEW PROPOSED COURSES

CONTINUE TO NEXT STEP

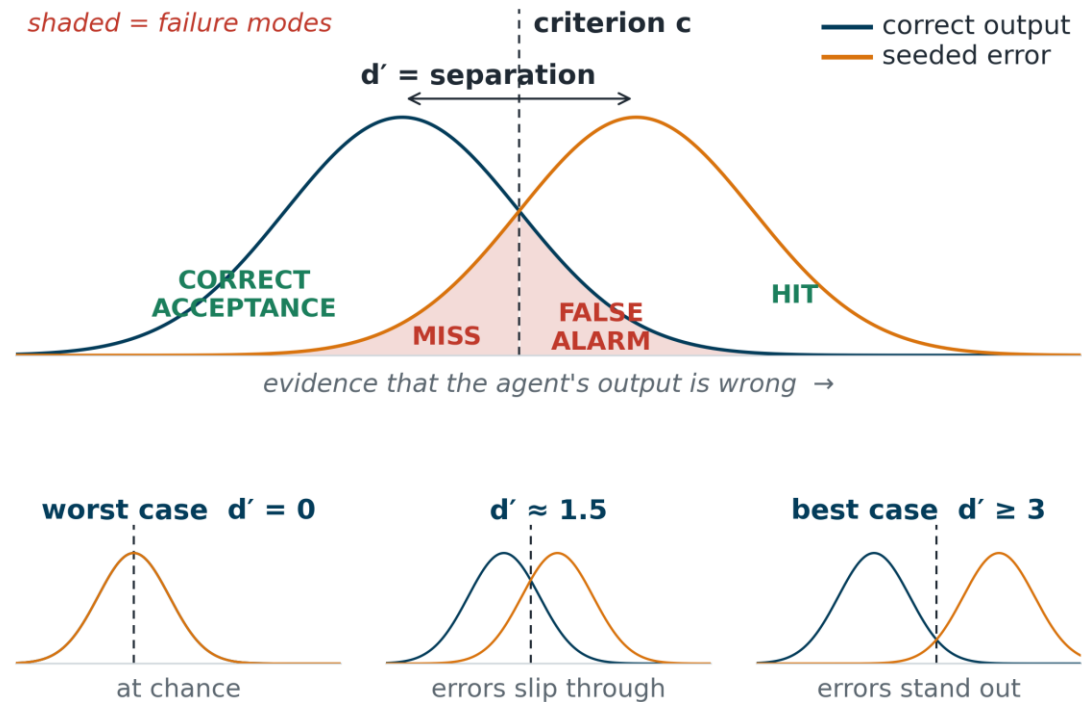
Tell the agent how to change this step...

You can hide the calendar here

Proposed courses by day and time; conflicts flagged

Measuring reliance: can overseers tell incorrect from correct agent outputs?

		WHAT THE PARTICIPANT DID	
		Intervened (edit / veto)	Accepted
AGENT OUTPUT	Seeded error signal	HIT ✓ caught the error MAE 2131, Spring-only, proposed for Fall → edited out at the checkpoint	MISS ✗ over-reliance: relied on wrong output MAE 3126, prerequisite not met → approved; rides through to submission
	Correct output noise	FALSE ALARM ✗ under-reliance: overrode good output APSC 2114, valid → swapped out "just in case"	CORRECT ACCEPTANCE ✓ appropriate reliance MAE 3127, valid → approved



d' = z(hit rate) – z(false-alarm rate) discrimination: how well the participant tells wrong output from right (primary measure)

c = –½ [z(hit rate) + z(false-alarm rate)] intervention bias: how often a participant edits agent responses, leaning toward intervening or accepting

Agenda

Introduction and motivations

Study Design

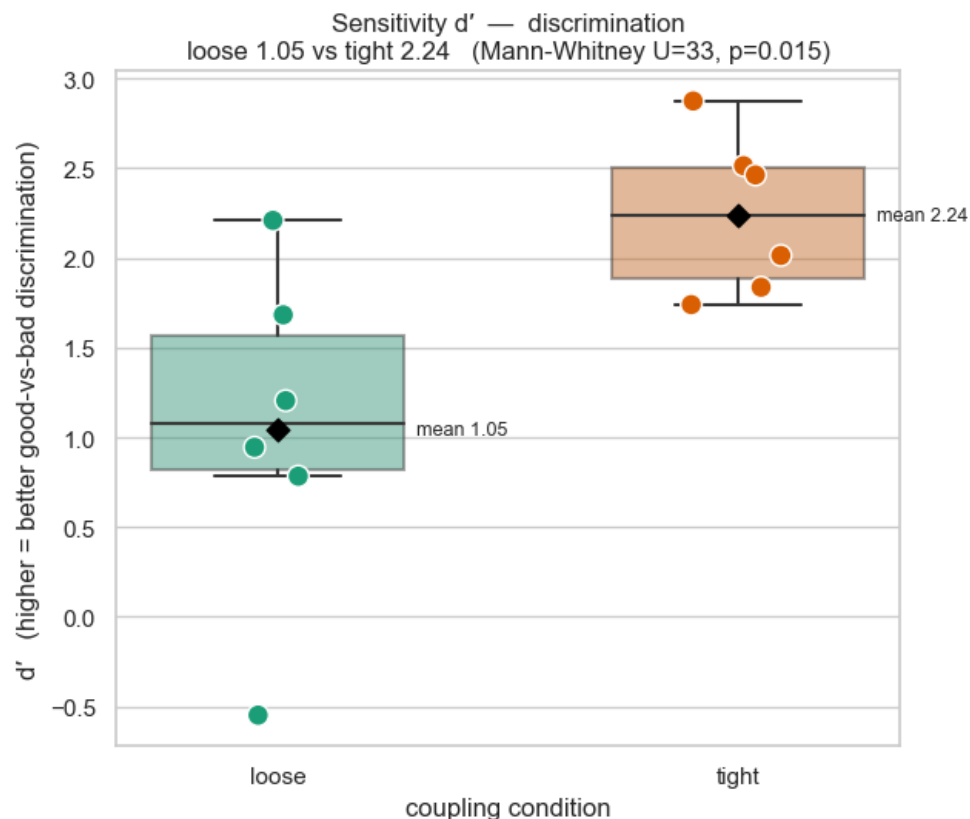
Results

Discussion

Conclusion and future work

Tight coupling raises overseer discrimination

RESULTS



Sensitivity d' by condition; box = IQR, $\blacklozenge$ = mean; $n = 12$.

2.24 vs 1.05

sensitivity d' , tight vs loose: higher discrimination under per-action coupling (Mann-Whitney $U = 33$, $p = .015$)

88% vs 69%

of seeded errors caught, tight vs loose; false alarms on correct output 11% vs 32%

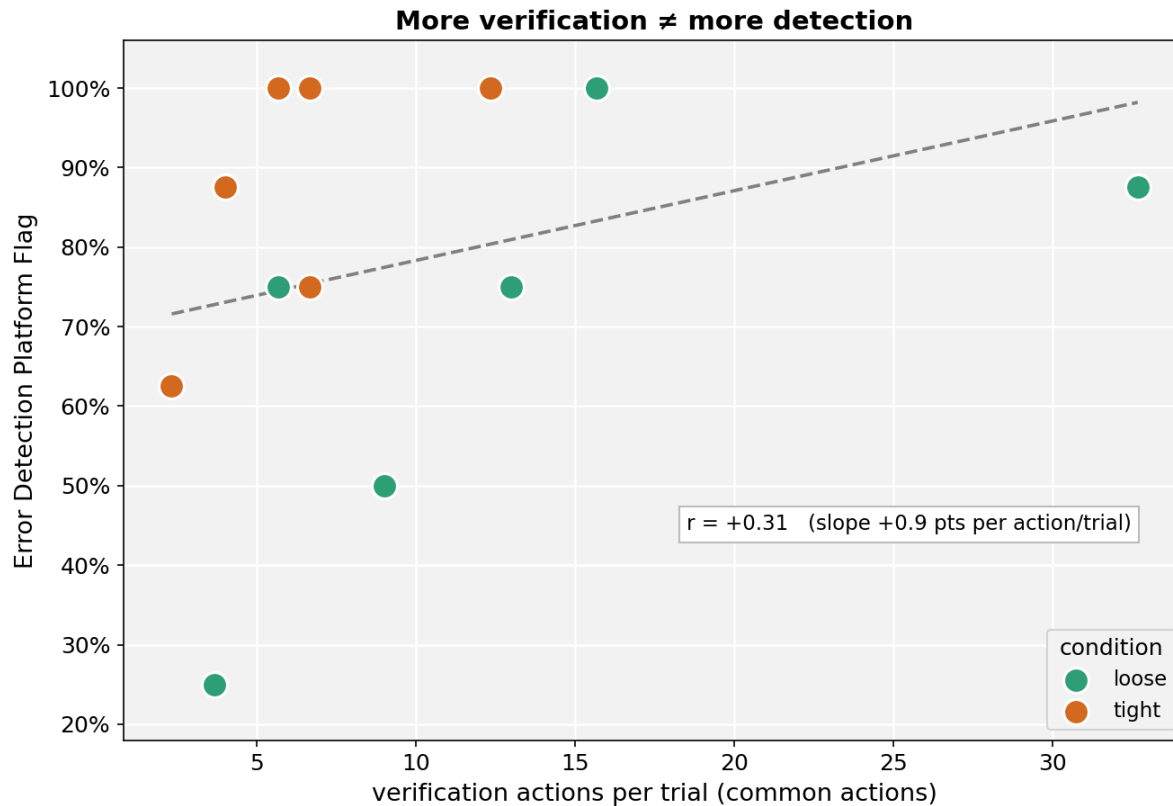
$c \approx 0$

intervention bias in both conditions ($p = 1.0$): tight reviewers judged better, they did not simply intervene more

Same agent, same task, same eight seeded errors: per-action approval improved participants' ability to tell wrong agent output from right.

Verification effort does not guarantee detection

RESULTS



13.3

verification actions / trial under LOOSE coupling

6.3

verification actions / trial under TIGHT coupling with higher detection

8.8 vs 8.6

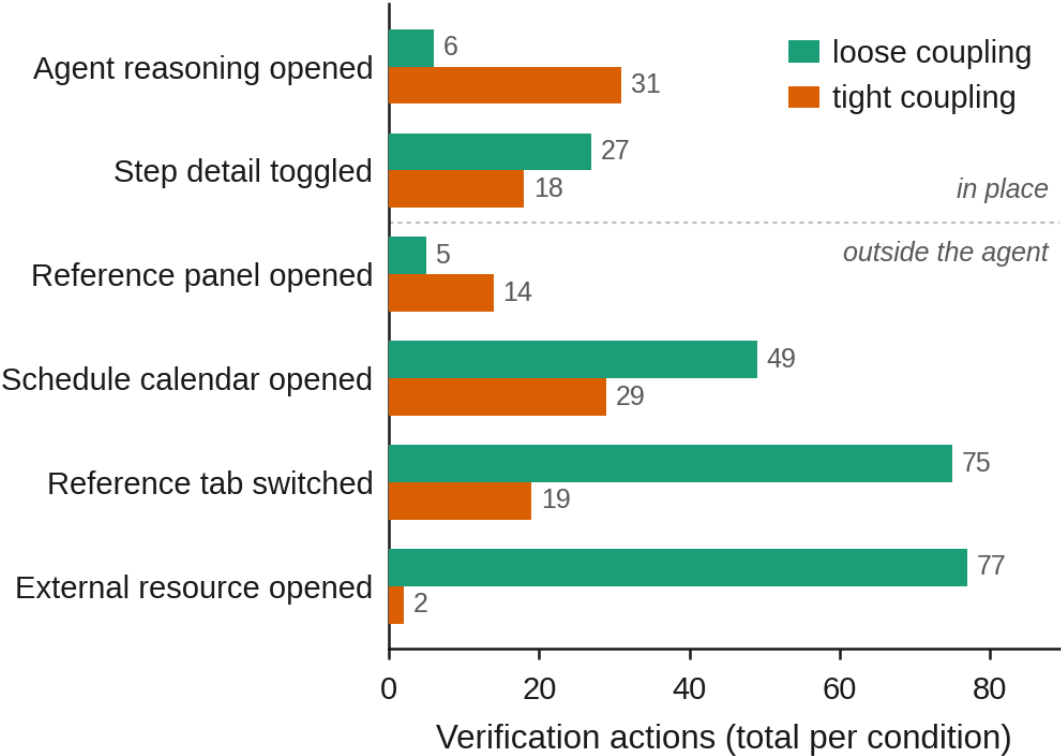
minutes per trial, TIGHT vs LOOSE (inside the same 12-minute budget)

Effort alone does not buy detection: tight coupling did not buy its advantage with time; per-action review made the same review time more effective.

Per-participant verification actions/trial vs platform-flag detection; $n = 12$, descriptive.

Where verification happens shifts with coupling

RESULTS



Verification actions by type, totals per condition (n = 12); dashed line = inside vs outside the agent.

77 vs 2

external-resource opens, LOOSE vs TIGHT: loose participants checked by leaving the agent

31 vs 6

opens of the agent's reasoning, TIGHT vs LOOSE: tight participants checked in place

75 vs 19

reference-tab switches, LOOSE vs TIGHT: a finished plan has to be reconstructed from references

The checkpoint arrives with the evidence: coupling changed what a unit of checking could accomplish, not how much.

Example: what participants typed to the agent

TIGHT "MAE 3128 needs to be removed. It violates the pre-req rules." → a verdict on a claim

LOOSE "Can you redo the plan so that I can take 17 credits." → a steer on the plan

Agenda

Introduction and motivations

Study Design

Results

Discussion

Conclusion and future work

Discussion and limitations

Discussion

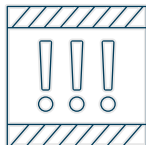


Coarser oversight reduced discrimination at no savings in user effort or review time



Per-action checkpoints keep review next to the evidence: an architectural cognitive forcing function

Limitations



- Small sample: $n = 12$, six per condition
- Participants role-played fictional personas; adherence varied, and groups were balanced on academic level but not on field
- One well-defined five-step workflow; tasks with many more steps may raise burden and habituation under per-action coupling

Agenda

Introduction and motivations

Study Design

Results

Discussion

Conclusion and future work

Conclusion and next steps

CONCLUSIONS

1

With agent, task, and errors held constant, coupling alone changed what overseers caught: tighter coupling, better discrimination

2

Coupling changed where verification happened: tight reviewers checked each claim in place; loose reviewers left the agent to reconstruct the plan from resources

3

The speed-for-oversight trade-off used to justify coarser approval did not materialize here: it should be measured, not assumed

FUTURE WORK

Transfer to workflows with the same structural signature: ticket triage, software change management, intelligence target development

Repeated exposure: does the coupling advantage persist, grow, or erode as the agent becomes familiar?

Map the wider granularity spectrum, add a no-AI baseline, and vary the other dimensions: control architecture and agent scope

Thank you!

CONCLUSIONS

1

With agent, task, and errors held constant, coupling alone changed what overseers caught: tighter coupling, better discrimination

2

Coupling changed where verification happened: tight reviewers checked each claim in place; loose reviewers left the agent to reconstruct the plan from resources

3

The speed-for-oversight trade-off used to justify coarser approval did not materialize here: it should be measured, not assumed

A dark blue square containing the white text "Q / A" in a large, bold, sans-serif font.

sarah.starr@gwu.edu

References

1. I. Ajzen and M. Fishbein, "Attitude-behavior relations: A theoretical analysis and review of empirical research," *Psychological Bulletin*, vol. 84, no. 5, pp. 888-918, 1977. doi: 10.1037/0033-2909.84.5.888
2. Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman, "Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems," in *Proc. 25th Int. Conf. on Intelligent User Interfaces (IUI)*, Mar. 2020, pp. 454-464. doi: 10.1145/3377325.3377498
3. D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*. New York: Wiley, 1966.
4. K. A. Hoff and M. Bashir, "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust," *Human Factors*, vol. 57, no. 3, pp. 407-434, May 2015. doi: 10.1177/0018720814547570
5. C. Kovesdi, Y. Yamani, A. Jackson, and L. Lin, "Application of Signal Detection Theory in Evaluating Trust of Information Produced by Large Language Models," *Proc. Human Factors and Ergonomics Society Annual Meeting*, vol. 69, no. 1, pp. 582-587, 2025. doi: 10.1177/10711813251368829
6. J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors*, vol. 46, no. 1, pp. 50-80, Mar. 2004. doi: 10.1518/hfes.46.1.50_30392
7. M. Madsen and S. Gregor, "Measuring Human-Computer Trust," in *Proc. 11th Australasian Conf. on Information Systems*, 2000.
8. K. Okamura and S. Yamada, "Adaptive trust calibration for human-AI collaboration," *PLoS ONE*, vol. 15, no. 2, art. e0229132, Feb. 2020. doi: 10.1371/journal.pone.0229132
9. R. Parasuraman and V. Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors*, vol. 39, no. 2, pp. 230-253, Jun. 1997. doi: 10.1518/001872097778543886
10. M. Raees and K. Papangelis, "From Trust to Appropriate Reliance: Measurement Constructs in Human-AI Decision-Making," *arXiv:2604.23896*, Apr. 2026. doi: 10.48550/arXiv.2604.23896
11. M. Schemmer, P. Hemmer, N. Köhl, C. Benz, and G. Satzger, "Should I Follow AI-based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making," *arXiv:2204.06916*, Apr. 2022. doi: 10.48550/arXiv.2204.06916
12. A. Singh and Z. Szajnfarder, "Architecting Human-AI Systems for Effective Collaboration and Oversight: Making Sense of Human/AI-in/on/Over/Under/Along-the-Loop," *Systems Engineering*, vol. 29, no. 2, pp. 337-353, 2026. doi: 10.1002/sys.70024
13. H. Stanislaw and N. Todorov, "Calculation of signal detection theory measures," *Behavior Research Methods, Instruments, & Computers*, vol. 31, no. 1, pp. 137-149, 1999. doi: 10.3758/BF03207704

References (cont.)

14. B. B. Anderson, C. B. Kirwan, J. L. Jenkins, D. Eargle, S. Howard, and A. Vance, "How Polymorphic Warnings Reduce Habituation in the Brain: Insights from an fMRI Study," in Proc. CHI '15, 2015, pp. 2883-2892. doi: 10.1145/2702123.2702322
15. Anthropic, "Measuring AI agent autonomy in practice," 2026. <https://www.anthropic.com/research/measuring-agent-autonomy>
16. Anthropic, "Trustworthy agents in practice," 2026. <https://www.anthropic.com/research/trustworthy-agents>
17. G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld, "Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance," in Proc. CHI '21, 2021, pp. 1-16. doi: 10.1145/3411764.3445717
18. R. Böhme and S. Köpsell, "Trained to Accept? A Field Experiment on Consent Dialogs," in Proc. CHI '10, 2010, pp. 2403-2406. doi: 10.1145/1753326.1753689
19. Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making," Proc. ACM Hum.-Comput. Interact., vol. 5, CSCW1, pp. 1-21, 2021. doi: 10.1145/3449287
20. I. Gabriel, A. Manzini, G. Keeling, et al., "The Ethics of Advanced AI Assistants," Google DeepMind, arXiv:2404.16244, 2024.
21. E. Horvitz, "Principles of Mixed-Initiative User Interfaces," in Proc. CHI '99, 1999, pp. 159-166. doi: 10.1145/302979.303030
22. B. Huffman and Z. Szajnfarder, "Workflow Architecture: A New Construct in the Age of AI-Enabled Systems," AI4SE & SE4AI Workshop, 2026 (extended abstract, submitted).
23. C. Lai, C. Chen, A. Smith-Renner, Q. V. Liao, and C. Tan, "Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies," in Proc. FAccT '23, 2023, pp. 1369-1385. doi: 10.1145/3593013.3594087
24. N. G. Leveson, Engineering a Safer World: Systems Thinking Applied to Safety. Cambridge, MA: MIT Press, 2012.
25. D. Manzey, J. Reichenbach, and L. Onnasch, "Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience," J. Cognitive Engineering and Decision Making, vol. 6, no. 1, pp. 57-87, 2012. doi: 10.1177/1555343411433844
26. National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, 2023. doi: 10.6028/NIST.AI.100-1
27. L. Onnasch, C. D. Wickens, H. Li, and D. Manzey, "Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis," Human Factors, vol. 56, no. 3, pp. 476-488, 2014. doi: 10.1177/0018720813501549
28. R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," IEEE Trans. Systems, Man, and Cybernetics A, vol. 30, no. 3, pp. 286-297, 2000. doi: 10.1109/3468.844354
29. S. Passi, "Agentic AI has a Human Oversight Problem," SSRN preprint 5529058, Sept. 2025. doi: 10.2139/ssrn.5529058
30. J. Rasmussen, "Risk management in a dynamic society: A modelling problem," Safety Science, vol. 27, no. 2, pp. 183-213, 1997. doi: 10.1016/S0925-7535(97)00052-0
31. Y. Shavit, S. Agarwal, M. Brundage, et al., "Practices for Governing Agentic AI Systems," OpenAI, 2023. <https://openai.com/research/practices-for-governing-agentic-ai-systems>
32. Stanford HAI, "What is Agentic AI?," Stanford Institute for Human-Centered AI, 2026. <https://hai.stanford.edu/ai-definitions/what-is-agentic-ai>
33. D. Tinkoff, L. Van der Veken, M. Chui, and T. Balakrishnan, "The state of AI in 2026: On the road to ROI," McKinsey & Company / QuantumBlack, 2026.

Backup

Kitchen Renovation Analogy

HOMEOWNER A *Decisions in Front of You. You Stay in the Loop.*

ACTION SURFACE: SCOPED
What the contractor can decide.

Demo order Adhesive type

Wiring route Installation method

EXECUTION AUTHORITY: APPROVE-IN-ADVANCE
Work waits for your sign-off.

Default: No action You must approve to proceed.

COUPLING CADENCE: TIGHT
You see the work often.

Plan & Select Demo Walkthrough Rough-in Check Pre-Finish Walkthrough Final Sign-off

WHAT HAPPENS

You catch issues early, understand the tradeoffs, and build **calibrated reliance**. You don't need to be an expert — the process puts decisions and feedback in front of you.

Architecture builds **calibrated reliance** because it creates opportunities to evaluate.

HOMEOWNER B *Agree on the Plan. Work Moves Without You.*

ACTION SURFACE: BROAD
What the contractor can decide.

Demo order Adhesive type

Wiring route Installation method

Swap materials Shift brand or supplier

Adjust layout Change a finish or color

Choose upgrades within the budget

EXECUTION AUTHORITY: ACT-UNLESS-OVERRIDDEN
Work moves forward by default.

Default: Action happens You can intervene, but only if you notice in time.

COUPLING CADENCE: LOOSE
You see it at milestones.

Plan & Budget Agreed Mid-Project Check-in Final Walkthrough

WHAT HAPPENS

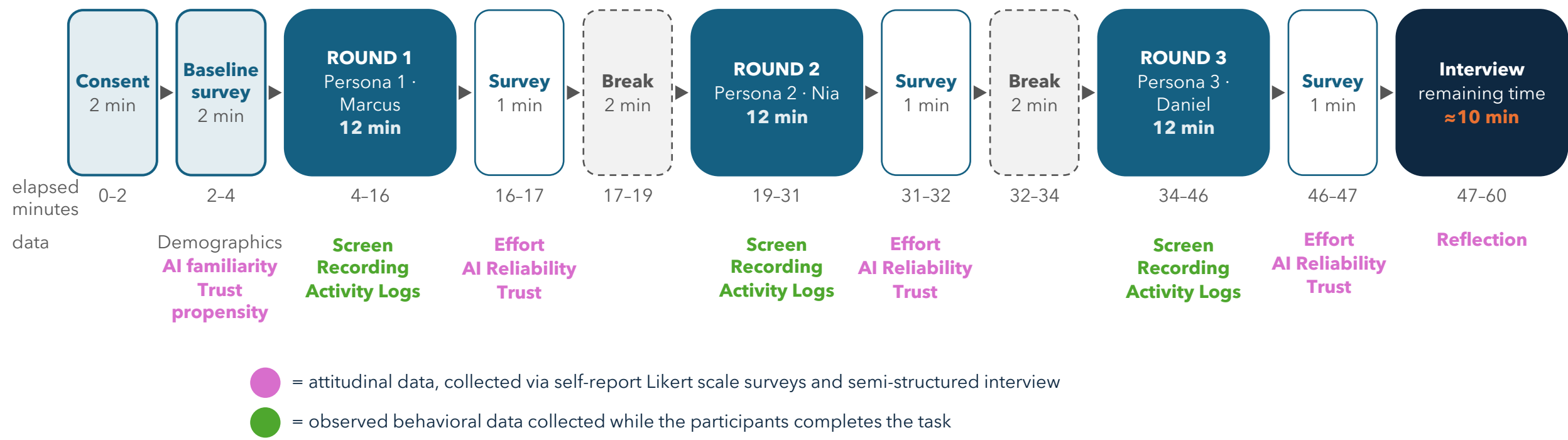
Work moves fast, but problems can hide behind the drywall. You're left guessing what's solid and what's not.

Architecture assumes **calibration** will happen on its own — but gaps stay hidden.

Participant flow: 60-minute in-person sessions

STUDY DESIGN

Participants complete the same set of personas in the same order each time, only the coupling granularity (tight vs loose) changes between subjects



Each student has a profile with medium level complexity

STUDENT PERSONAS

Student A

- Name: **Marcus Bennett**
- Medium complexity (Round 1)
- Rising-junior ME (BS), good standing, GPA 3.8
- On track through four semesters, lower-level sequence complete through Thermodynamics
- Ready to start the upper-level MAE core (no complications)

Student B

- Name: **Nia Coleman**
- Medium complexity (Round 2)
- Rising-junior ME (BS), good standing, GPA 3.4
- Switched into ME sophomore year, sequence runs about a semester behind her class year
- Still needs Analytical Mechanics II, Physics II, and Multivariable Calc

Student C

- Name: **Daniel Okonkwo**
- Medium complexity (Round 3)
- Rising-junior ME (BS), good standing, GPA 3.2
- Midway through the program, moving into the upper-level MAE core
- Deferred Linear Algebra I (MATH 2184) and still needs to fit it in

All three are rising junior B.S. Mechanical Engineering majors building Fall 2026 Schedule

Scoring detectability: how hard is each seeded error to catch?

Detectability = the baseline likelihood that an attentive participant catches the error from what the interface shows. Each seeded error is scored a priori on three dimensions (1 = more detectable → 3 = less detectable):

Scoring dimension	1 – easy to catch	2	3 – hard to catch
D1 • Cross-reference load	1 source: visible in the proposed action or audit alone	2 sources: must compare two fields or screens	3+ sources, or sustained tracking across steps
D2 • Knowledge / memory	On-screen contradiction; no outside knowledge needed	Requires catalog / policy knowledge not shown	Requires recalling earlier-stated info or a prior step
D3 • Salience / plausibility	Obvious anomaly that jumps out	Moderate	Plausible-but-wrong; the action looks reasonable

Composite = D1 + D2 + D3 (range 3-9), binned into tiers:

HIGH • composite 3-4
easy to catch

MED • composite 5-6
moderate

LOW • composite 7-9
hard to catch – plausible-but-wrong

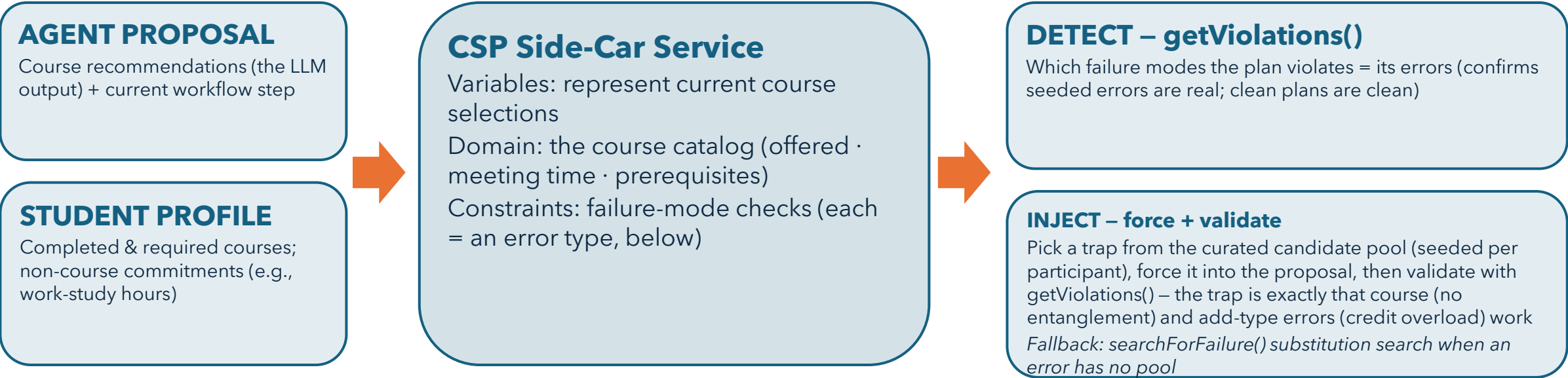
Course Registration Workflow: Seeded Errors by Round

Persona	list_candidates	verify_offered	verify_schedule	verify_prereqs	submit
Persona 1	–	–	Timing overlap (HIGH) Pool: {CE 2220 + CE 3604}, {APSC 2114 + MATH 3740} – pair shares a slot; verify falsely reports no conflict	Missing prereq (MED) Pool: MAE 3120 / CE 3250 / MAE 3155 – prereqs not completed; verify falsely passes it	–
Persona 2	Hallucinated course (HIGH) Pool of 5 fabricated codes (MAE 2010 / 2410 / 3210 / 3450 / 2230) – realistic credits + meeting time	Wrong semester (HIGH) Pool: MAE 2131 / MAE 3128 / CE 2210 – Spring-only; verify falsely marks it offered in Fall	–	Missing prereq (MED) Pool: MAE 3126 / APSC 2114 / MAE 3187 – prereqs not completed; verify falsely passes it	–
Persona 3	Repeat credit (MED) Pool: CE 2220 / MAE 2117 / APSC 2058 – already completed; proposal-level (no falsify)	–	Timing overlap (HIGH) MAE 3162 + MAE 3166W – both MWF 2:00; verify falsely reports no conflict	–	Credit overload (MED) Pool: {MAE 3145 + APSC 3115}, {CE 3720 + MATH 3410} – +6 cr over 18; submit shows it with no warning

Cell color = a-priori detectability tier: green HIGH · amber MED · red LOW

The error-harness CSP: how agent errors are defined, injected, and checked

RESEARCH SETTING

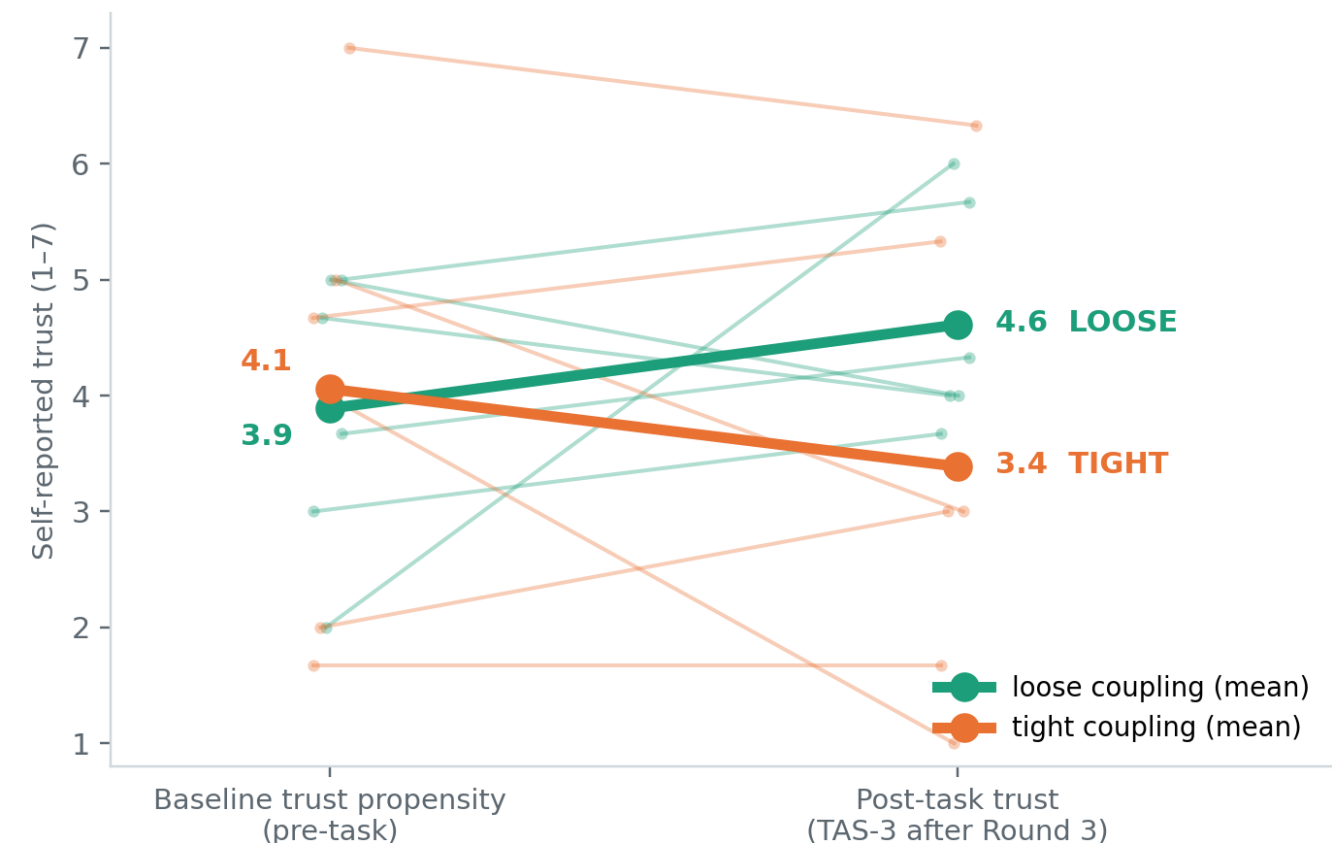


Constraint set: each constraint defines one seeded-error type

Missing prerequisite	Missing corequisite	Already-earned credit	Hallucinated course
Wrong semester (not offered)	Course-course time conflict	Conflict with non-course time	Credit overload (> 18)

Stated trust moved opposite to performance

Self-reported trust (TAS-3 after each round, 1-7) was collected as a covariate and moved in the opposite direction from performance.



Thin lines: individual participants (6 per condition); thick lines: condition means. Baseline propensity = three items adapted from McKnight et al. 2002; post-task = TAS-3 short form after Round 3.

Groups started equivalent.

Baseline trust propensity 4.1 (tight) vs 3.9 (loose): no pre-differentiation on the attitude most likely to drive reliance.

Trust moved opposite to performance.

Post-task trust 3.4 under tight vs 4.6 under loose (averaged over the three rounds: 3.8 vs 4.4). The group that caught more errors (88% vs 69%) came away trusting the agent less.

Stated trust is a misleading design signal.

A team choosing between these designs on post-task trust or satisfaction would pick the one that produced worse oversight, extending Bućinca et al.'s warning about subjective measures to agentic systems.