



The MESH Environment: Multi-level Evaluation of Sociotechnical HAI-workflows

Zoe Szajnfarter, Sarah Starr, Bryce Huffman and
Ryan Watkins

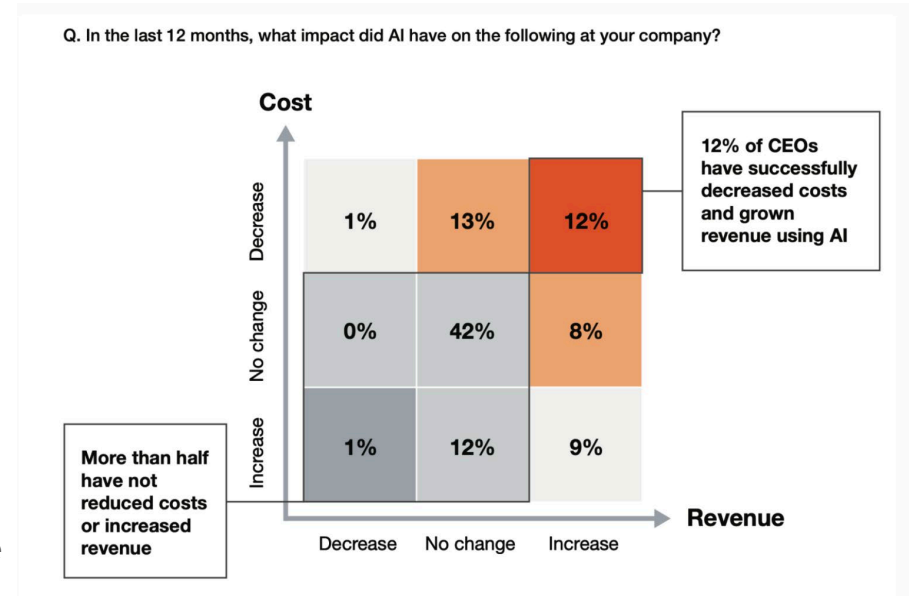
The George Washington University

Funding acknowledgement: (1) NSF NRT-AI-FW-HTF: Co-Design of Trustworthy AI in Systems (DTAIS), (2) the NIST-NSF Institute for Trustworthy AI in Law and Society (TRAILS) (3) GW TAI



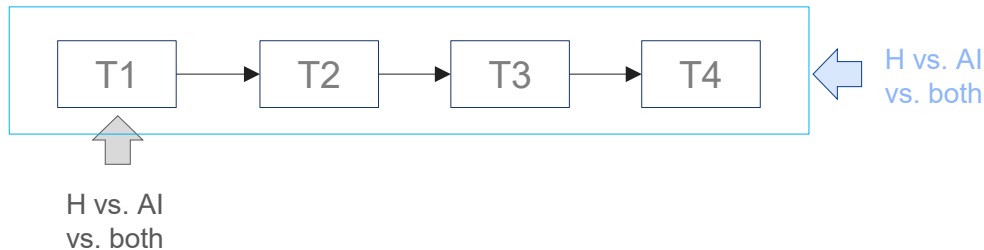
Motivation: Why we need a new kind of testbed

- Despite wide “adoption” most AI-adoption is failing to generate enterprise value.
 - **Individual** => work system
 - **Complex change in work**
 - **Socio-technical factors not being captured pre-deployment**
- Need a way to **envision, test and scaffold** joint technology and workforce transformation.

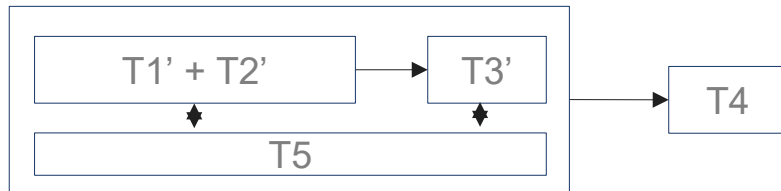


Core issue: more than a process automation problem

Typical view:



Restructured view:



Why might this matter from a testbed perspective?

- Model eval focuses on the input-process-output from the existing process, so even just getting data from the re-architected process can be difficult.
- Human experience can change significantly (and unpredictably) which can impact overall process outcomes.

Our Solution: The MESH Environment

Reference architecture: Reusable Generic Infrastructure

Six reusable computational and physical layers that are needed for any sociotechnical evaluation

- | | | |
|---|---------------------------------|--|
| 1 | DATA HANDLING | stores and protects data with different levels of access based on actor type |
| 2 | TASK SCOPE | templates each task, which can be assigned to an actor, with an objective, completion criteria, handoffs, and tool/data access |
| 3 | ACTORS AND USERS | representative subjects, simulated agents, or AI models |
| 4 | INFORMATION PATHWAYS | three classes to transfer information: actor to actor, one-way retrieval, and public broadcast (i.e., shared with everyone) |
| 5 | OBSERVATION | built-in sensors to track behavior and outputs in the environment (e.g., telemetry, voice, and survey instruments) |
| 6 | EXPERIMENT ORCHESTRATION | connects the elements into workflow architectures |

Abstract covers current implementation details
(U.S. Provisional Patent Application No. 64/091,886)

VISTA process

Step 1:

Extract contextual workflows

Step 2:

Design evaluation for stated objective

Step 3:

Specify the workflow and task environment

Step 4:

Build the VISTA (UX) in MESH

Step 5:

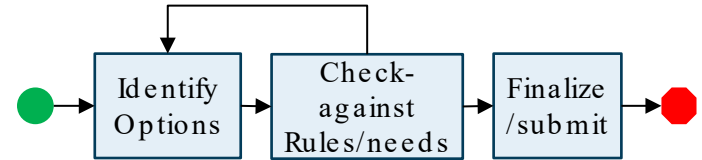
Customize sensors and data collection as needed

Illustrative Use Case: Academic Advising

"I pursued a career in student affairs because of my passion for helping students...
Currently, I spend a significant amount of my time doing literal paperwork"



Specific Workflow: Semesterly course registration



Potential actors:

- Students, professional advisors, faculty advisors, AI chatbots and/or agents

Potential "tools":

- Open websites (Bulletin), Student-specific databases (DegreeMap), Chatbots, advisors.

Potential channels:

- In-person, email, chat, one-way retrieval

AI can help, but how?

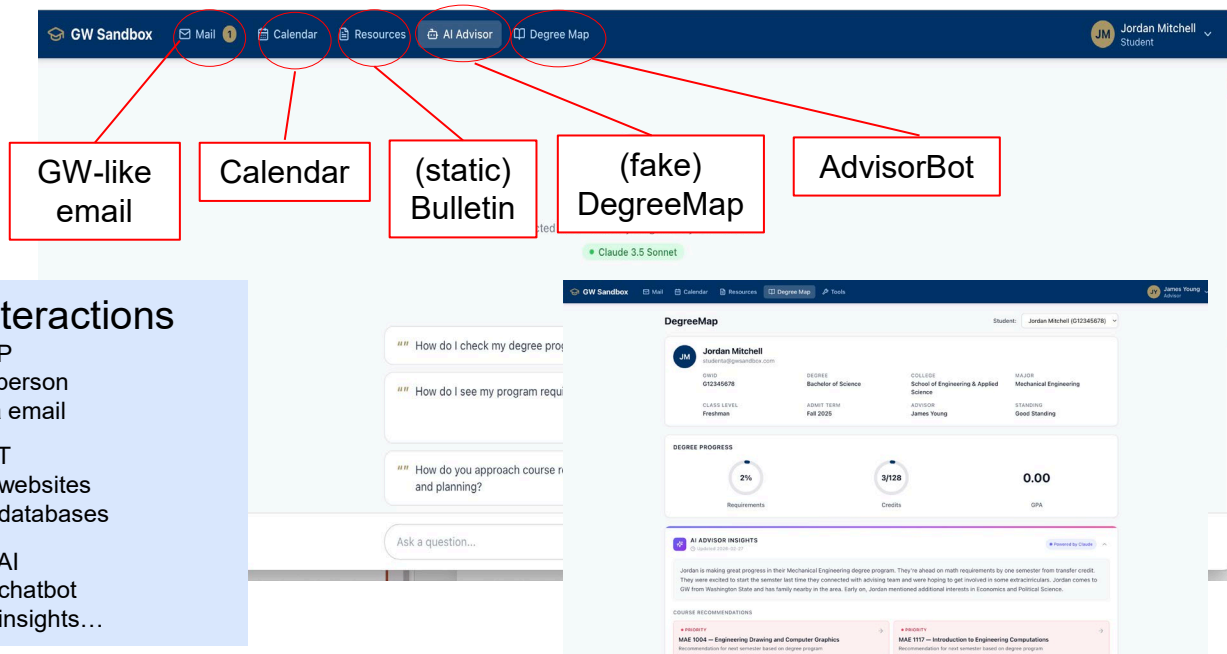
Images generated by Gemini 2.5 Flash

MESH+VISTA environment build-out

2-location Physical Space



Virtual space for technology-mediated interactions



Interactions

P2P
In-person
Via email

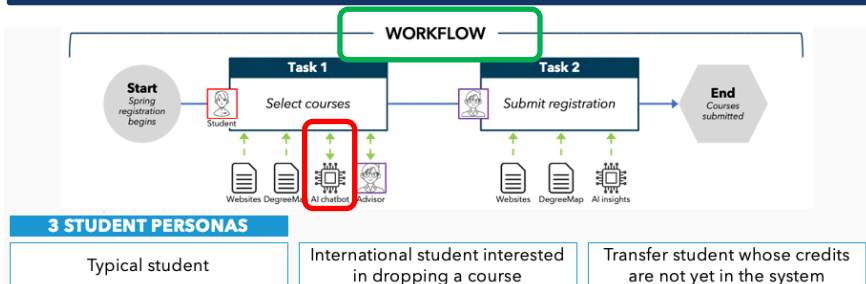
P2T
w/ websites
w/ databases

P2AI
w/ chatbot
w/ insights...

Observational data: virtual platform telemetry,
content of interactions, custom surveys and check-ins.

Is MESH useful?

Tech Pilot: Does MESH “see” differently?



Conducted evaluations at three levels

System-level evaluation in the field

unit under test: submitted registration (output)

Correctness of the submitted registrations

All three correct

Field testing at model-level

unit under test: AI chatbot

78 chatbot interactions observed across the pilot run

At least 6 AI failures

hallucinated course requirements, factually correct but contextually incorrect responses, and inconsistent responses

Pre-deployment benchmarking

unit under test: AI chatbot

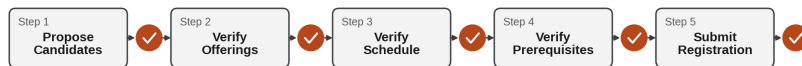
Test set of 30 questions with ground-truth answers

100% accurate

on the fact-based test cases compared to ground-truth

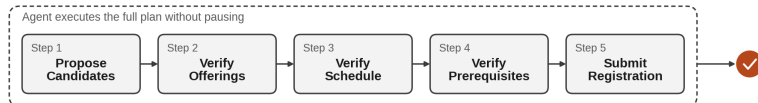
Research Pilot: Does “coupling” change reliance?

(a) Tight coupling: per-action approval, five checkpoints per trial



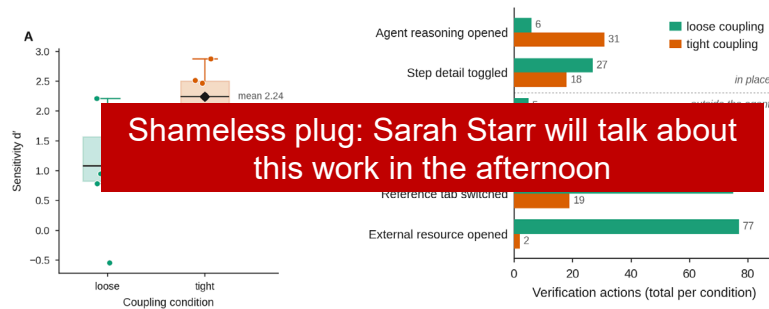
The agent pauses after every step; the participant approves, edits, or vetoes that step before the agent continues.

(b) Loose coupling: plan-level approval, one checkpoint per trial



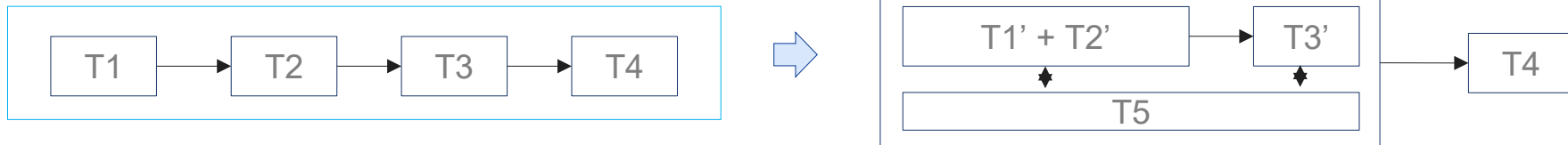
The completed five-step plan is presented once; the participant approves, edits, or vetoes the plan as a whole.

Approval checkpoint: the agent cannot proceed or commit until the participant responds. Workflow steps, agent, and seeded errors are identical in both conditions.



Conclusion and next steps

- Set out to solve two related problems:



- (1) Current evaluation overfocuses on model level (misses socio-technical dynamics)
 - (2) Need basis for thinking through meaningful process re-engineering including difficult to predict impact on human roles.
- Presented the MESH reference architecture and VISTA build out process as a solution.
 - Demonstrated value through two pilots
 - Next steps...
 - > rigidize platform
 - > onboard external problem owners

Questions or Comments

zszajnfa@gwu.edu

