WHAT NEEDS ATTENTION?
PRIORITIZING DRIVERS OF DEVELOPERS' TRUST AND ADOPTION OF GENAI

Rudrajit Choudhuri /ROOD-ruh-jit CHOH-dur-ee/

PhD Student & MOSIP Research Fellow

Oregon State University College of Engineering | Microsoft Research

Generative AI is revolutionizing SE

COMMUNICATIONS OF THE ACM

The Paradigm Shifts in Artificial Intelligence

Even as we celebrate AI as a technology that will have far-reaching benefits for humanity, trust and alignment remain disconcertingly unaddressed.

Survey: The AI wave continues to grow on software development teams

TECH

Google CEO says more than a quarter of the company's new code is created by AI

Hugh Langley Oct 29, 2024, 9:36 PM UTC Share Save

Trust in AI? *What does it mean? Why does it matter?*

"the attitude that an AI agent will achieve an individual's goals in a situation characterized by uncertainty and vulnerability"

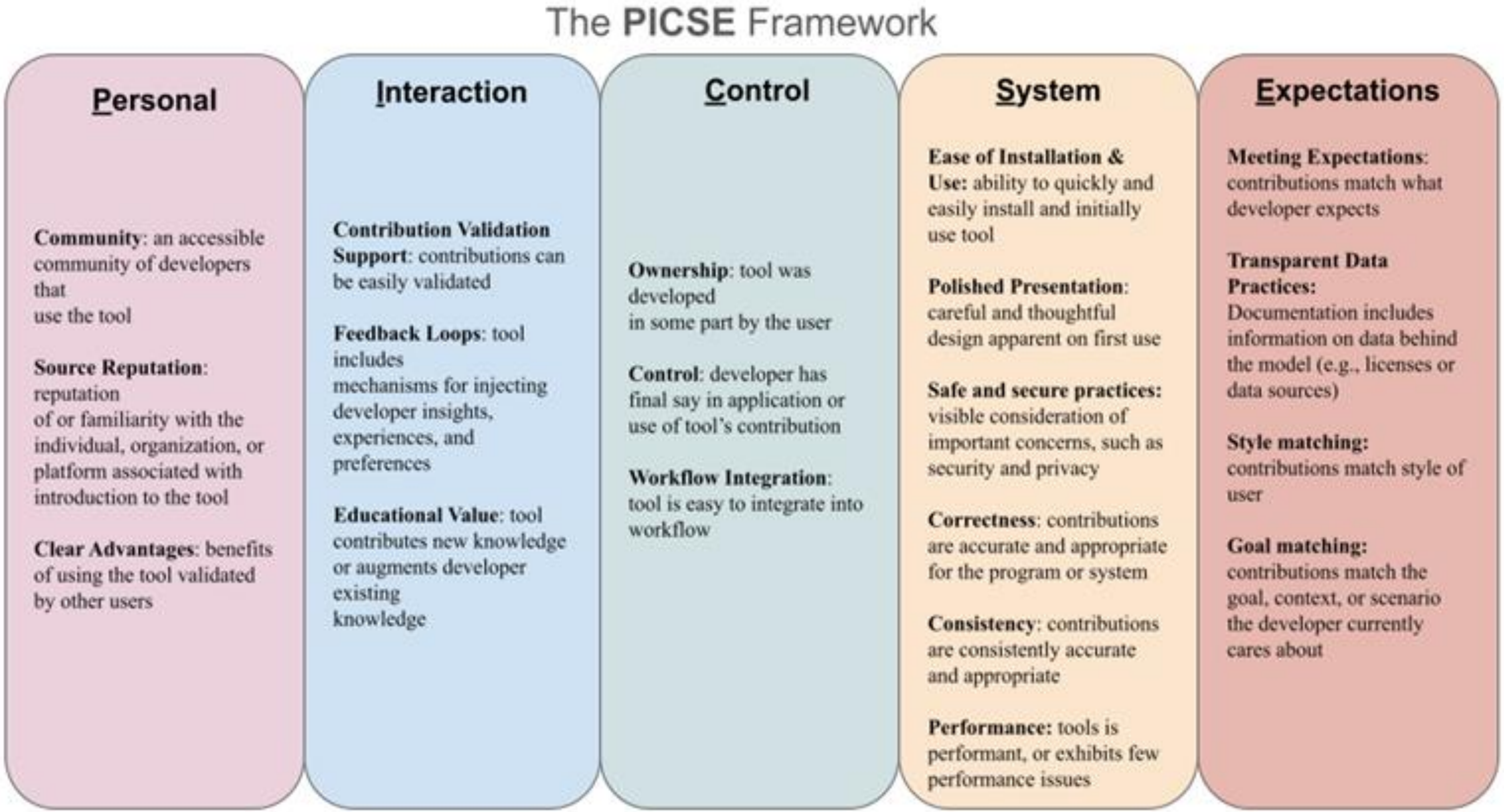
Trust being an ***attitude*** is a psychological construct that is not directly observable & should be:

- captured through *psychometrically validated instruments*
- distinguished from observable measures *such as reliance*

A *foundational design requirement for supporting effective human-AI interactions:*

- Miscalibrated levels of trust can lead developers to:
  - Overlook AI-induced errors and risks in work
  - Eschew its use altogether

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50-80.



Johnson, B., Bird, C., Ford, D., Forsgren, N., & Zimmermann, T. (2023, May). Make your tools sparkle with trust: The PICSE framework for trust in software tools. ICSE-SEIP (pp. 409-419). IEEE.

What to prioritize in tool design for trust?

It is important to establish an understanding of

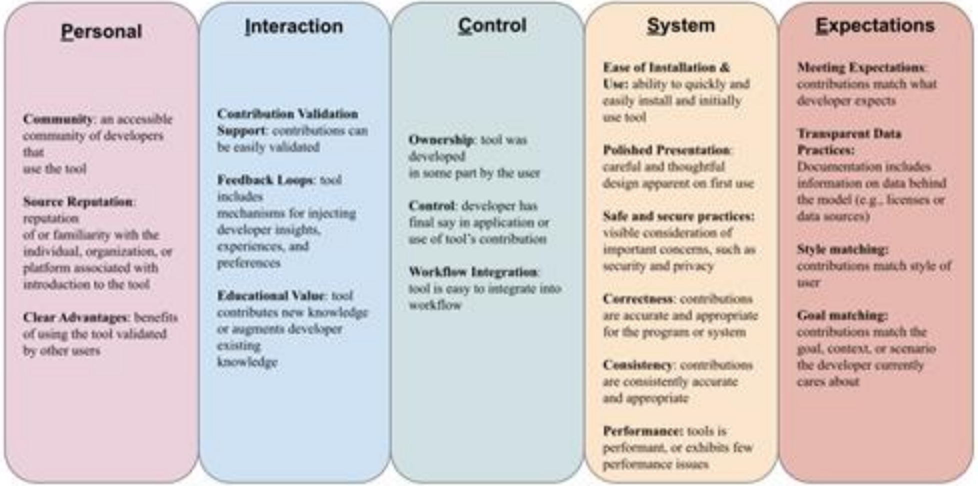
RQ1: How these *multitude of factors* affect developers' *trust in genAI tools*?

- A validated instrument for
  - capturing different trust-RELATED factors in human-genAI interaction contexts
    - through a psychometric analysis of the PICSE framework
- The strength & significance of these factors' association with developers' trust in genAI tools

Survey with software developers (N=238) at GitHub Inc. & Microsoft

Psychometric Analysis

Psychometric quality refers to the objectivity, reliability, and validity of an instrument



- PICSE was qualitatively developed → Its psychometric quality had not been assessed
- We conducted psychometric analysis of the framework to empirically:
  - refine its factor groupings,
  - which were then evaluated for their association with trust

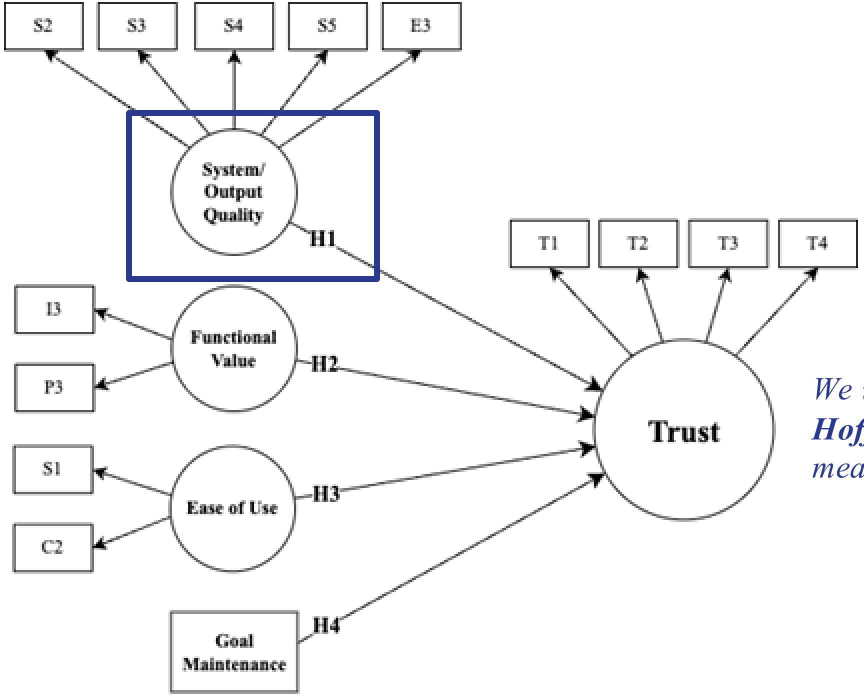
Validated PICSE Instrument *for capturing trust-RELATED factors in HAI context*

Construct	ID	Items
System/Output Quality	S2	Presentation/Interaction design
	S3	Safety/security practices
	S4	Consistent contextual accuracy
	S5	Performance in tasks
	E3	Style matching of contributions

Building the structural (theoretical) model

Construct	ID
	I3
Functional Value	P3
	S1
Ease of use	C2
Goal Maintenance	E4

H1: System/Output quality of genAI is positively associated with developers' trust in these tools.



*We used the validated **Hoffman's TXAI** to measure the **trust construct***

* *Note: TXAI is*

(a) derived from validated trust scales specifically for HAI interactions,
(b) psychometrically validated, and (c) is widely used to capture the trust construct.

The exclusionary nature of AI

often fails to support all users adequately

Design and Human-Computer Interaction, Language Processing, Machine Learning

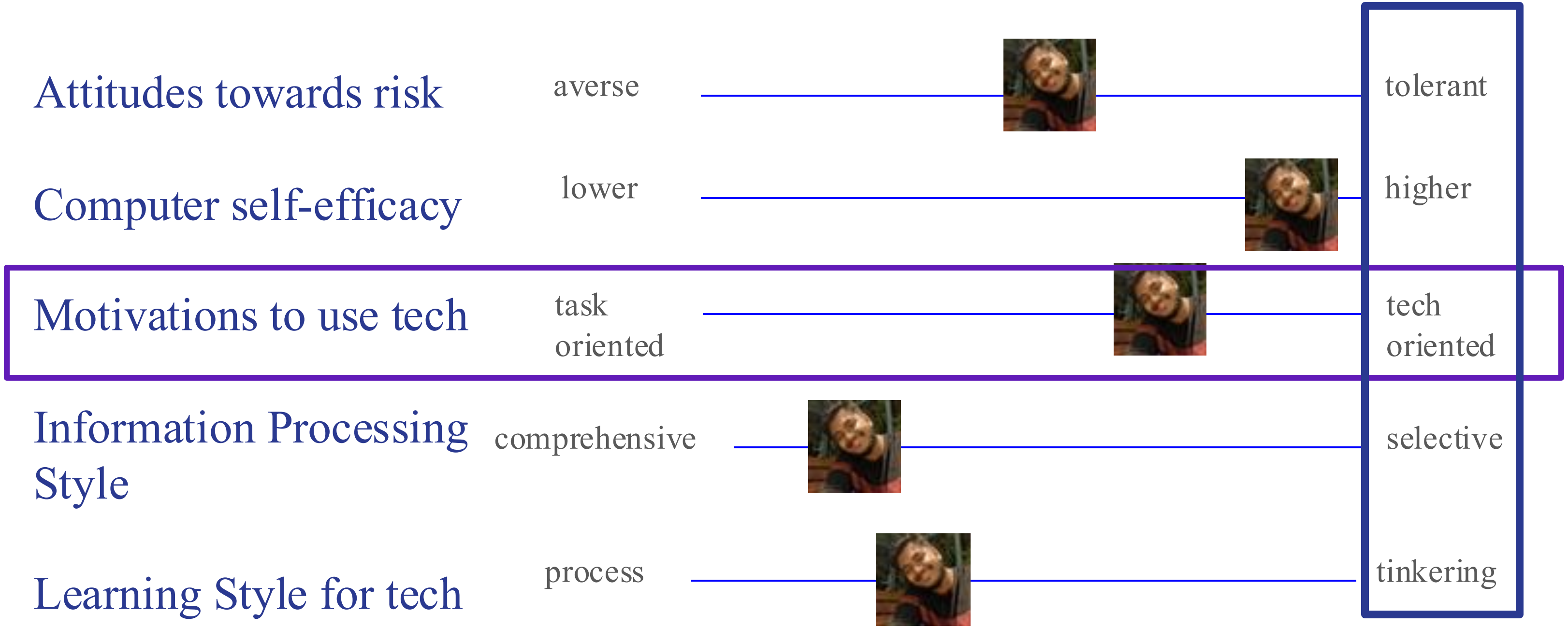
AI-Detectors Biased Against Non-Native English Writers

- Substantial body of work exists in modeling technology adoption,
    - These studies *don't* consider the inclusivity of the software design
    - One such aspect of inclusivity is supporting **cognitive diversity**:
        - Fosters ***divergence in perceptions and interaction styles with technology***
        - No particular style is inherently better or worse
    - When an user's cognitive style is unsupported (or misaligned) by software:
        - Additional "cognitive tax" everytime they use that software
        - Additional barriers to usage and adoption

RQ2: How are developers' *trust and cognitive styles* associated with their *intentions to use genAI tools*?

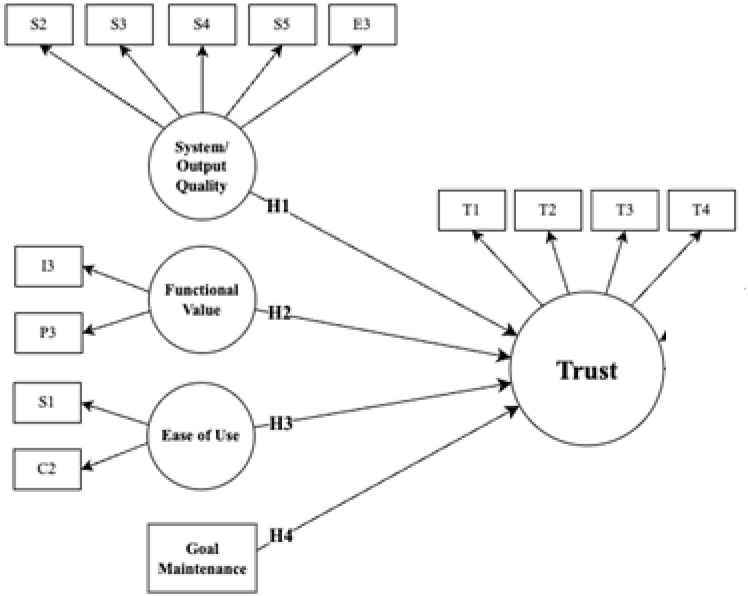
Cognitive Diversity, *i.e. variations in cognitive styles*

diverse ways users perceive, process, and interact with information & technology, as well as their approach to problem-solving



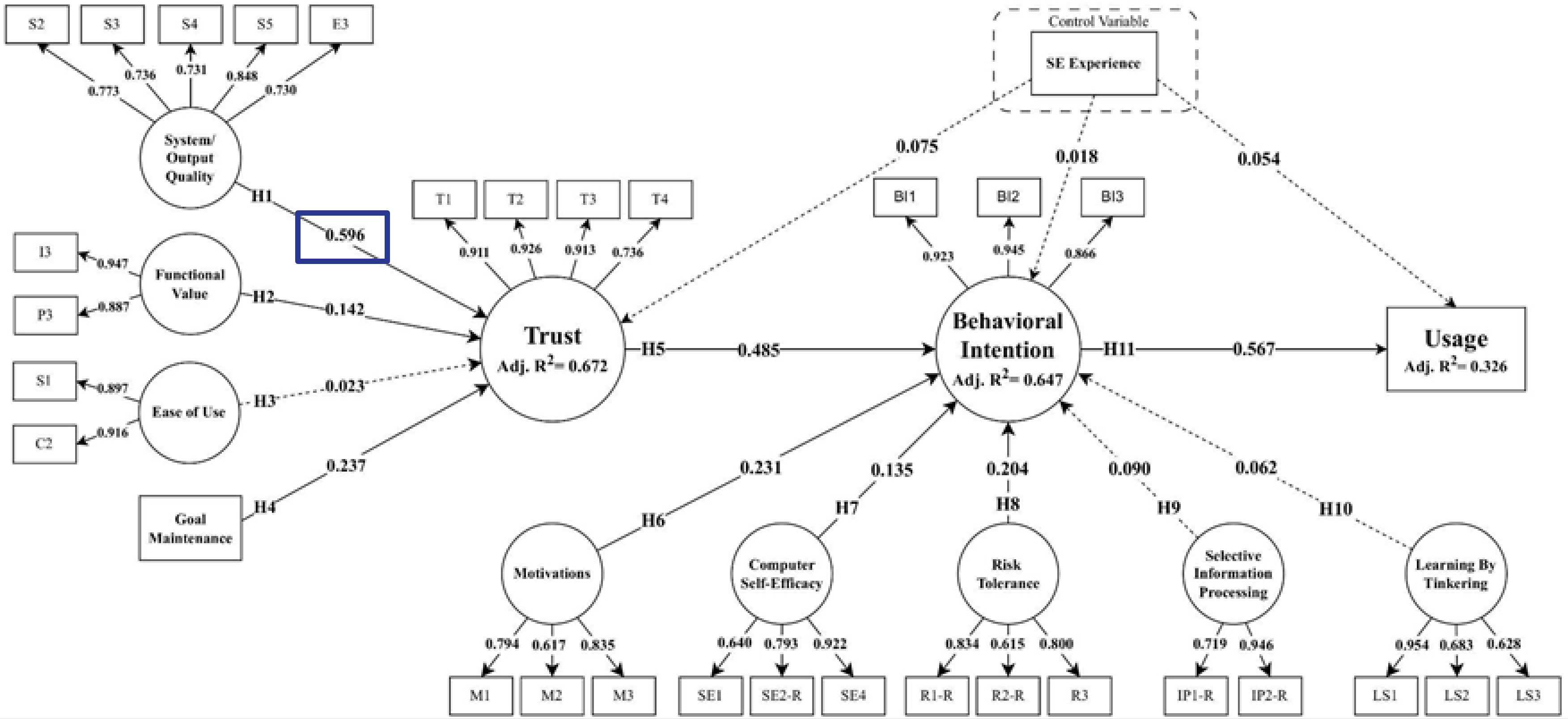
Burnett, M., Stumpf, S., Macbeth, J., Makri, S., Beckwith, L., Kwan, I., ... & Jernigan, W. (2016). GenderMag: A method for evaluating software's gender inclusiveness. *Interacting with computers*, *28*(6), 760-787.

Building the structural (theoretical) model (contd.)



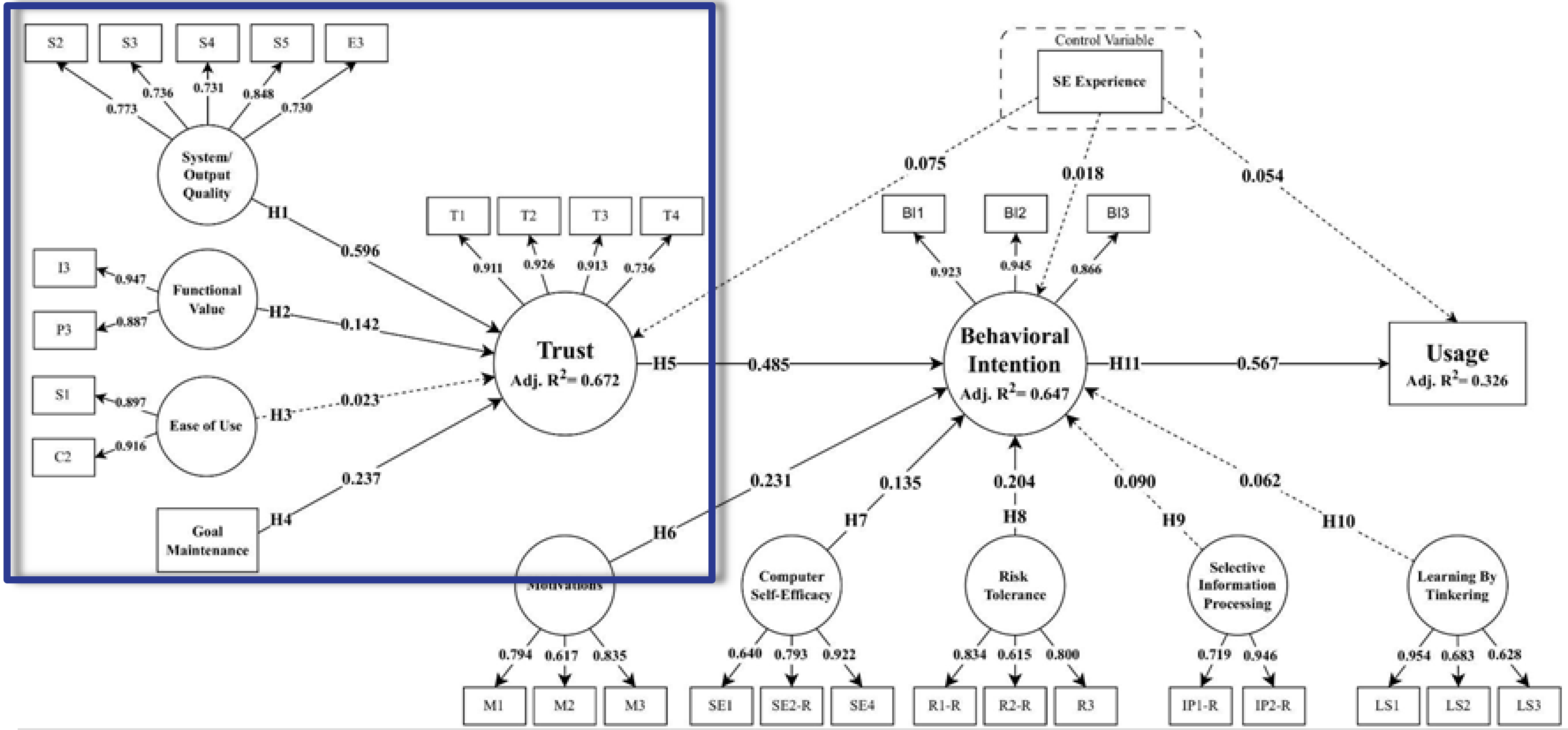
* *Note: We used the validated GenderMag facet survey to capture the five cognitive styles*
We used components of the UTAUT model to capture the behavioral intention and usage constructs

Structural Model (PLS-SEM) – What matters?



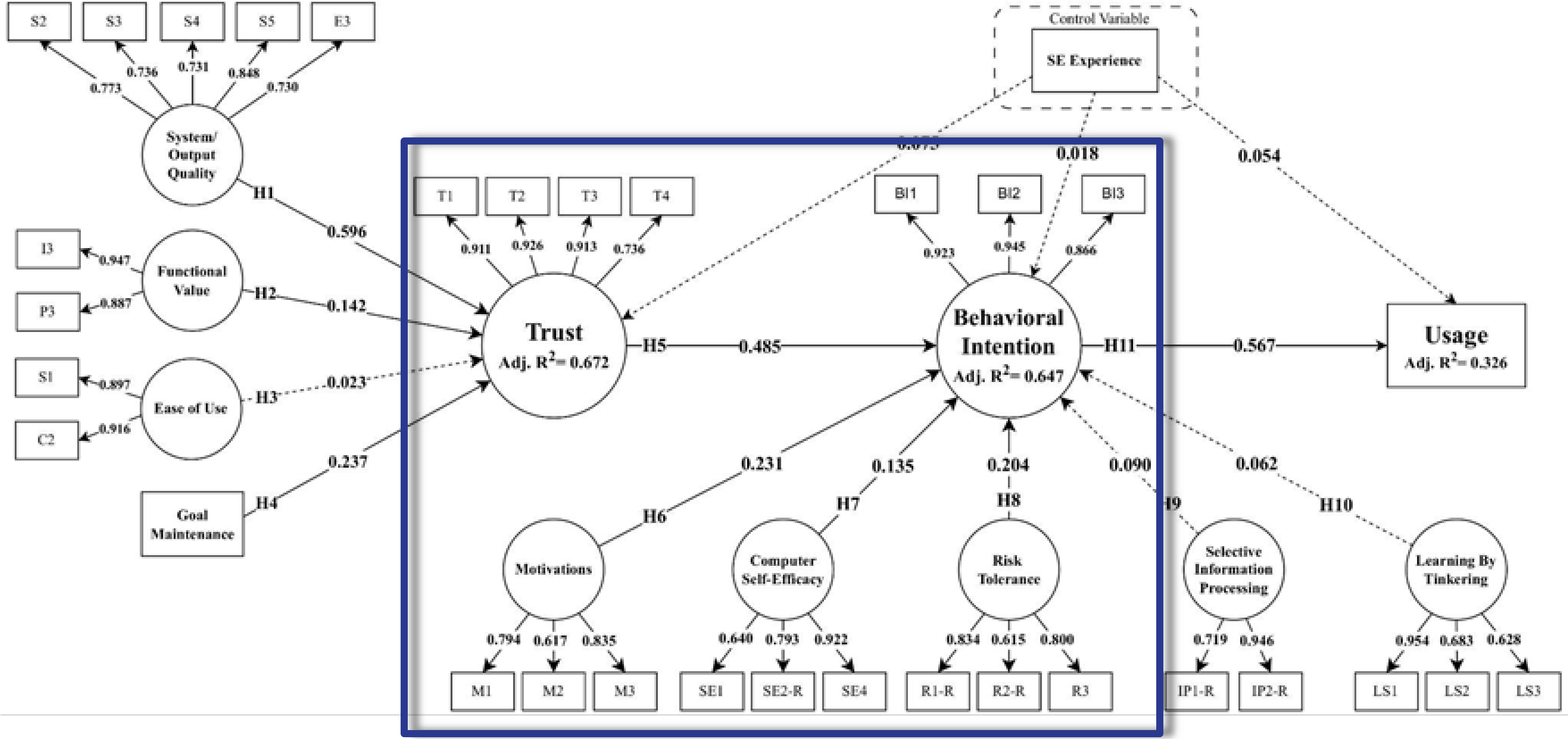
PLS-SEM Model: Solid lines indicate item loadings and path coefficients ($p < 0.05$); dashed lines represent non-significant paths. Reverse-coded items are suffixed with '-R' (e.g., SE2-R). Latent constructs are depicted as circles and adjusted R^2 (Adj. R^2) values are reported for endogenous constructs

Factors associated with trust (RQ1)



PLS-SEM Model: Solid lines indicate item loadings and path coefficients ($p < 0.05$); dashed lines represent non-significant paths. Reverse-coded items are suffixed with '-R' (e.g., SE2-R). Latent constructs are depicted as circles and adjusted R^2 (Adj. R^2) values are reported for endogenous constructs

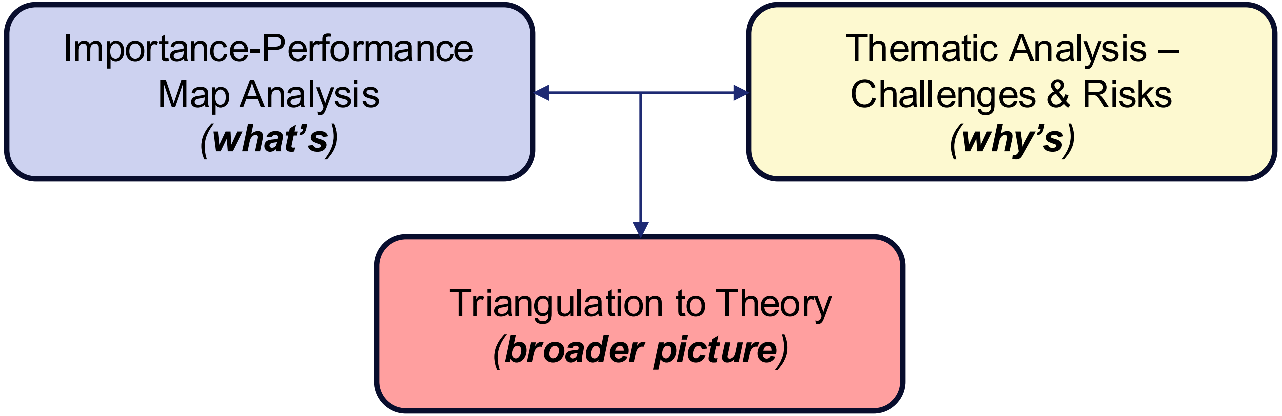
Factors associated with behavioral intentions (RQ2)



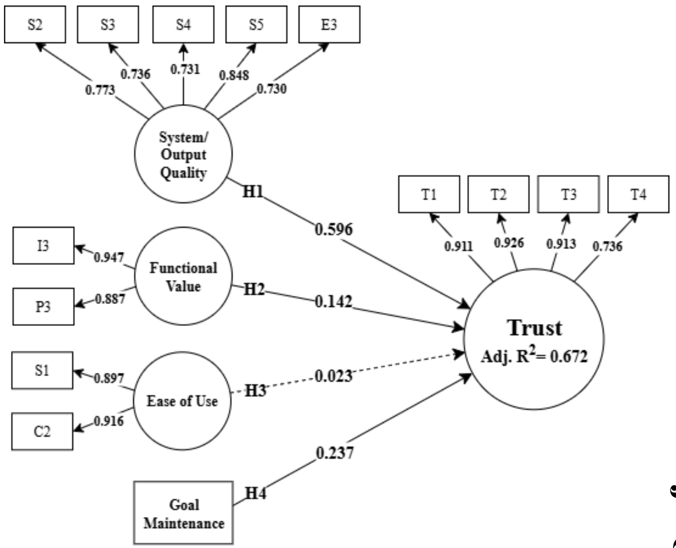
PLS-SEM Model: Solid lines indicate item loadings and path coefficients ($p < 0.05$); dashed lines represent non-significant paths. Reverse-coded items are suffixed with '-R' (e.g., SE2-R). Latent constructs are depicted as circles and adjusted R^2 (Adj. R^2) values are reported for endogenous constructs

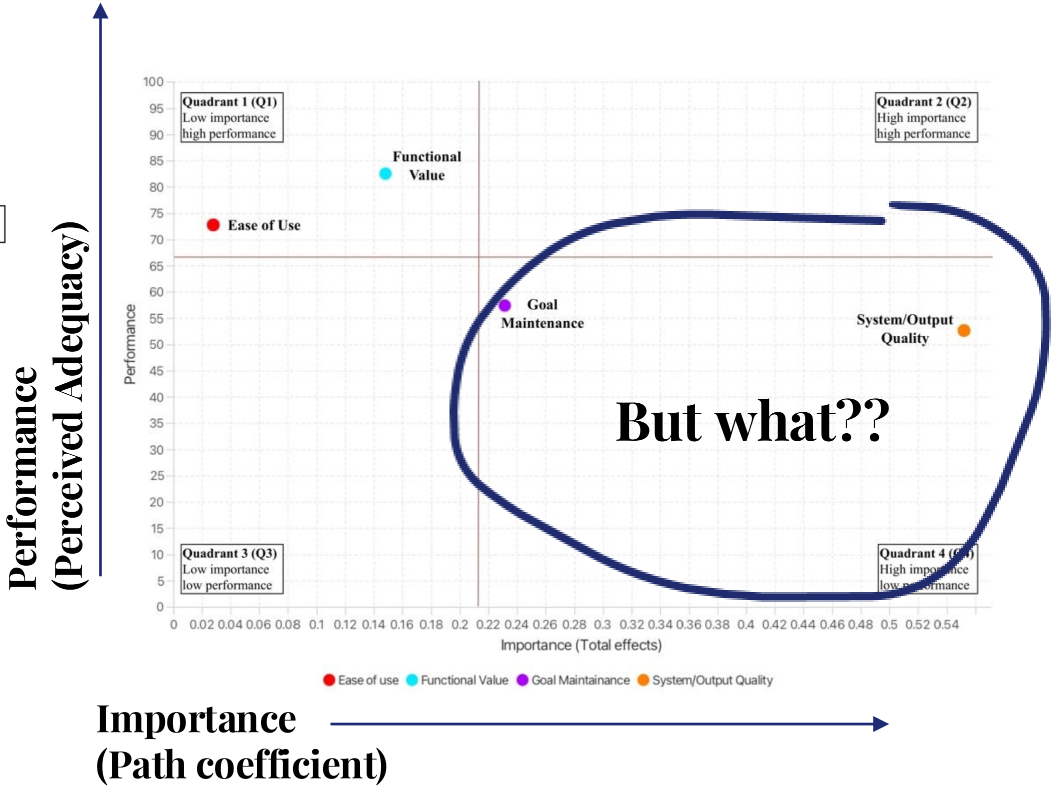
What Needs Attention and Why?

Prioritizing Drivers of Trust & Adoption of GenAI

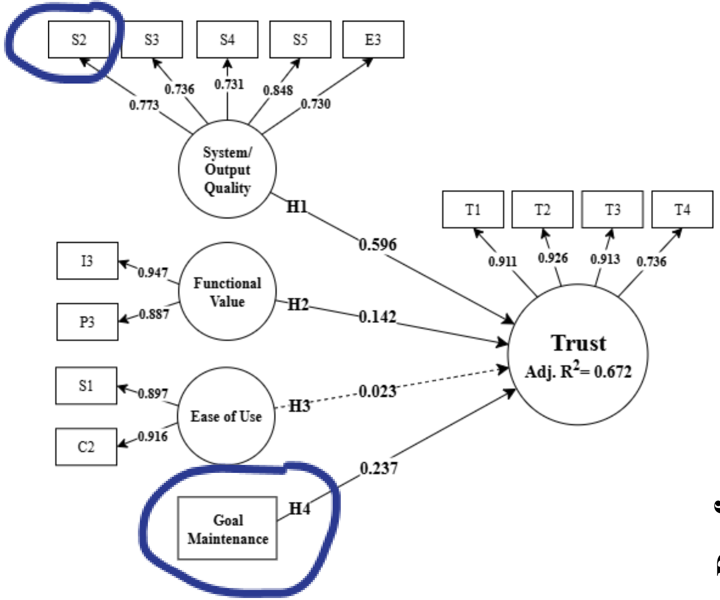


From what's important to what needs attention - Trust

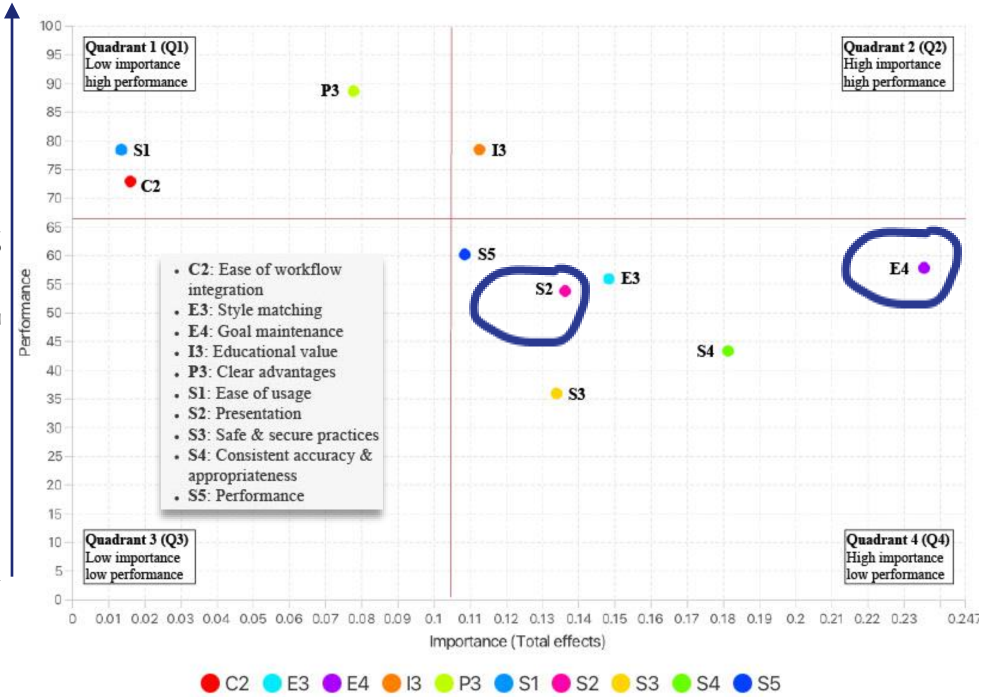




From what's important to what needs attention - Trust

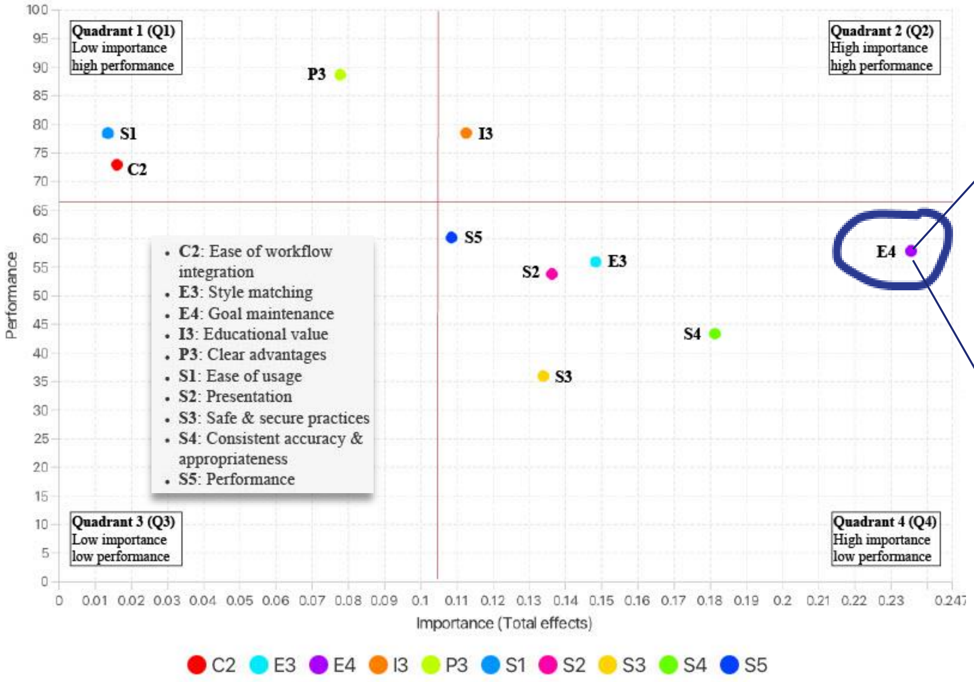


Performance (Perceived Adequacy)



Importance
(Path coefficient)

What's wrong with gAI's goal maintenance?



Misalignment with task objectives

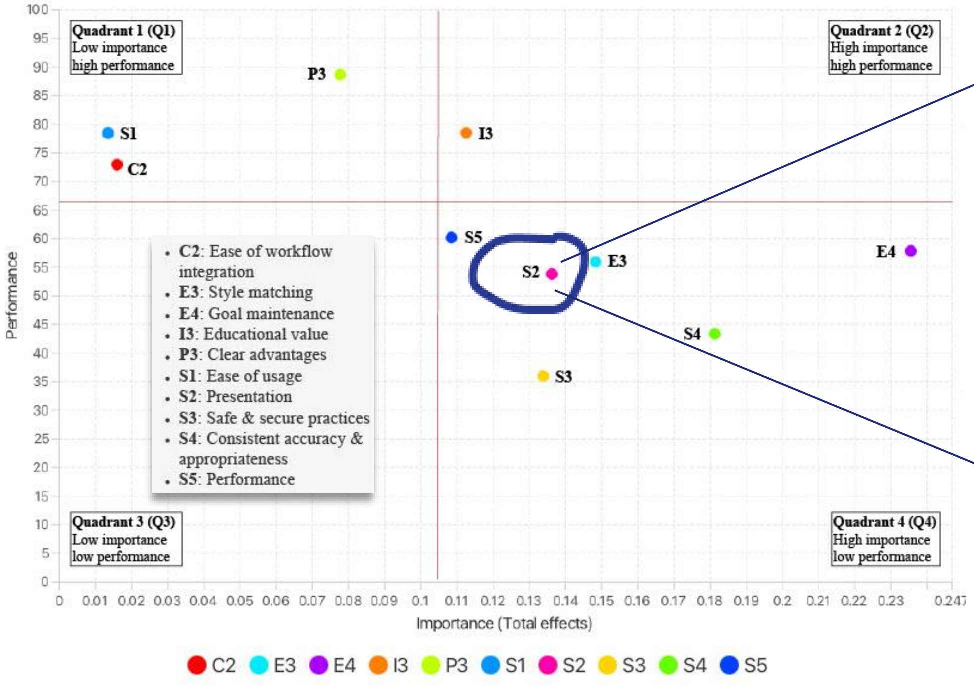
Verification burden

High effort in prompting

High effort in modifying AI responses

Loose coupling & coordination between genAI and dev's cognitive abilities
→ **limited *cognitive support* & induced *extrinsic cognitive load***

What's wrong with gAI's interaction design?



Poor feedback mechanisms/affordances

Constrained modes & interaction frictions

Excessive verbosity of outputs/contributions

"The interaction feels limiting, there's no easy way to organize information intuitively…it's hard to explore ideas using [dominant] chat" (P117)

Eroded <u>*competence*</u> & <u>*autonomy*</u>
→ ***uncertainty***, **reduced *agency, satisfaction, & trust***

And there is so much more (to fix)!

...

Safety & Security
(e.g., data handling, misinformation, ethical concerns)

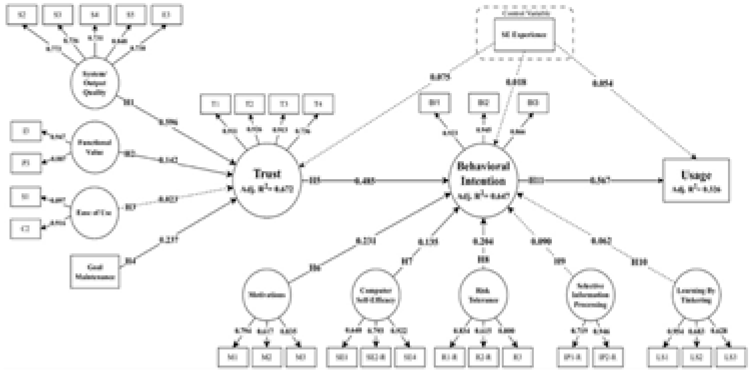
Accuracy & Appropriateness
(e.g., relevance, predictability, correctness)

Style matching
(e.g., task specific, overall)

Performance
(e.g., efficiency, error handling & recovery)

...

How can I (researcher/practitioner) use the work?



Use the model &/or the validated instrument to improve understanding of AI adoption dynamics

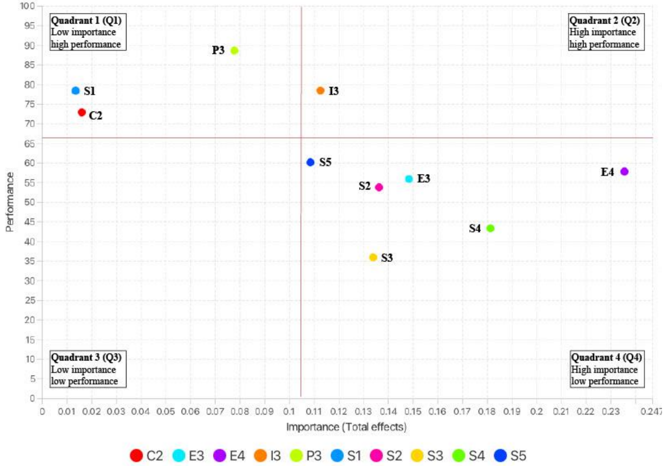
“Is this model tool-specific? How relevant is it in 2026, 27, 28,…”

Factors (read lenses) are based on dev-genAI interactions
- With tool improvements, expectations & perceptions co-evolve

Specifics: Prioritize goal maintenance, transparency, & agency
Overall: Guide design (improv.) with cognitive factors in mind;
→ Design for inclusive HAI-UX

“Is it a one-time thing? Can I design once and for all”

Tool-smithing & (re)design needs to co-evolve as well.
- Essential to build these tools to not only assist with tasks but also meaningfully support the people who use it.



THANK YOU!
QUESTIONS?

choudhru@oregonstate.edu

Rudrajit Choudhuri /ROOD-ruh-jit CHOH-dur-ee/
PhD Student & MOSIP Research Fellow

Oregon State University College of Engineering | Microsoft Research

Check out the paper

Check out my profile

Generative AI?

The term **"generative AI"** is broad, encompassing systems with different capabilities,

- core algorithm (such as the transformer),
- a particular instantiated model (such as GPT-4);
- a productised system comprised of an ensemble of multiple models together with prompt engineering, safety heuristics, and user interface affordances (such as ChatGPT).

So what is it **by definition**?

Generative AI refers to artificial intelligence systems that create new content, such as text, images, music, or code, by learning patterns from existing data.

For this study, when we talk about genAI, we are referring to it as a term for tools such as ChatGPT, Copilot, Claude, Gemini, etc.

i.e., a productised system comprised of an ensemble of multiple models together with prompt engineering, safety heuristics, and user interface affordances.

Sarkar, Advait. "Will Code Remain a Relevant User Interface for End-User Programming with Generative AI Models?." *ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*. 2023.

Consistent Accuracy and Appropriateness (S4)	(a) Lack of contextual appropriateness in outputs:
	(b) Incorrect or irrelevant outputs
	(c) Low predictability of output quality
Style Matching of AI Contributions (E3)	(a) Mismatch with task-specific or project styles (project settings, coding convention)
	(b) Mismatch with individual styles (Problem-solving/development style)
Presentation (S2)	(a) Poor feedback mechanism/unclear affordances (Prompt-Output traceability)
	(b) Constrained interaction modes
	(c) Excessive verbosity in outputs
Safe and Secure Practices (S3)	(a) Input data privacy risk: (Risks of data exposure or leakage, Limited transparency in data handling)
	(b) Misinformation risks
	(c) Legal/Ethical concerns
Performance of AI (S5)	(a) Efficiency issues in complex or niche tasks
	(b) Poor error handling and recovery mechanisms

Evaluating the model

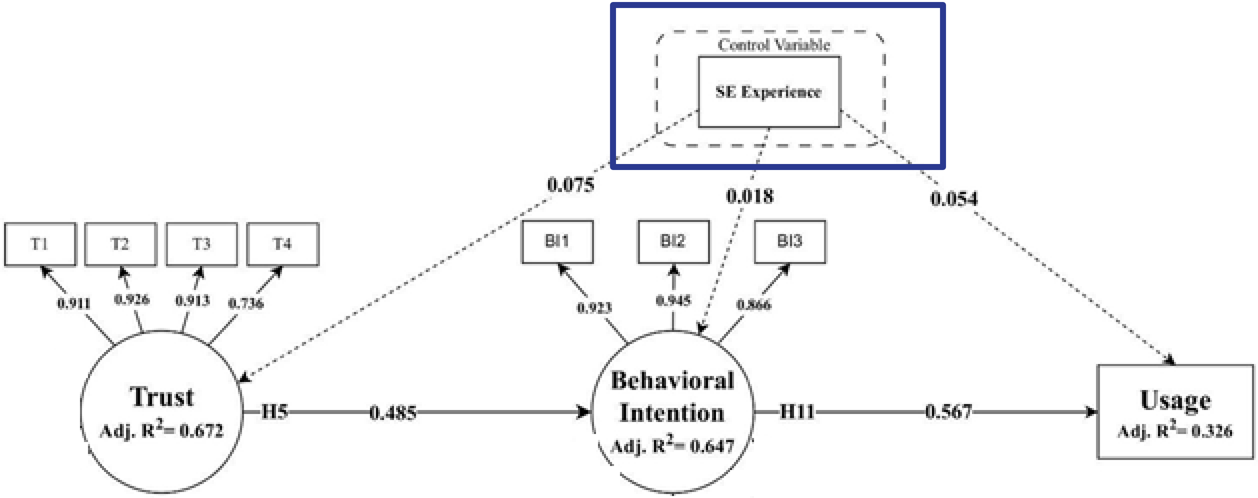
We used **Partial Least Squares-Structural Equation Modeling (PLS-SEM)** to test our theoretical model.

- PLS-SEM is a **second-generation multivariate data analysis technique** that has gained traction in empirical SE studies investigating complex phenomena
    - Allows for **simultaneous analysis of relationships among constructs** & **addresses multiple interconnected research queries** in one comprehensive analysis
    - Particularly **suited for exploratory studies** due to its **flexibility in handling model complexity** while accounting for measurement errors in latent variables (constructs)
    - **Does not require data to meet distributional assumptions.**
        - Instead, it uses a **bootstrapping approach** to determine the statistical significance of path coefficients (i.e., relationships between constructs)
        - The PLS path model is estimated for a large number of random subsamples (usually 5000), generating a bootstrap distribution, which is then used to make statistical inferences

PLS-SEM: Measurement Model Evaluation

Convergent Validity (AVE, Factor loadings)	Examines how a measure correlates with alternate measures of the same construct, focusing on the correlations between indicators (questions) and their corresponding construct.
Internal Consistency Reliability (Cronbach's α, Composite Reliability)	Examines that the indicators are consistent with one another and that they consistently and reliably measure the same construct.
Discriminant Validity (HTMT, FL Criterion)	Examines the distinctiveness of each construct in relation to the others (how different really is a construct compared to a different construct)
Collinearity Assessment (VIF)	Examines the correlation between predictor variables, ensuring they are independent to avoid potential bias in the model path estimations

Control variables



- Familiarity with genAI, was excluded as a control variable due to a highly skewed distribution of responses.
- We evaluated the model for detecting the presence of unobserved heterogeneity
  - Confirmed absence of any group differences in the model caused by unmeasured criteria

PLS-SEM: Structural Model Evaluation