Structural Pitfalls of Processing AI-generated Information

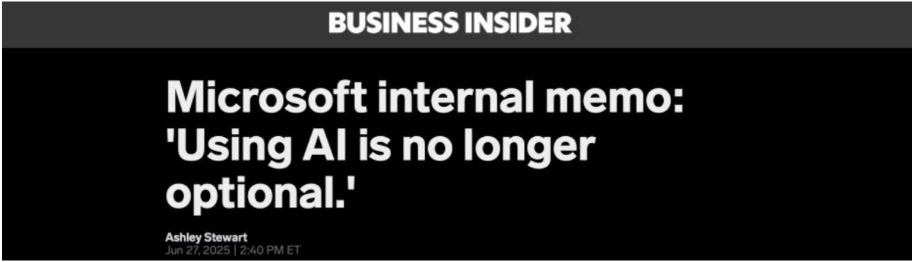
Jungpil Hahn
(with Junjie Zhou)

National University of Singapore

December 12, 2025

AI: Information Producer

- Organizations are increasingly faced with the need to process and act on information generated by AI systems



BUSINESS INSIDER

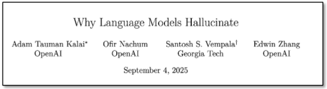
Microsoft internal memo: 'Using AI is no longer optional.'

Ashley Stewart
Jun 27, 2025 | 2:40 PM ET

Microsoft makes AI use mandatory for employees (July 2025)

AI: Information Producer

- Variability in quality of AI outputs

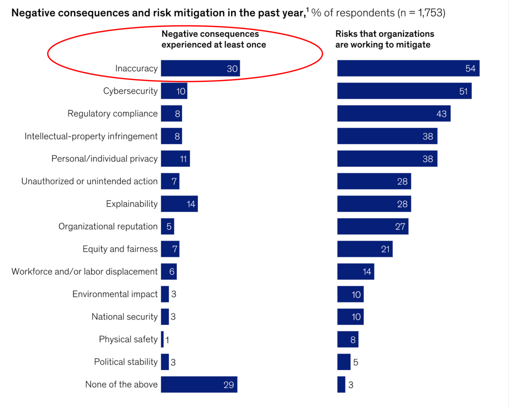


Why Language Models Hallucinate

Adam Tauman Kalai* OpenAI · Ofir Nachum OpenAI · Santosh S. Vempala† Georgia Tech · Edwin Zhang OpenAI

September 4, 2025

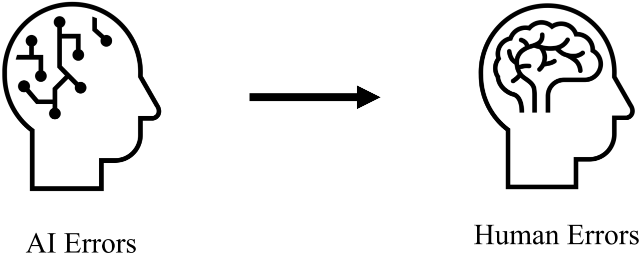
Hallucination



Inaccuracy

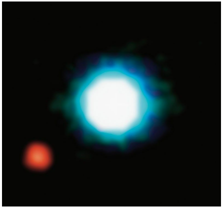
To Err is Human

- Human actors are constrained by their limited information processing capability and cognitive capacity
  - Expert vs. novice (Brynjolfsson et al., 2025)
  - Algorithmic appreciation vs. aversion (Jussupow et al., 2021)

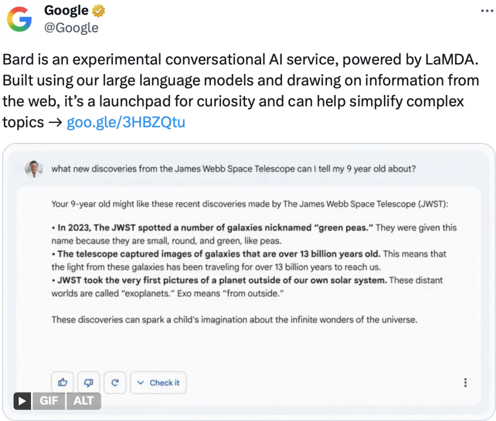


Type II Error

- Inaccurately accept inferior information



First exoplanet by VLT (2004)



Google
@Google

Bard is an experimental conversational AI service, powered by LaMDA. Built using our large language models and drawing on information from the web, it's a launchpad for curiosity and can help simplify complex topics → goo.gle/3HBZQtu

what new discoveries from the James Webb Space Telescope can I tell my 9 year old about?

Your 9-year old might like these recent discoveries made by The James Webb Space Telescope (JWST):

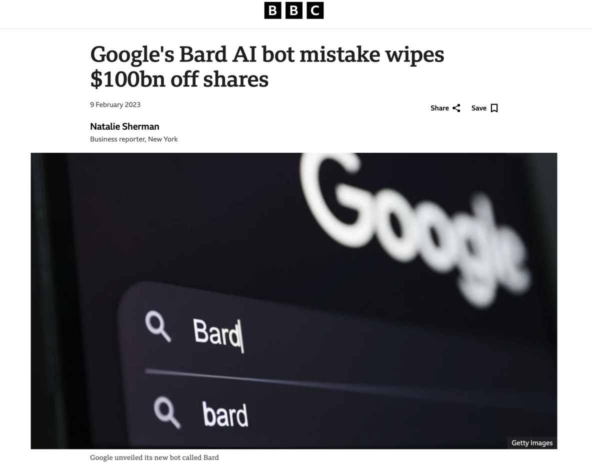
- **In 2023, The JWST spotted a number of galaxies nicknamed "green peas."** They were given this name because they are small, round, and green, like peas.
- **The telescope captured images of galaxies that are over 13 billion years old.** This means that the light from these galaxies has been traveling for over 13 billion years to reach us.
- **JWST took the very first pictures of a planet outside of our own solar system.** These distant worlds are called "exoplanets." Exo means "from outside."

These discoveries can spark a child's imagination about the infinite wonders of the universe.

Bard delivers inaccurate answer (Feb 2023)

Type II Error

- Inaccurately accept inferior information



BBC

Google's Bard AI bot mistake wipes $100bn off shares

9 February 2023

Share Save

Natalie Sherman

Business reporter, New York

Google unveiled its new bot called Bard

Alphabet loses $100 billion in market value (Feb 2023)

Type I Error

- Reject superior information
  - Verification bias: the correctness of the algorithmic outcome is verified only when the information is accepted and translated into action

    (de Véricourt and Gurkan, 2023)

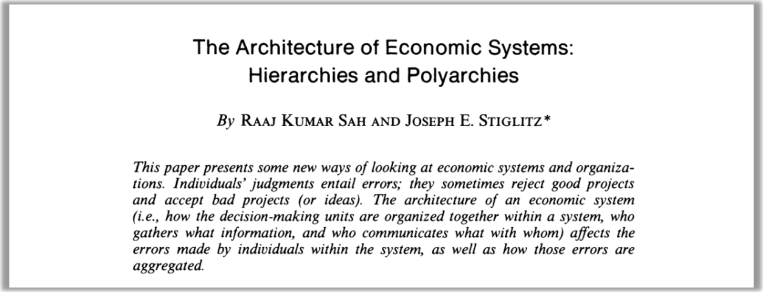
Type I error is silent and undermeasured.

Validation of AI-generated Information

- To fully leverage AI-generated information $\implies$ accept it when it is correct and reject it when it is incorrect
- Only recently have studies questioned the assumption that AI outputs are always correct
  - The cognitive process by which decision makers monitor both AI system performance and their own performance (Jussupow et al., 2021)
  - The design of algorithm that decision makers rely on to validate outcomes (Wang et al., 2025)
- **Gap**: individual level decision $\implies$ organizational outcome?

How Organizations Validate AI-generated Information?

- The information processing view in economics (Stiglitz, 1985; Sah and Stiglitz, 1986)



The Architecture of Economic Systems:
Hierarchies and Polyarchies

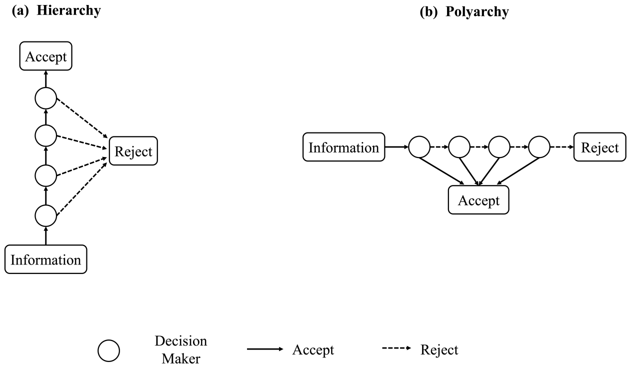
By RAAJ KUMAR SAH AND JOSEPH E. STIGLITZ*

This paper presents some new ways of looking at economic systems and organizations. Individuals' judgments entail errors; they sometimes reject good projects and accept bad projects (or ideas). The architecture of an economic system (i.e., how the decision-making units are organized together within a system, who gathers what information, and who communicates what with whom) affects the errors made by individuals within the system, as well as how those errors are aggregated.

Sah and Stilitz (AER 1986)

How Organizations Validate AI-generated Information?

- Two representative architectures (information processing structures)
  - Hierarchy reduces Type II errors; while polyarchy is deemed to minimize Type I errors (Sah and Stiglitz, 1986)

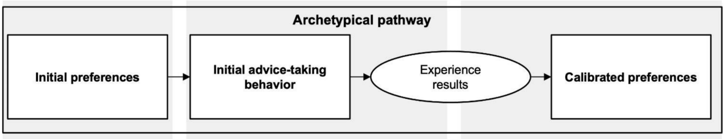


How Organizations Validate (AI-generated) Information?

- Two representative architectures (information processing structures):
  - Hierarchy and polyarchy represent the most strict and the most loose structures, respectively.
  - Alternative structures lie in between – e.g., committee (Sah and Stiglitz, 1988), hybrid structures (Christensen and Knudsen, 2010)

Backward Propagation also Exists

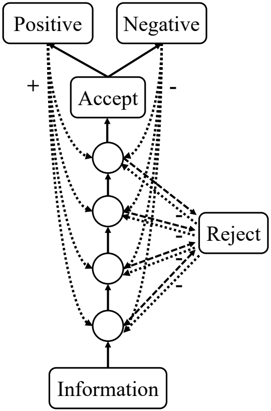
- Individuals' initial appreciation / aversion toward AI-generated information evolve as they experience the outcomes of accepting or rejecting it (Jussupow et al., 2024; Turel and Kalhan, 2023)
  - ▶ When individuals perceive AI systems as more (less) experienced $\Longrightarrow$ disproportionately prefer (disfavor) AI-generated information $\Longrightarrow$ systematic algorithmic appreciation (aversion) (Bigman and Gray, 2018; Hou and Jung, 2021; Dietvorst et al., 2015)
  - ▶ Belief vs. preference based bias (Bohren et al., 2019; Hu et al., 2025)



Jussupow et al. (MISQ 2024)

Endogenous Preference Calibration

- Information processing structures shape the extent to which individuals experience the associated outcomes
  - Asymmetries in exposure to AI-generated information
  - Asymmetric distribution of positive and negative experiences



Information Processing Structure with Preference Updating

This Study

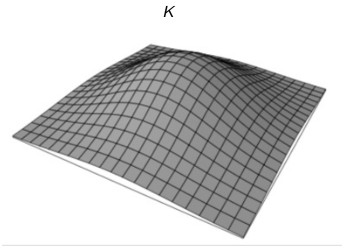
- **Central Idea**: Organizational validation of AI-generated information: information-processing structures filter AI-generated information and simultaneously shape how preferences are updated.
- **Outline**
  - A computational model captures the feedback loop
  - A series of experiments to examine how different information processing structures influence the validation of AI-generated information

A Computational Model

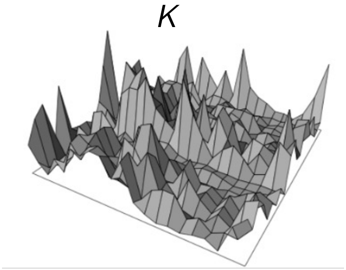
- Task Environment
- Agents
  - AI and AI-generated Information
  - Individual Evaluator
- Interaction
  - Information Processing Structure
  - Information Processing and Preference Calibration

Task Environment: A Space of Alternatives

- Alternative $\mathbf{d} \implies$ solution, software, strategic plan, etc.
  - $\mathbf{d} = \langle d_1, d_2, ..., d_N \rangle$
  - Each d_i is interdependent with K others
- Each alternative is associated with a fitness value $v(\mathbf{d})$



Smooth Landscape



Rugged Landscape

AI and AI-generated Information

- AI ⟹ an agent capable of searching the space at scale and generating advice in response to human input at a certain level of accuracy
    - Human input: an organization's initial location within the alternative space $\mathbf{d}$
    - Scale: the AI agent explores neighboring configurations $\mathbf{d}'$ of $\mathbf{d}$ within a Hamming distance of B
    - Accuracy: the AI agent returns one alternative from the top $100 \times (1 - A)\%$ explored configurations
    - AI-generated Information: an alternative $\mathbf{d}'$ to change the organization's status quo $\mathbf{d}$

Real-life scenarios

- Human actors prompt AI models to produce solutions in a cost-efficient way (Boussioux et al., 2024)

Individual Evaluator

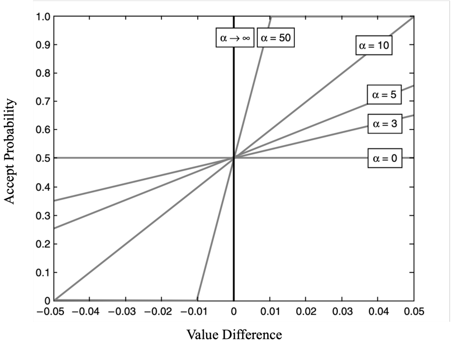
- Individuals $\implies$ fallible agents capable of distinguishing between AI-generated information $\mathbf{d}'$ and the status quo $\mathbf{d}$
- The likelihood of making errors is negatively associated with the differences in value between the generated information and the status quo $v(\mathbf{d}') - v(\mathbf{d})$

Real-life scenarios

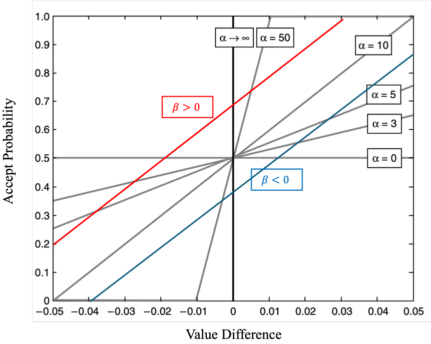
- An AI-generated containing many obvious factual error can be quickly rejected vs. one provides a slightly less accurate interpretation of data

Individual Evaluator

- Individuals are characterized as screening functions
  $f = \alpha(v(\mathbf{d}') - v(\mathbf{d})) + \beta$
  - α: Judgmental ability
  - β: Judgmental bias



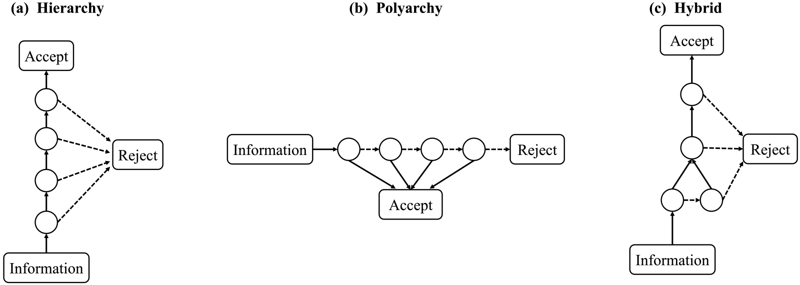
Judgmental ability α



Judgmental bias β

Information Processing Structure

- F: the aggregation of n individual evaluators
  - Hierarchy: $F = f^n$
  - Polyarchy: $F = 1 - (1 - f)^n$
  - Hybrid: mixture of hierarchy and polyarchy
- Formally, the extent to which an information processing structure approximates hierarchy or polyarchy is determined by the number of information processing layers l $(1 \leq l \leq n)$



Information Processing and Preference Calibration

- The information processing structure operates repeatedly within a consistent task environment
  - At each step t, the organization (along with its individuals) is located at a status quo $\mathbf{d}$
  - The AI agent then generates information on changing the status quo $\mathbf{d}$ to an alternative configuration $\mathbf{d}'$
  - Individuals evaluate the generated information according to the information processing structure and decide whether to adopt the alternative $\mathbf{d}'$
    - If adopted, the value associated with $\mathbf{d}'$ is realized
    - If not adopted, the value associated with $\mathbf{d}$ is retained

Information Processing and Preference Calibration

- Individuals calibrate their preferences toward the AI system according to outcome $\sigma_{i,t}$ at time t
  - $\sigma_{i,t} \in \{-1, 0, 1\}$
    - $\sigma_{i,t} = 1$ if the processed information is verified to enhance the overall value $v(\mathbf{d}') > v(\mathbf{d})$
    - $\sigma_{i,t} = -1$ if the processed information fails to do so (verified to lower the value realized; rejected)
    - $\sigma_{i,t} = 0$ if individual i does not process the information at t
  - Calibrate preference $\beta_{i,t+1} = \beta_{i,t} + \frac{1}{m_{i,t}+1}(\sigma_{i,t} - \beta_{i,t})$
    - $m_{i,t}$: the number of prior experiences; $\beta_{i,t+1}$: the average of all prior experiences

Validation of the Model

- Model Components
- Validation Experiments
  - AI accuracy $(A) \uparrow \implies$ individual performance $\uparrow$
  - AI accuracy $(A) \uparrow (\downarrow) \implies$ individual preferences shift toward appreciation (aversion)
  - The performance of information processing structures without AI-generated information
- Sensitivity Analyses

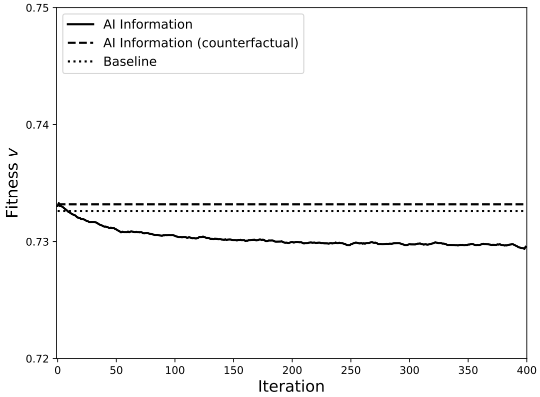
Experiment Design

- Analyses
  - Experiment 1: the performance of hierarchy
  - Experiment 2: the impact of AI accuracy (A)
  - Experiment 3: alternative information processing structures (I)
    - 400 iterations (t)
    - 10,000 replications
    - Number of individuals $n = 6$
    - Judgmental ability $\alpha = 5$

Parameter space

Parameter	**Value**	**Robustness Checks**
Ruggedness of the alternative space	$K = 11$	$[1, 3, 5, 7, 9]$
Breadth of AI search	$B = 2$	$[4, 6, 8, 10, 12]$

Baseline: A Hierarchical Information Processing Structure

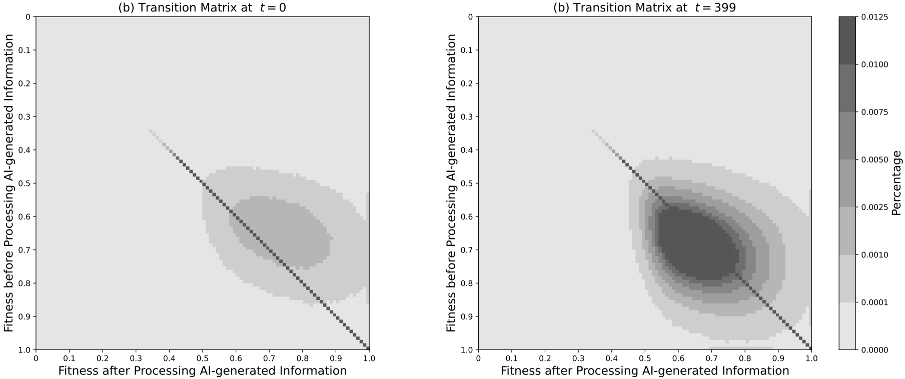


Performance of Hierarchy over Time ($A = 0.6$)

Findings

- Hierarchy + AI-generated information $\implies$ performance ↑
- Hierarchy fails to maintain the advantage

Performance Transition Matrix

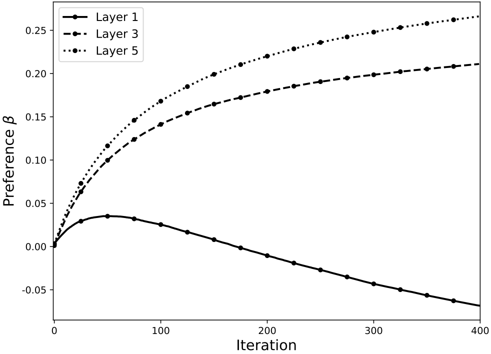


Performance Transition Matrix

Breakdown of performance drop

- Individuals calibrate preference $\implies$ hierarchy adopts AI-generated information ↑
- Inferior AI-generated information ↑

Preference Calibration



Preferences in Hierarchy over Time

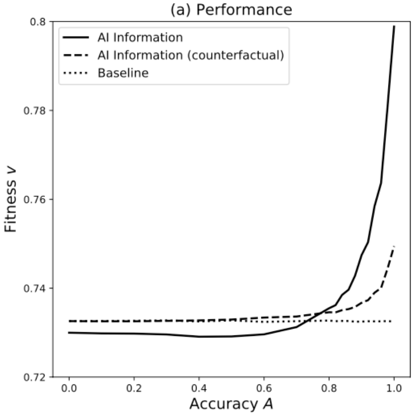
Mechanism

- Asymmetric updating of preferences
  - Higher-layer individuals still exhibit appreciation even when lower-layer individuals begin to exhibit aversion

Baseline: A Hierarchical Information Processing Structure

- A temporal misalignment:
    - Hierarchy $\Longrightarrow$ adopting AI-generated information $\leftrightarrow$ preference calibration

The Impact of AI accuracy

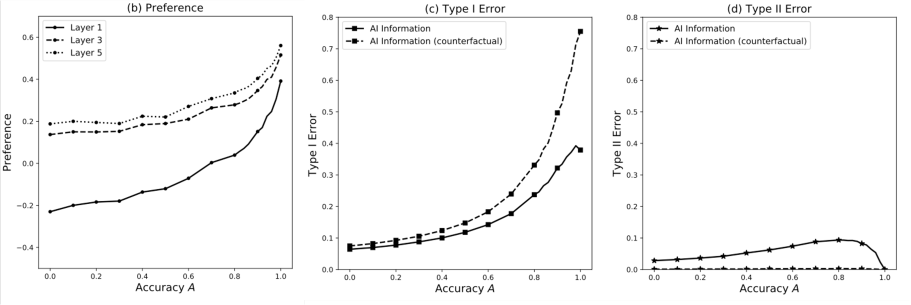


Performance of Hierarchy ($t = 400$)

Finding

- With preference updating, AI accuracy $\uparrow \implies$ performance $\cup$

The Impact of AI Accuracy



Preferences, Type I and Type II errors in Hierarchy ($t = 400$)

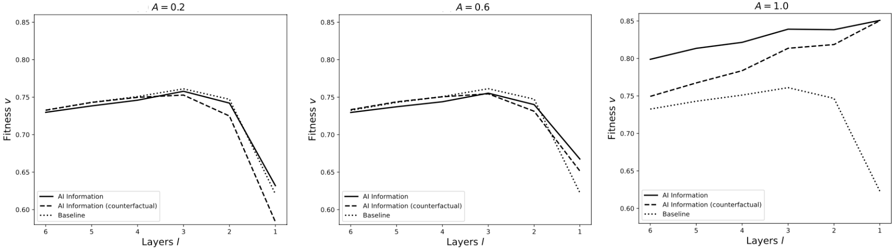
Mechanism

- AI accuracy ↑ | preference updating $\implies$ Type II error $\cap$; Type I error ↓

The Impact of AI Accuracy

- A temporal misalignment:
  - Hierarchy $\implies$ adopting AI-generated information $\leftrightarrow$ preference calibration
  - As AI accuracy (A) increases, temporal misalignment first becomes more likely, then declines once accuracy becomes very high
    - ⊙ Better AI systems do not necessarily lead to higher performance

Alternative Information Processing Structures



Performance across Information Processing Structures

Findings

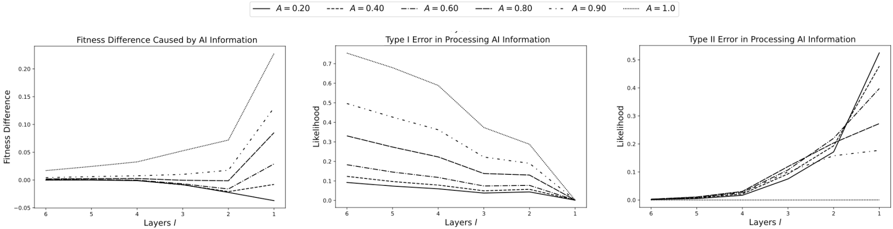
- AI accuracy ↑ $\implies$ hybrid → polyarchy
- hybrid → polyarchy $\implies$ performance decline ↓

Dissecting the Impacts

- Total Impact = Impact of AI + Impact of Preference Updating
  - Impact of AI = AI w/t preference updating (counterfactual) − no AI (baseline)
  - Impact of Preference Updating = AI w/ preference updating (treatment) − AI w/t preference updating (counterfactual)

Difference Caused by AI-generated Information

- AI w/t preference updating − no AI (counterfactual − baseline)



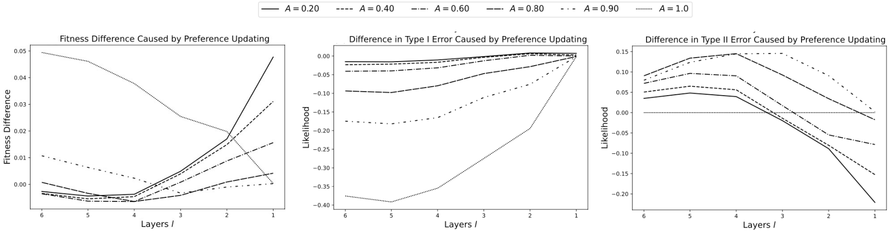
Difference Caused by AI-generated Information

Findings

- hierarchy → polyarchy $\implies$ performance difference ↑ (↓) | AI accuracy
- AI accuracy ↑ $\implies$ difference in Type I error ↑; difference in Type II error ↑

Difference Caused by Preference Updating

- AI w/ preference updating − AI w/t preference updating (treatment − counterfactual)



Fitness Difference Caused by Preference Updating

Difference in Type I Error Caused by Preference Updating

Difference in Type II Error Caused by Preference Updating

Difference Caused by AI-generated Information

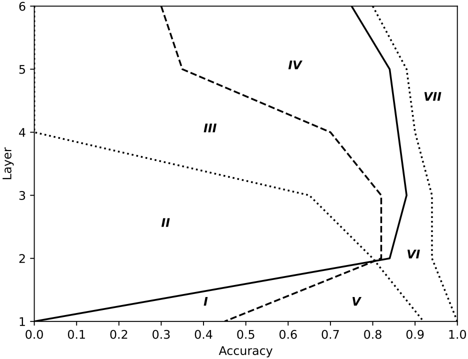
Findings

- hierarchy → polyarchy $\Longrightarrow$ performance difference $\cup$ ($\downarrow$) | AI accuracy
- hierarchy → polyarchy $\Longrightarrow$ Type II error $\cup$ | AI accuracy

Alternative Information Processing Structures

- A temporal misalignment:
  - Hierarchy $\implies$ adopting AI-generated information $\leftrightarrow$ preference calibration
  - As AI accuracy A increases, temporal misalignment first becomes more likely, then declines once accuracy becomes very high
    - ⊙ Better AI systems do not necessarily lead to higher performance
  - Hybrid structure is more likely to be adversely impacted
    - ⊙ A U-shaped function, with polyarchy being immune to the performance decline

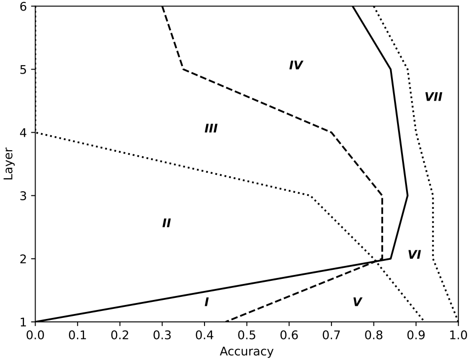
Summary



Setup

- Dashed line: initial performance ≥ without AI-generated information
- Solid line: eventual performance ≥ without AI-generated information
- Dotted lines: eventual performance ≥ initial performance

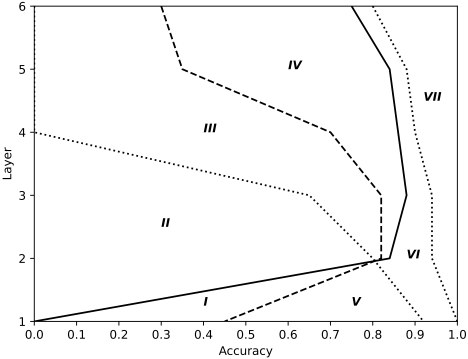
Summary



Region

- I: initial performance ↓, preference updating ↑, eventual performance ↑
- II: initial performance ↓, preference updating ↑, eventual performance ↓
- III: initial performance ↓, preference updating ↓, eventual performance ↓

Summary



Region

- IV: initial performance ↑, preference updating ↓, eventual performance ↓
- V and VII: initial performance ↑, preference updating ↑, eventual performance ↑
- VI: initial performance ↑, preference updating ↓, eventual performance ↑

Contribution to Theory

- A structural perspective on validating AI-generated information
  - The dual impact of information processing structure – i.e., filtering the generated information and endogenously shaping individuals' preference calibration
  - Uncover the temporal complexity and the boundary condition in validating AI-generated information
- The role of structure in shaping the evolution of algorithmic appreciation and aversion
- The interplay between the design of organization structure and human actors' adaptation on influencing the effectiveness of leveraging AI

Practical Implication

- The optimal design of information processing structure in processing AI-generated information
- The evaluation of AI-generated information across different horizons
  - Poor performance does not necessarily imply that the AI system is poorly designed

Future Work

- Empirical testing
  - Survey in a manufacturing company
- Institutional factors
  - The extent to which individuals reveal or conceal their usage of AI is influenced by institutional factors (Zhou et al., 2025; Reif et al., 2025)
- Prescription
  - The design of feedback mechanism so that individuals at intermediate levels receive additional feedback

Thank You!

Questions?
jungpil@nus.edu.sg