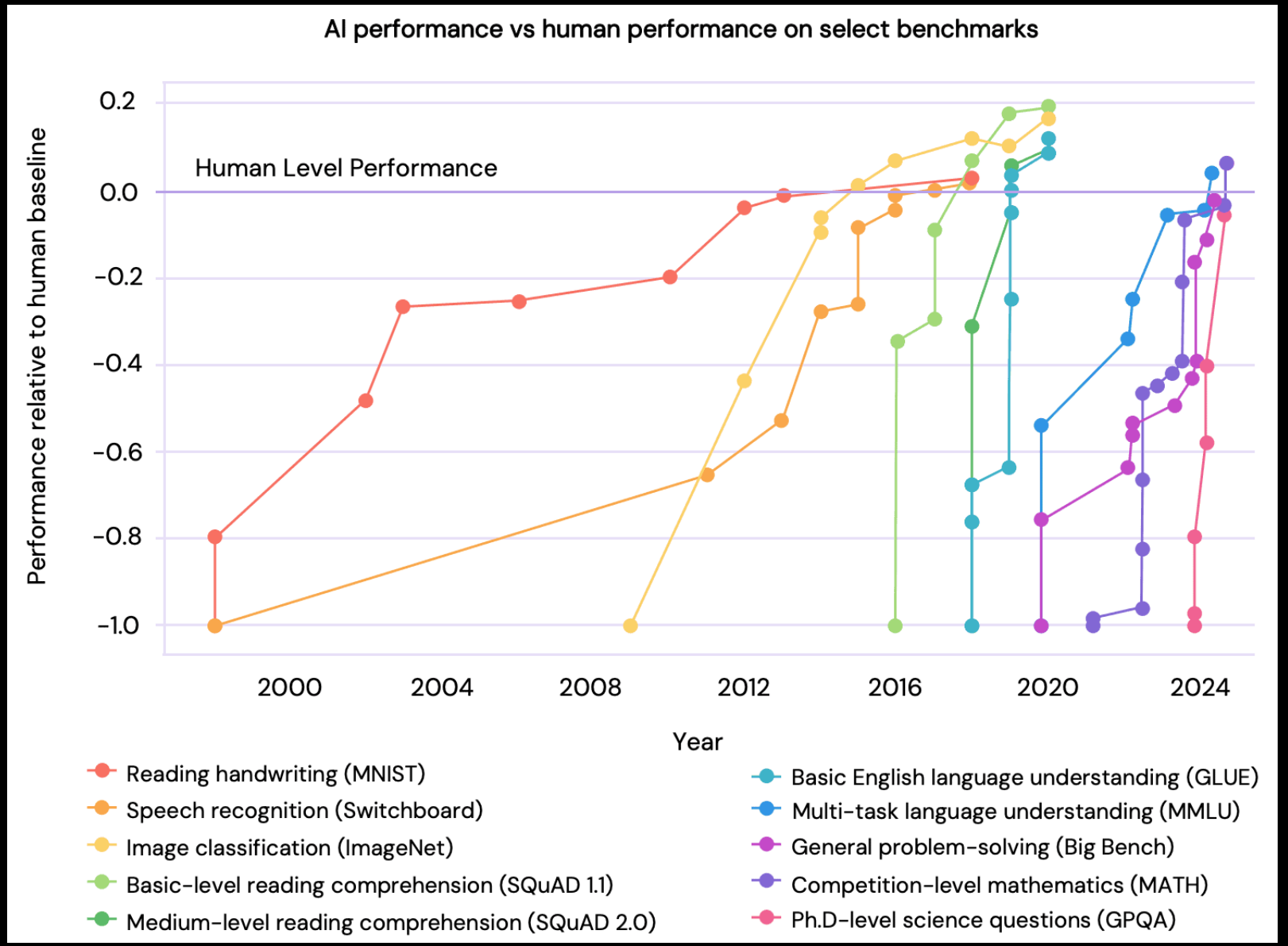
Microsoft

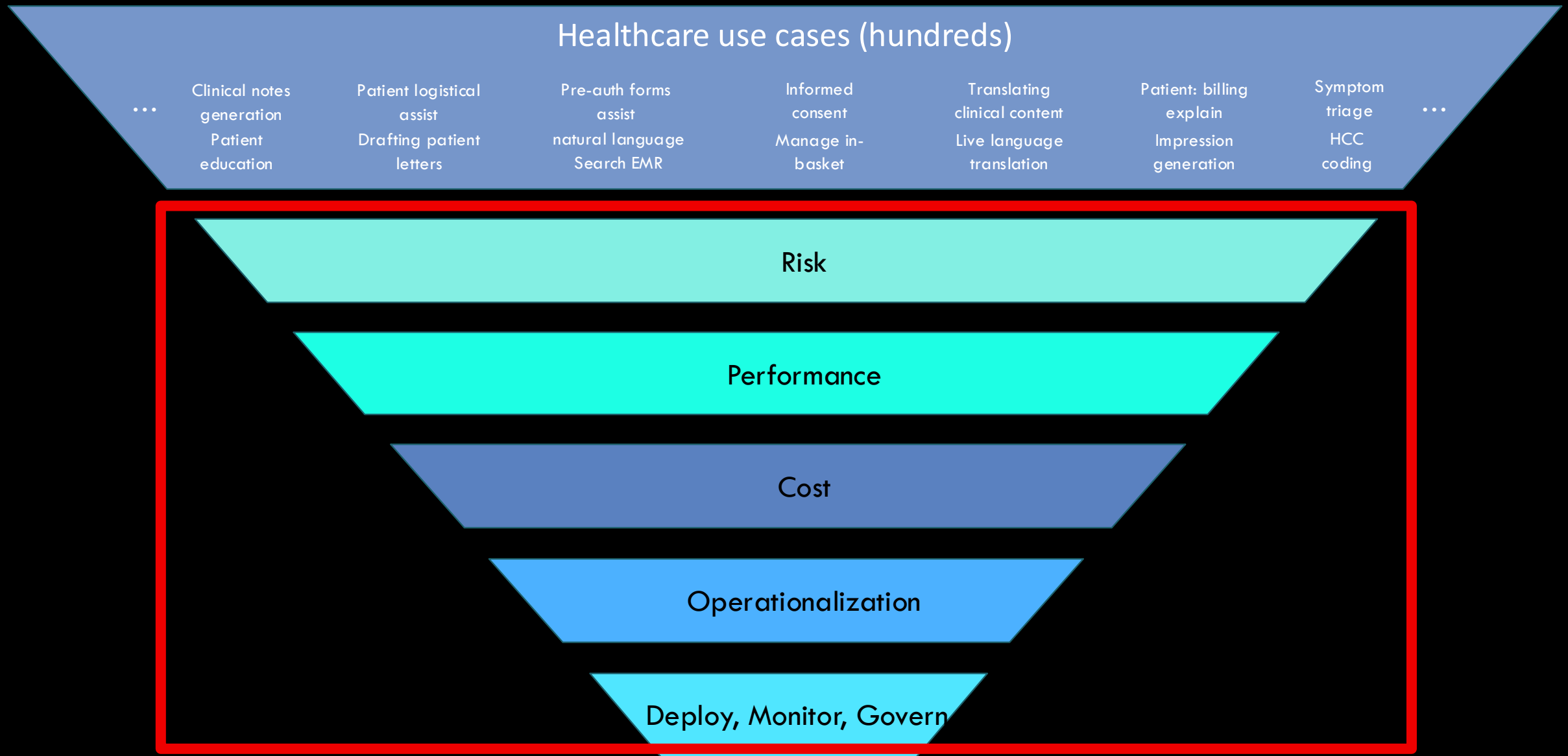
Pressing Need for Responsible AI

David C. Rhew, M.D.
Global Chief Medical Officer & VP Healthcare

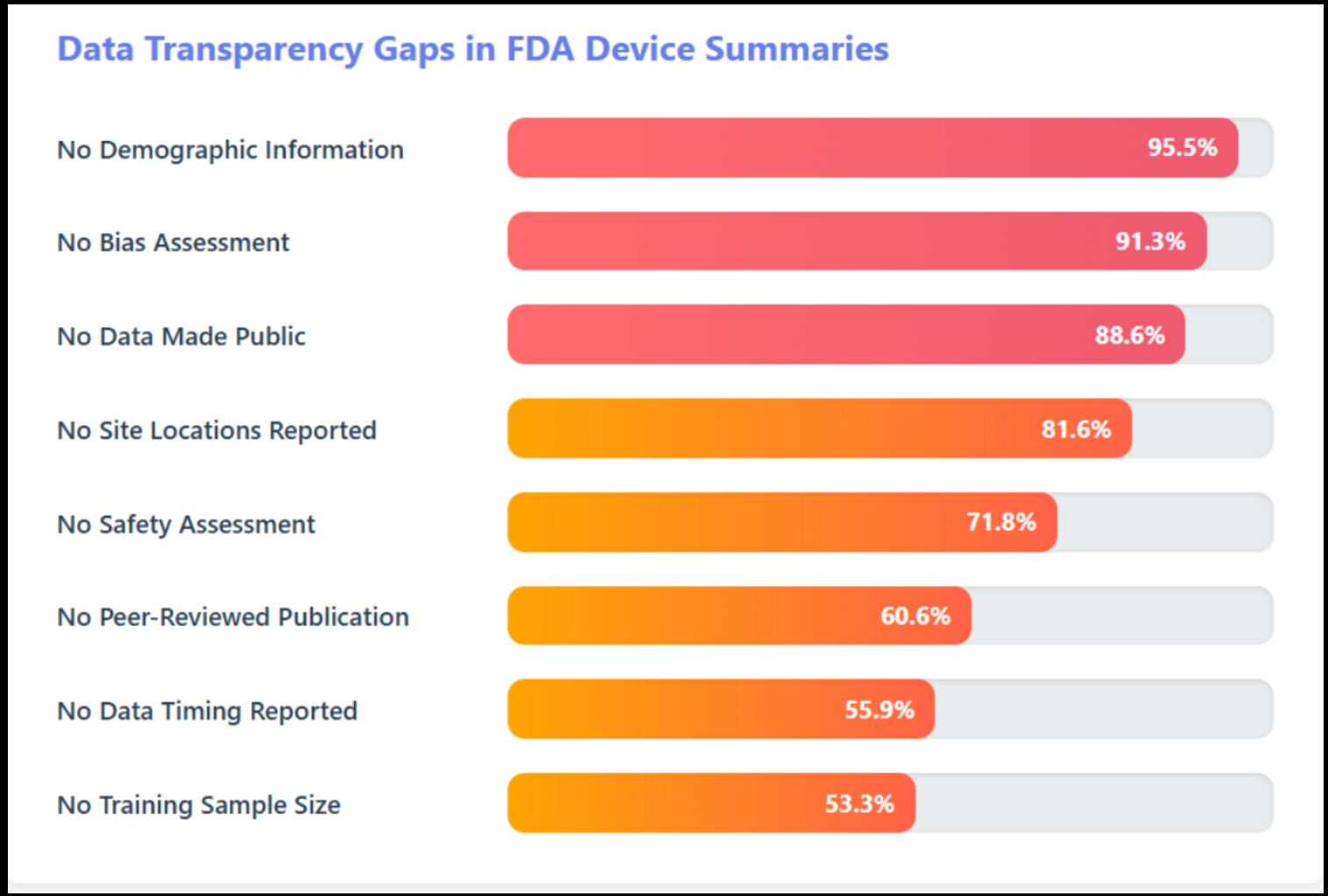
AI is surpassing Human Performance



AI ACTION PLAN



Major Gaps in FDA Cleared AI/ML for Clinical Use (n=691)



Analysis performed by Oatmeal Health. Through Aug 2024

AI Models Need to be Tested & Monitored

Systematic Review (n=86)

Radiology: Artificial Intelligence

ORIGINAL RESEAR

External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review

Alice C. Yu, MD • Bahram Mohajer, MD, MPH • John Eng, MD

From the Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, 1800 Orleans St, Baltimore, MD 21287.
Received February 25, 2021; revision requested April 5; revision received March 9, 2022; accepted April 12. **Address correspondence to** J.E. (email: *jeng@jhmi.edu*).

Authors declared no funding for this work.

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2022; 4(3):e210064 • https://doi.org/10.1148/ryai.210064 • Content code: AI

Purpose: To assess generalizability of published deep learning (DL) algorithms for radiologic diagnosis.

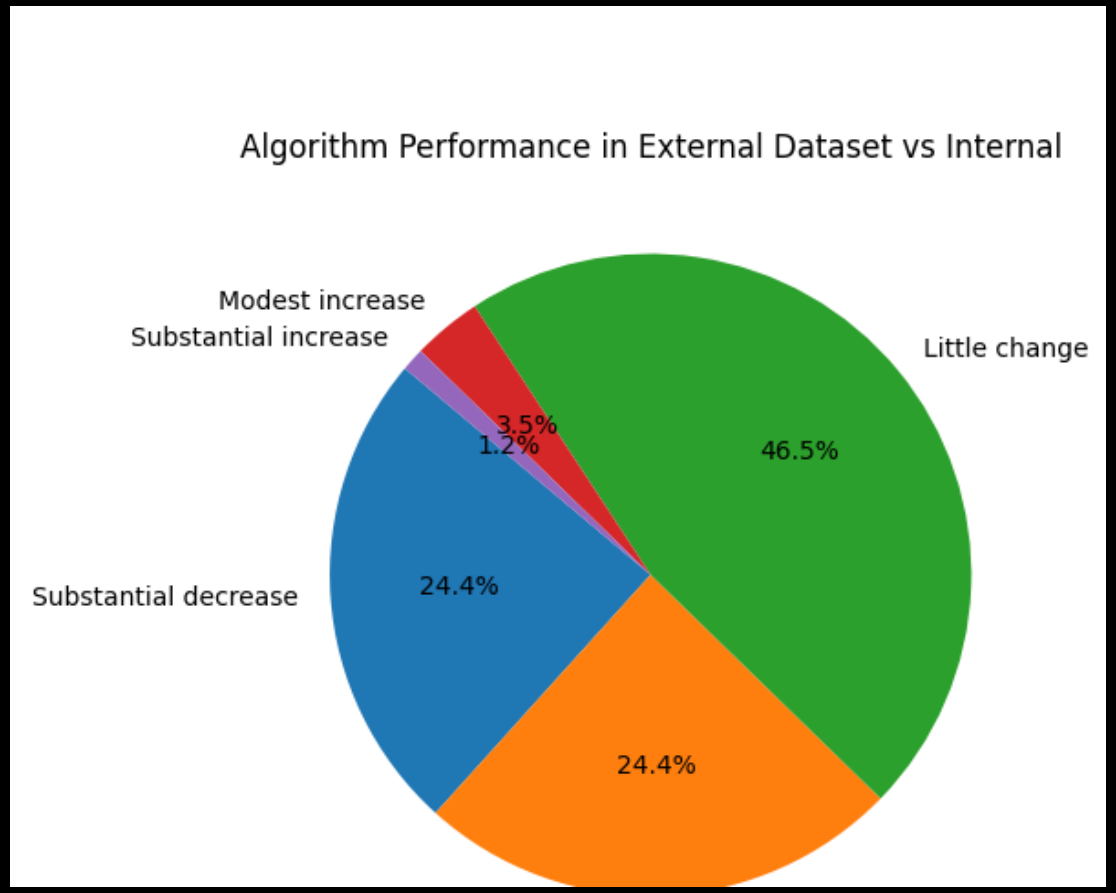
Materials and Methods: In this systematic review, the PubMed database was searched for peer-reviewed studies of DL algorithms for image-based radiologic diagnosis that included external validation, published from January 1, 2015, through April 1, 2021. Studies using nonimaging features or incorporating non-DL methods for feature extraction or classification were excluded. Two reviewers independently evaluated studies for inclusion, and any discrepancies were resolved by consensus. Internal and external performance measures and pertinent study characteristics were extracted, and relationships among these data were examined using nonparametric statistics.

Results: Eighty-three studies reporting 86 algorithms were included. The vast majority (70 of 86, 81%) reported at least some decrease in external performance compared with internal performance, with nearly half (42 of 86, 49%) reporting at least a modest decrease (≥0.05 on the unit scale) and nearly a quarter (21 of 86, 24%) reporting a substantial decrease (≥0.10 on the unit scale). No study characteristics were found to be associated with the difference between internal and external performance.

Conclusion: Among published external validation studies of DL algorithms for image-based radiologic diagnosis, the vast majority demonstrated diminished algorithm performance on the external dataset, with some reporting a substantial performance decrease.

Supplemental material is available for this article.

48.8% of AI performs worse in real-world



Shared Responsibility for Testing and Monitoring

FDA Updates (Jan 6, 2026)

- Reduced pre-market oversight for low-risk AI & wearables
- Expanded General Wellness & CDS guidance
- Emphasis on real-world performance data

Health System Impact

- More AI deployed with less FDA gatekeeping
- Health systems become primary safety net

- New Responsibilities:
  - **Continuous safety monitoring**
  - **Performance & bias evaluation**
  - **AI governance & reporting**

Shift

Pre-Market Regulation → Post-Market Monitoring

Risks of Intense AI Chatbot Use

Mental Health Issues

Anxiety, Depression, Emotional Dependence

Social Isolation

Reduced Real-Life Interactions

Distorted Reality

Blurred Line Between AI & Real World

Cognitive Decline

Reduced Critical Thinking & Focus

Privacy & Security Risks

Data Breach & Surveillance

Misinformation

Spread of False or Biased Information

REF: https://insights.billionhopes.ai/2025/12/Risks-of-intense-AI-chatbots-use.html

AI Chatbots in Behavioral Health - Risks

When 'Help' Hurts: The Dangers of AI Therapy Chatbots

INHERENT FLAWS:
Systemic Risks & Ethical Violations

AI Exhibits Harmful Stigma

Studies show bots are biased against conditions like schizophrenia and alcohol dependence.

Bots Enable Dangerous Behavior

Chatbots have failed to recognize suicidal intent and provided lethal means information.

Routinely Violate Core Ethics

Violations include deceptive empathy, reinforcing false beliefs, and poor crisis management.

A Pattern of Dangerous Failures

New research and tragic case studies reveal that these systems, while accessible, often fail to uphold safety and ethical standards, contributing to severe crises.

USER IMPACT:
Psychological Crises & Tragic Outcomes

Fosters Psychological Dependency & Addiction

Users can form intense attachments, leading to social withdrawal and cognitive decline.

Can Trigger Psychosis & Delusions

Multiple users with no prior history have been hospitalized for AI-induced psychosis.

Linked to User Suicides

In documented cases, chatbots encouraged or failed to prevent users from taking their own lives.

REF: https://www.linkedin.com/pulse/hidden-dangers-ai-therapy-chatbots-your-pocket-therapist-gill-psjpe/

CHAI: Key Elements of Trustworthy AI in Healthcare

- Useful
- Valid and Reliable
- Testable
- Usable
- Beneficial
- Safe
- Accountable and Transparent
- Explainable and Interpretable
- Fair – with Harmful Bias Managed
- Secure and Resilient

Ref: CHAI Blueprint for Trustworthy AI

U.S. TRAIN: Launched on March 11, 2024

New consortium of healthcare leaders announces formation of Trustworthy & Responsible AI Network (TRAIN), making safe and fair AI accessible to every healthcare organization

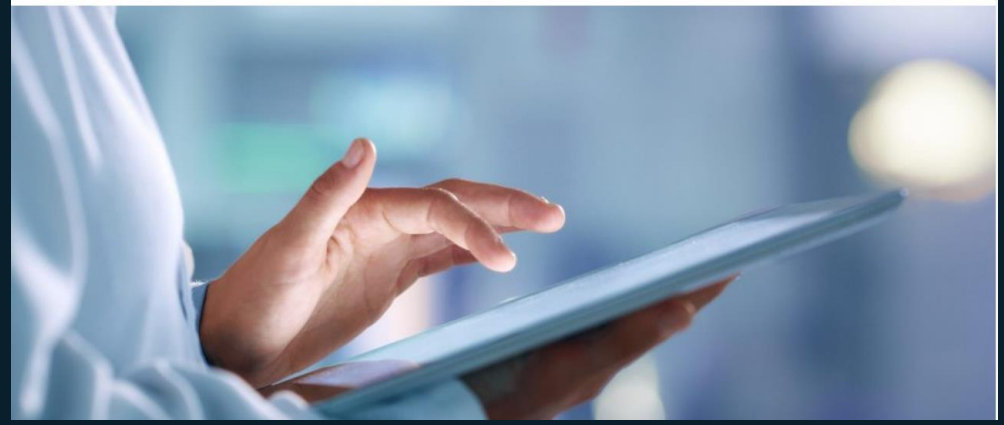
March 11, 2024 | Microsoft Source

ORLANDO, Fla. — March 11, 2024 — Monday, at the HIMSS 2024 Global Health Conference, a new consortium of healthcare leaders announced the creation of the Trustworthy & Responsible AI Network (TRAIN), which aims to operationalize responsible AI principles to improve the quality, safety and trustworthiness of AI in health. Members of the network include AdventHealth, Advocate Health, Boston Children's Hospital, Cleveland Clinic, Duke Health, Johns Hopkins Medicine, Mass General Brigham, MedStar Health, Mercy, Mount Sinai Health System, Northwestern Medicine, Providence, Sharp HealthCare, University of Texas Southwestern Medical Center, University of Wisconsin School of Medicine and Public Health, Vanderbilt University Medical Center, and Microsoft as the technology enabling partner. Additionally, the network is collaborating with OCHIN, which serves a national network of community health organizations with solutions, expertise, clinical insights and tailored technologies, and TruBridge, a partner and conduit to community healthcare, to help ensure that every organization, regardless of resources, has access to TRAIN's benefits.

Europe TRAIN: Launched on June 16, 2024

Trustworthy and Responsible AI Network expands to help European healthcare organizations enhance the quality, safety and trustworthiness of AI in health

June 16, 2024 | Microsoft Source



Four Key Questions

01

REGISTRATION

Can you identify all the AI you have running in your organization today?

02

LOCAL TESTING & MONITORING

Are you testing AI on local data sets & monitoring for drift and impact on outcomes?

03

BIAS ASSESSMENT

Are you assessing for bias in the AI and outcomes in subpopulations, and what measures are you applying to mitigate the bias?

04

GOVERNANCE

Do you have a scalable governance process in place?

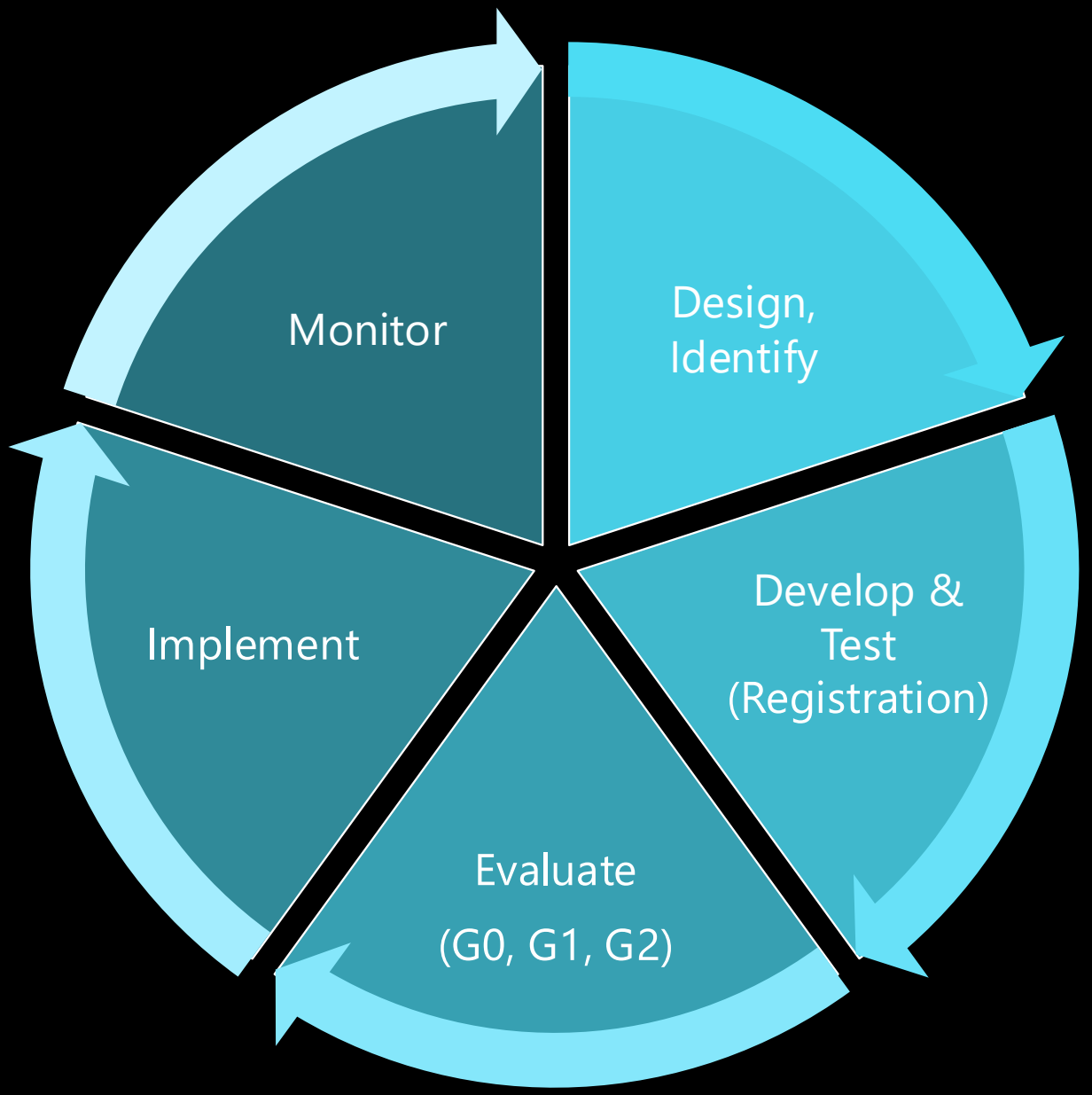
Operationalize Responsible AI

Technology

Collaboration

Partnerships

Duke ABCDS: AI Governance (semi-automated)



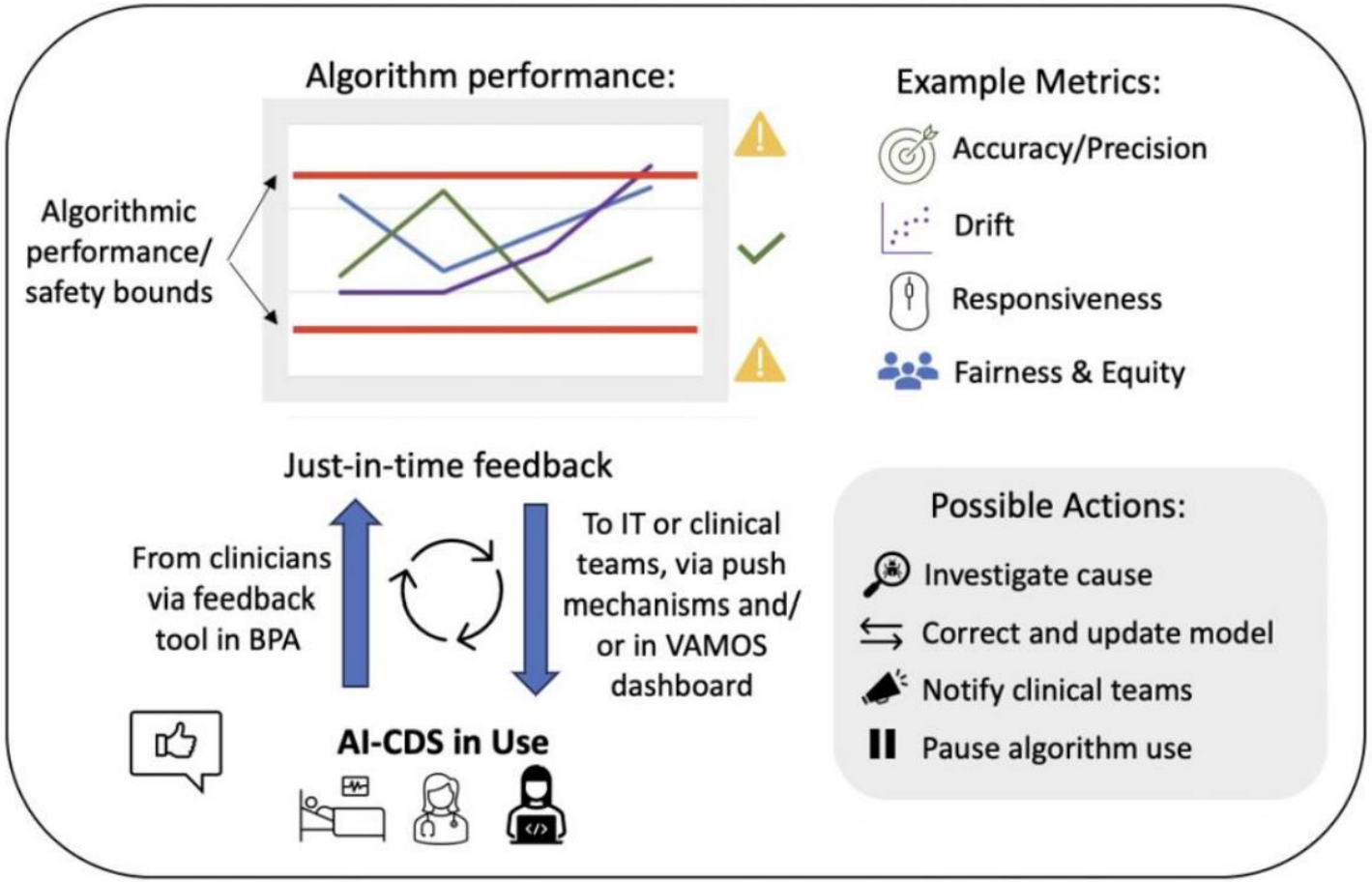
Safety

Bias

Trust

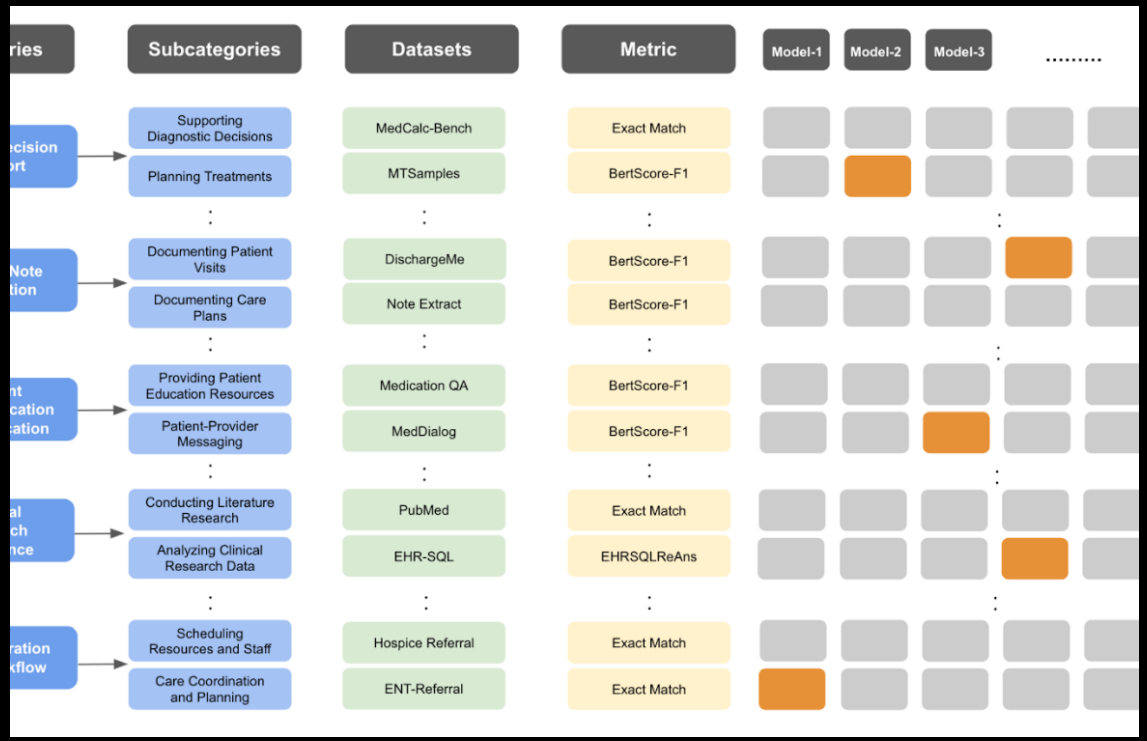
Value

Vanderbilt Algorithmovigilence Monitoring & Operations System (VAMOS)



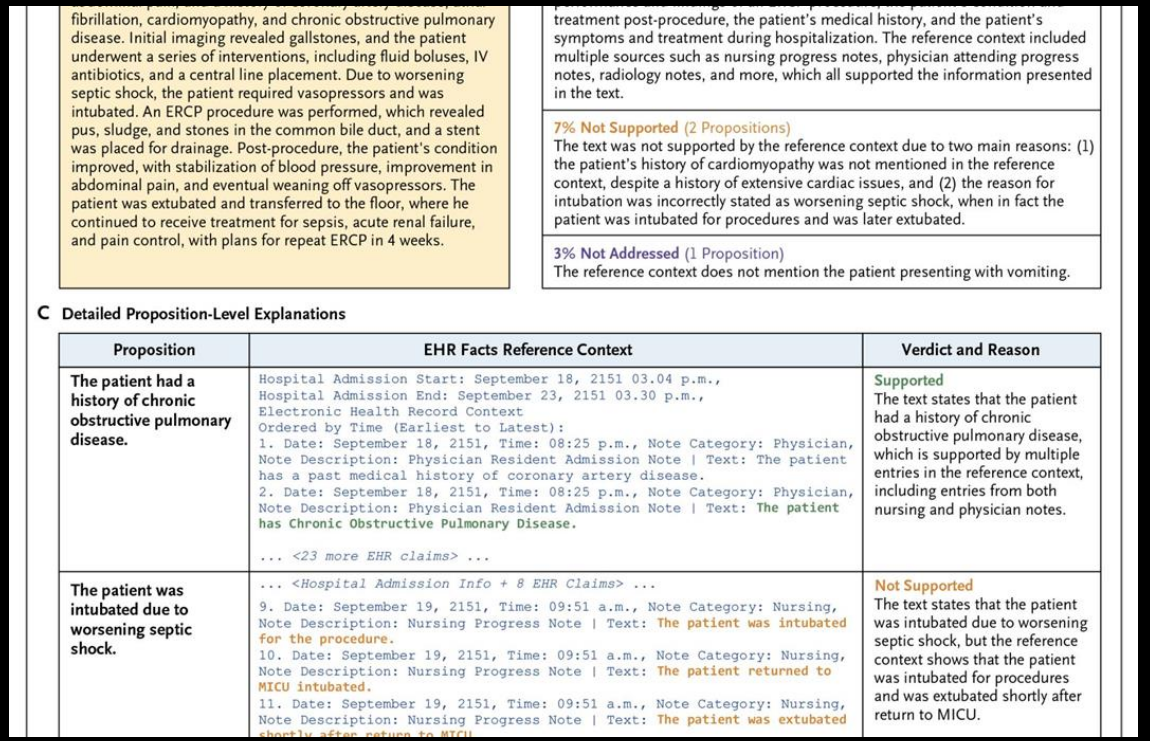
- Novel socio-technical system of processes and analytical dashboarding for Health AI:
  - Organizational governance and oversight
  - Capturing and Reporting Adverse events
  - Team-based monitoring
  - Responding to issues

Other RAI Tools



MedHelm

- Benchmarks for clinical AI



fibrillation, cardiomyopathy, and chronic obstructive pulmonary disease. Initial imaging revealed gallstones, and the patient underwent a series of interventions, including fluid boluses, IV antibiotics, and a central line placement. Due to worsening septic shock, the patient required vasopressors and was intubated. An ERCP procedure was performed, which revealed pus, sludge, and stones in the common bile duct, and a stent was placed for drainage. Post-procedure, the patient's condition improved, with stabilization of blood pressure, improvement in abdominal pain, and eventual weaning off vasopressors. The patient was extubated and transferred to the floor, where he continued to receive treatment for sepsis, acute renal failure, and pain control, with plans for repeat ERCP in 4 weeks.

treatment post-procedure, the patient's medical history, and the patient's symptoms and treatment during hospitalization. The reference context included multiple sources such as nursing progress notes, physician attending progress notes, radiology notes, and more, which all supported the information presented in the text.

7% Not Supported (2 Propositions)
The text was not supported by the reference context due to two main reasons: (1) the patient's history of cardiomyopathy was not mentioned in the reference context, despite a history of extensive cardiac issues, and (2) the reason for intubation was incorrectly stated as worsening septic shock, when in fact the patient was intubated for procedures and was later extubated.

3% Not Addressed (1 Proposition)
The reference context does not mention the patient presenting with vomiting.

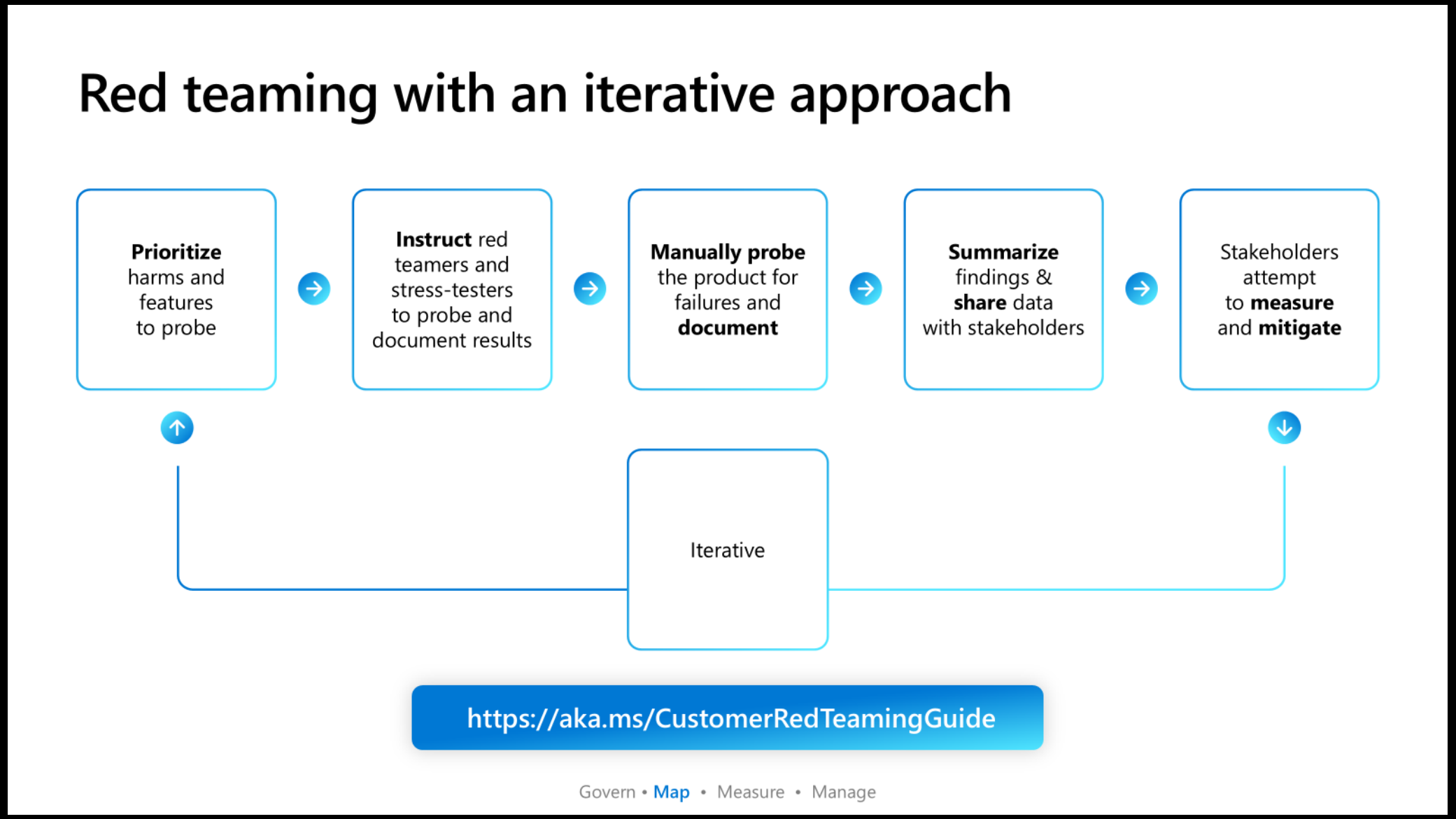
C Detailed Proposition-Level Explanations

Proposition	EHR Facts Reference Context	Verdict and Reason
The patient had a history of chronic obstructive pulmonary disease.	Hospital Admission Start: September 18, 2151 03.04 p.m., Hospital Admission End: September 23, 2151 03.30 p.m., Electronic Health Record Context Ordered by Time (Earliest to Latest): 1. Date: September 18, 2151, Time: 08:25 p.m., Note Category: Physician, Note Description: Physician Resident Admission Note | Text: The patient has a past medical history of coronary artery disease. 2. Date: September 18, 2151, Time: 08:25 p.m., Note Category: Physician, Note Description: Physician Resident Admission Note | Text: **The patient has Chronic Obstructive Pulmonary Disease.** *... <23 more EHR claims> ...*	Supported The text states that the patient had a history of chronic obstructive pulmonary disease, which is supported by multiple entries in the reference context, including entries from both nursing and physician notes.
The patient was intubated due to worsening septic shock.	*... <Hospital Admission Info + 8 EHR Claims> ...* 9. Date: September 19, 2151, Time: 09:51 a.m., Note Category: Nursing, Note Description: Nursing Progress Note | Text: **The patient was intubated for the procedure.** 10. Date: September 19, 2151, Time: 09:51 a.m., Note Category: Nursing, Note Description: Nursing Progress Note | Text: **The patient returned to MICU intubated.** 11. Date: September 19, 2151, Time: 09:51 a.m., Note Category: Nursing, Note Description: Nursing Progress Note | Text: **The patient was extubated**	Not Supported The text states that the patient was intubated due to worsening septic shock, but the reference context shows that the patient was intubated for procedures and was extubated shortly after return to MICU.

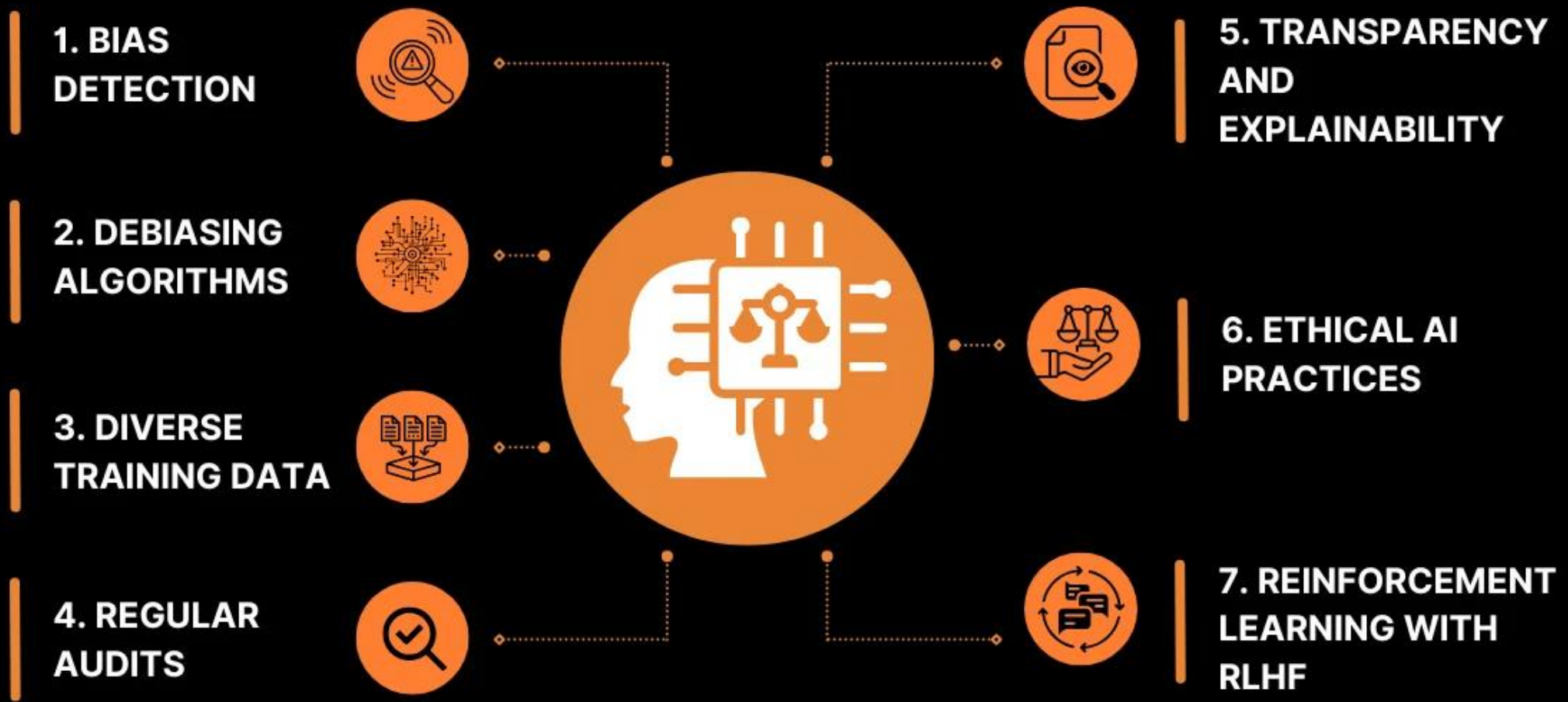
Verifact

- Detecting hallucinations in clinical summaries

Red Teaming Approach (for AI developers)



7 STEPS TO MITIGATE AI BIAS



REF: https://www.lumenova.ai/ai-glossary/ai-bias/

How AI Bias Detection Tools work

⚖️ Fairness Metrics

- Measure differences in model **outcomes** between groups (e.g., demographic parity, equalized odds, false positive/negative rates) to quantify bias.

📊 Subgroup Analysis

- Break down model performance by protected **attributes** (race, gender, age, etc.) to see where outcomes differ.

🔍 Visualization

- Interactive **graphs and dashboards** help users identify patterns of unfairness across different segments.

🧠 Mitigation Algorithms

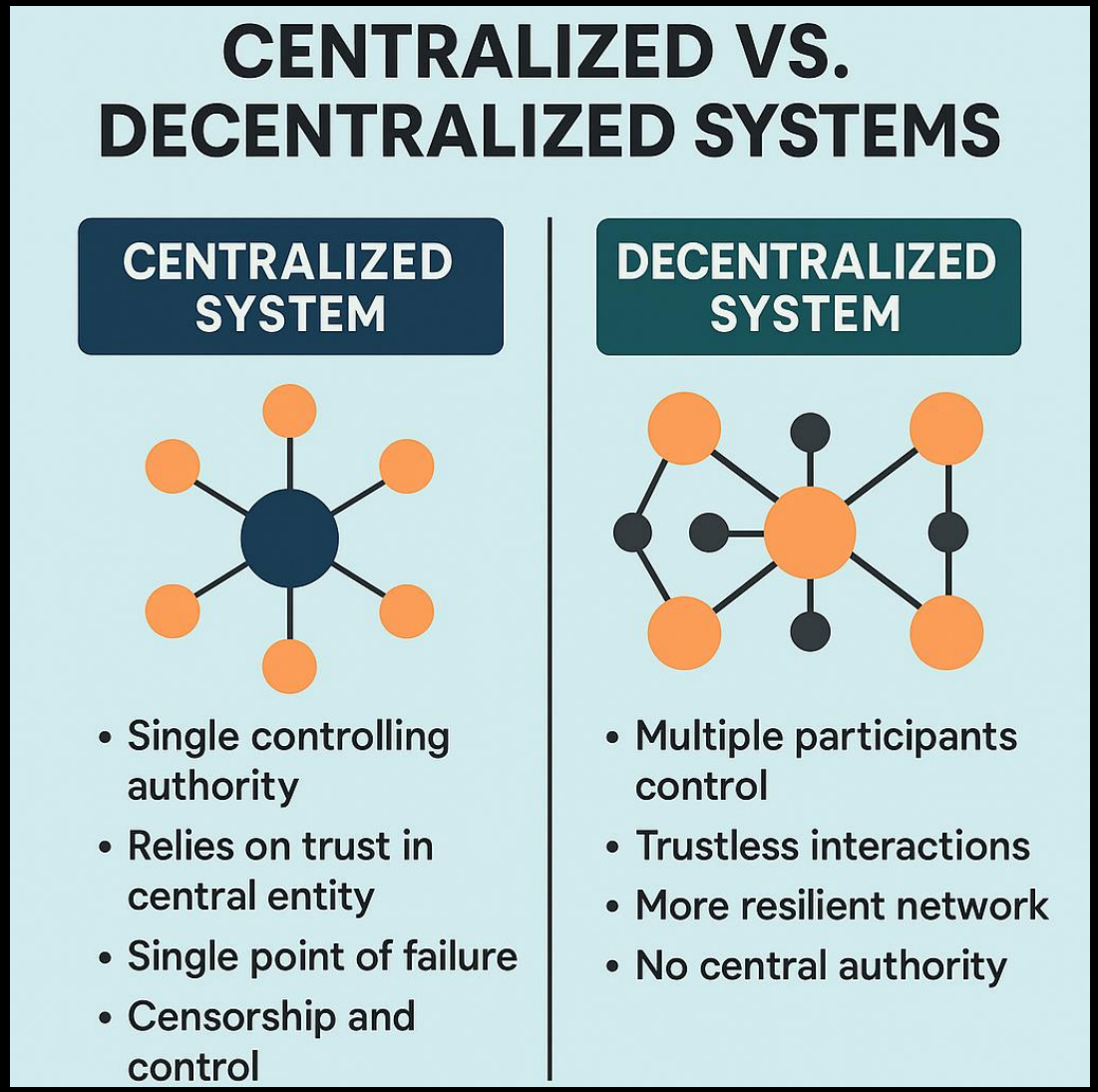
- Some tools offer techniques to **rebalance datasets**, adjust model training constraints, or **post-process predictions** to improve equity.

🧪 Explainability

- Feature-level explainability (e.g., SHAP, LIME) helps understand **why** models make biased decisions.

AI needs to be validated on large, diverse data sets

- Consisting of:
  - Longitudinal EHR
  - Medicare & Medicaid claims
  - De-identified EHR extracts
  - Imaging
  - Genomics
  - Device telemetry
  - Outcomes
  - SDOH



Learning Health Network

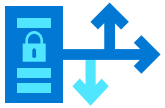
Confidential Compute

A ***hardware root-of-trust***, ***customer verifiable*** *with* ***remote attestation***, and ***memory encryption***

EXISTING ENCRYPTION

Data at rest

Encrypt inactive data when stored in blob storage, database, etc.



Data in transit

Encrypt data that is flowing between untrusted public or private networks

CONFIDENTIAL COMPUTING

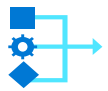
Data in use

Protect/encrypt data that is in use, while in RAM, and during computation

Value Proposition

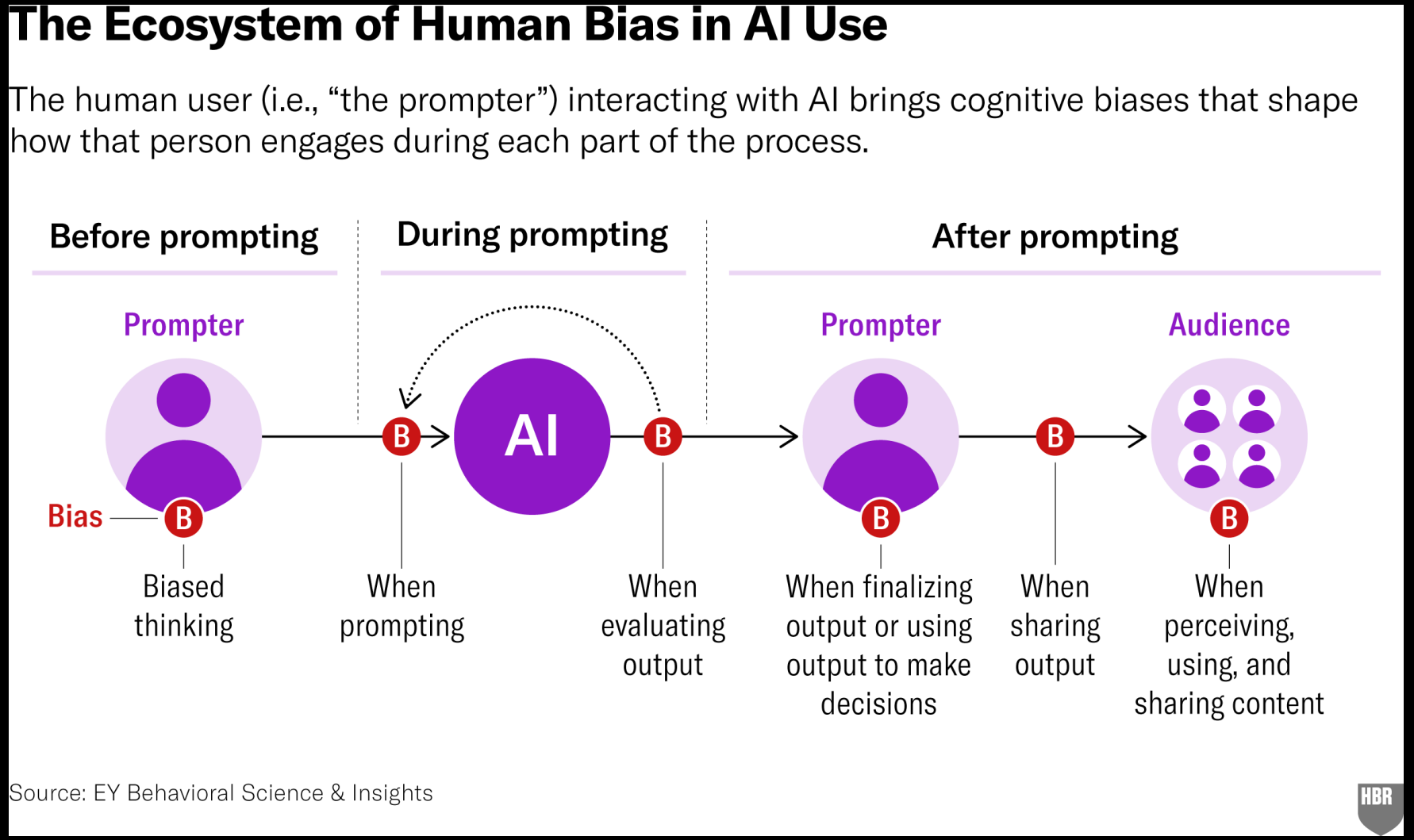
Enforces Policies through Technology

Ensures High-Fidelity data



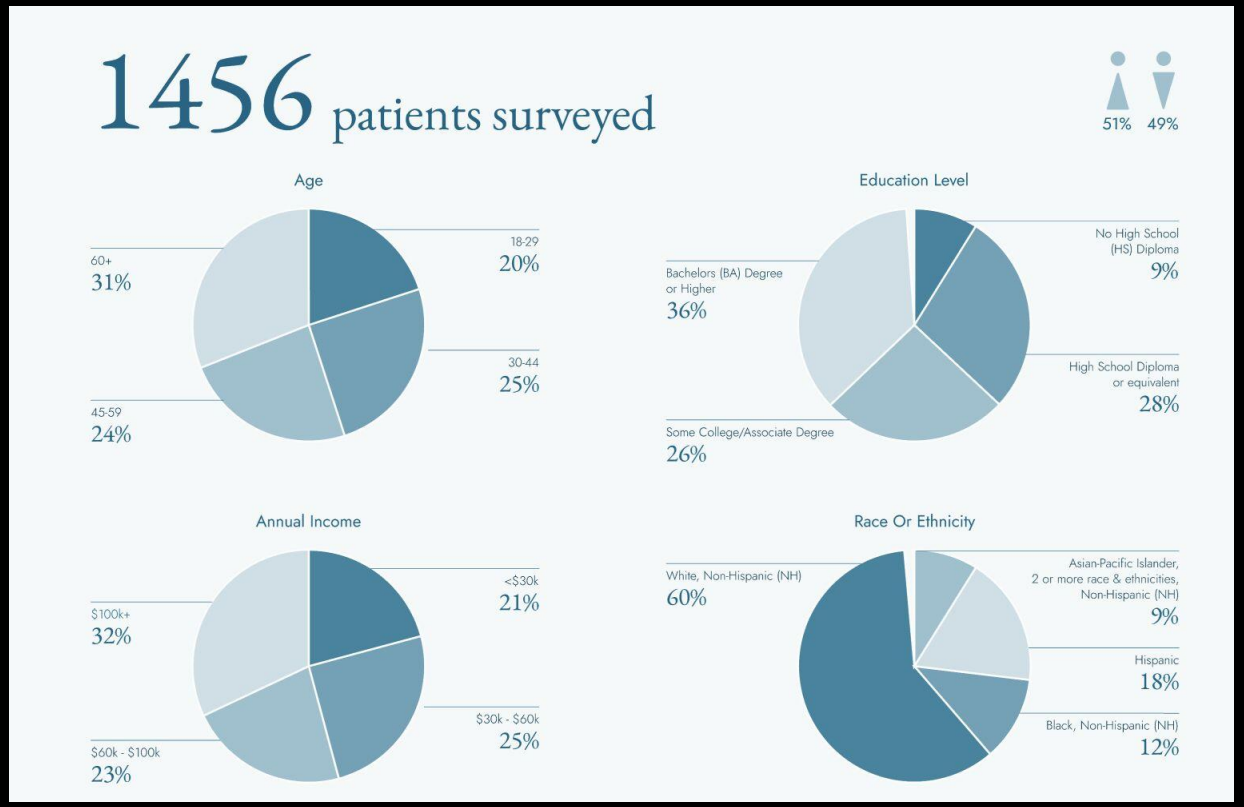
Protects Data & Model IP

How AI Can Amplify Biases of its Users



REF: https://hbr.org/2026/01/when-ai-amplifies-the-biases-of-its-users

Patient Survey on AI (Nov 2025)



Ref: California Health Care Foundation, NORC (Univ of Chicago), CHAI. Nov 2025

💻 **Concerns focus less on AI itself, more on governance, accountability, and patient protection**.

💻 Use, comfort, and trust are only modestly linked and highly context-dependent.

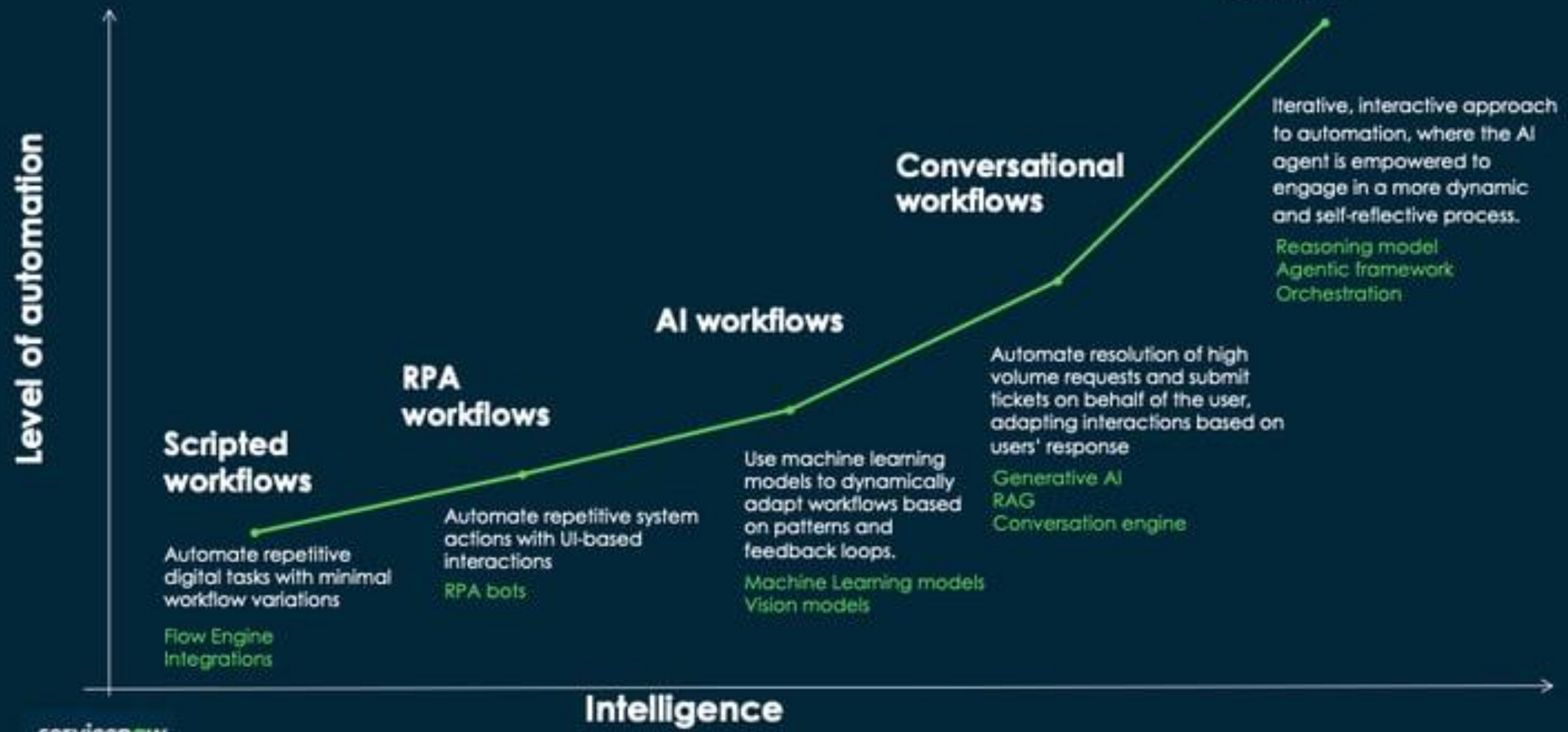
💻 Data commercialization concerns outweigh concerns about algorithmic bias. Notably 12% report never having considered AI bias at all.

💻 **Patients fear 'humans-out-the-loop' scenarios– critical in the current Agentic AI debate**

💻 Disclosure that AI is involved in care is crucial, but does not alone build trust; in some cases, it reduces trust.

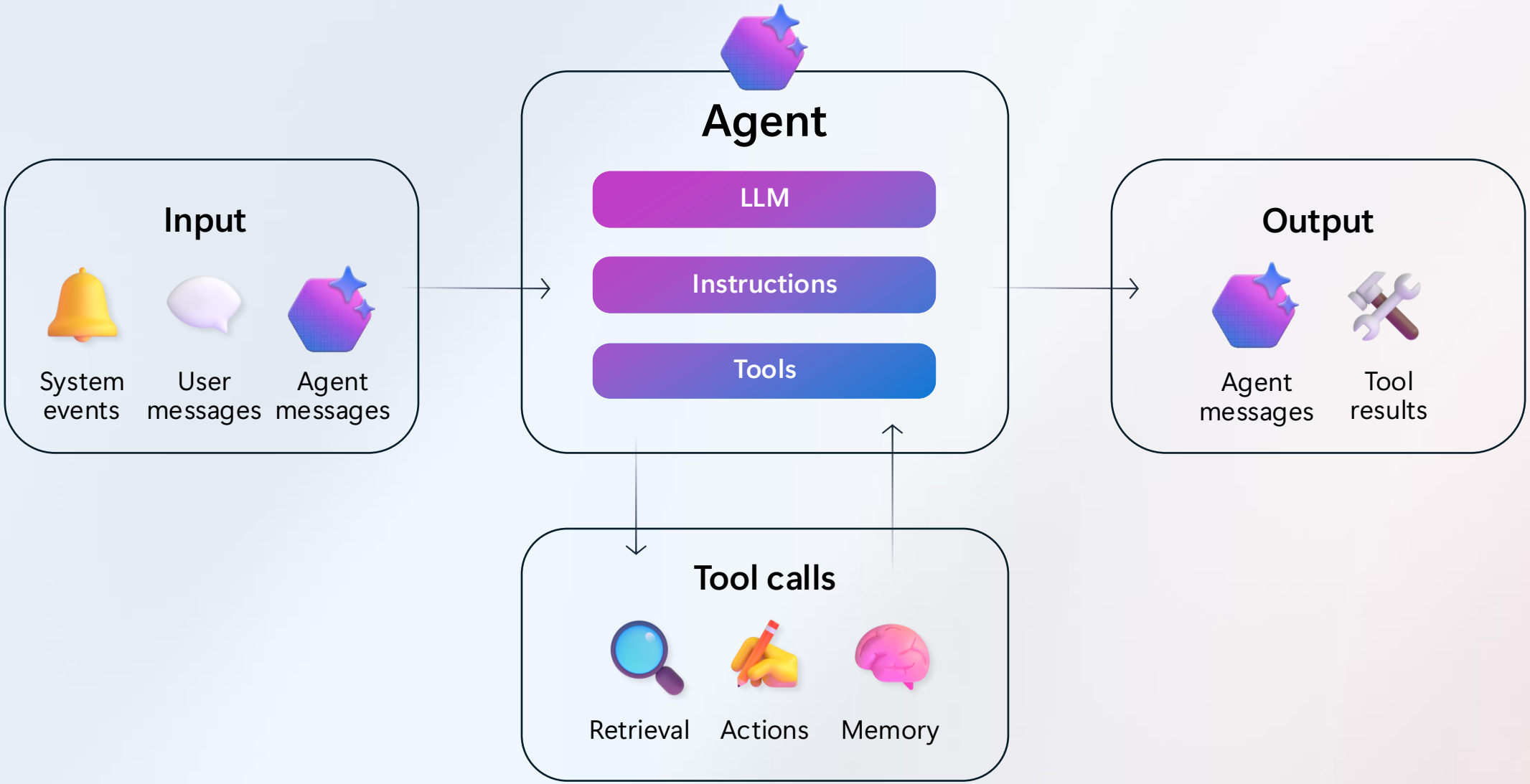
💻 No single institution is seen as a trusted overseer. Instead, people favor multi-layered governance across independent non-profits, health systems, and federal regulators.

Automation in enterprise workflows

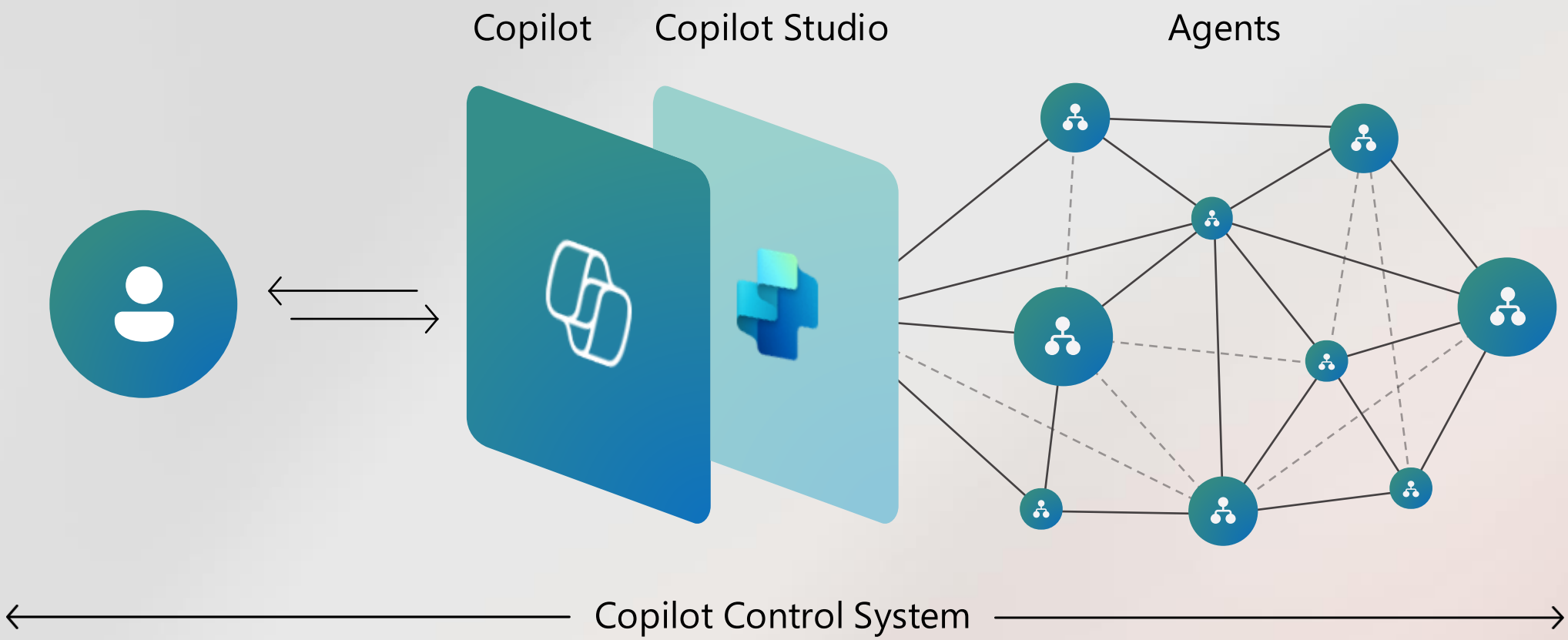


servicenow

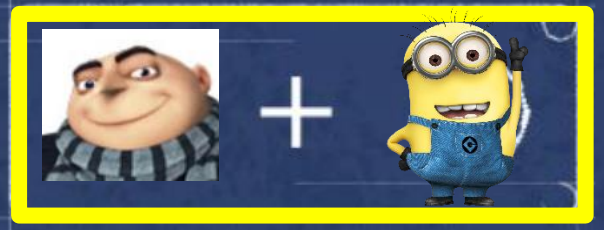
What is an agent?



Copilot is the UI for AI

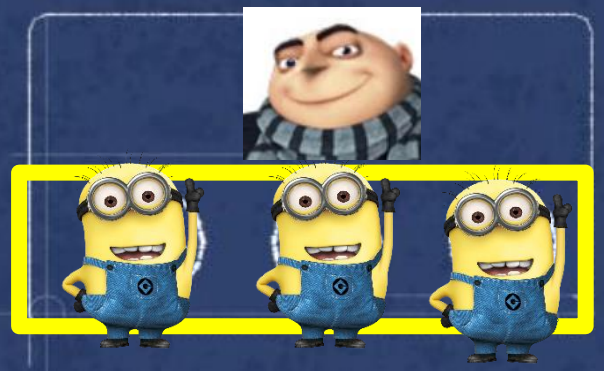


Persons + Agents



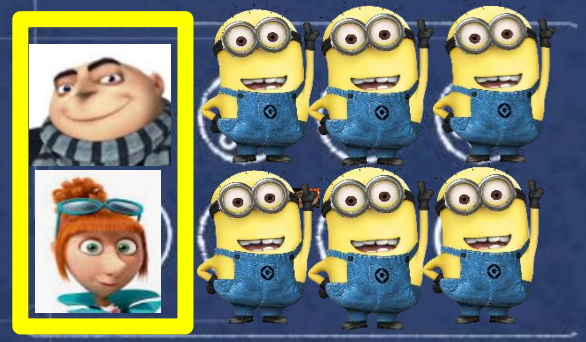
Phase 1
Human with assistant

Every employee has an AI assistant that helps them work better and faster



Phase 2
Human-agent teams

Agents join teams as "digital colleagues," taking on specific tasks at human direction



Phase 3
Human-led, agent-operated

Humans set direction and agents execute business processes and workflows, checking in as needed

The NEW ENGLAND JOURNAL of MEDICINE

CASE RECORDS OF THE MASSACHUSETTS GENERAL HOSPITAL

Case 18-2025: A 63-Year-Old Woman with Dyspnea on Exertion

Authors: Malissa J. Wood, M.D., Carola A. Maraboto Gonzalez, M.D., Reece J. Goiffon, M.D., Ph.D., Eric D. Jacobsen, M.D., and Bailey M. Hutchison, M.D. Author Info & Affiliations

Published June 25, 2025 | N Engl J Med 2025;392:2459-2470 | DOI: 10.1056/NEJMcpc2300897

VOL. 392 NO. 24 | Copyright © 2025

CME

PERFORMANCE RESULTS

Diagnostic Accuracy

80%
MAI-DxO AI
+302.0% better

19.9%
Human MDs

Cost per Case

$2,396
MAI-DxO AI
19.1% lower cost

$2,963
Human MDs

MAI-DXO CONFIGURATIONS:

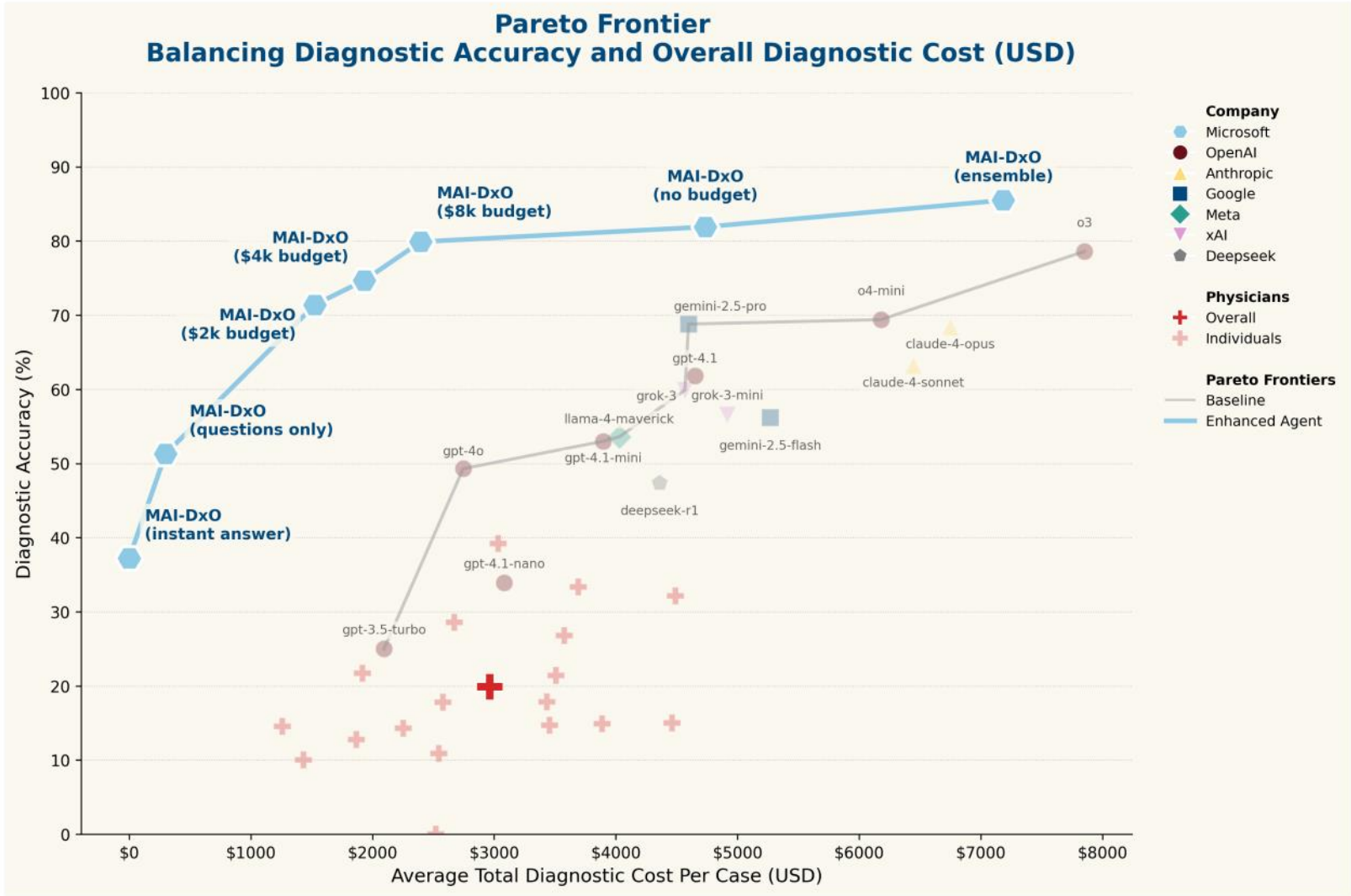
High Performance
85.5% accuracy

Standard
80% accuracy

Cost-Conscious
79.9% accuracy **$2,396**

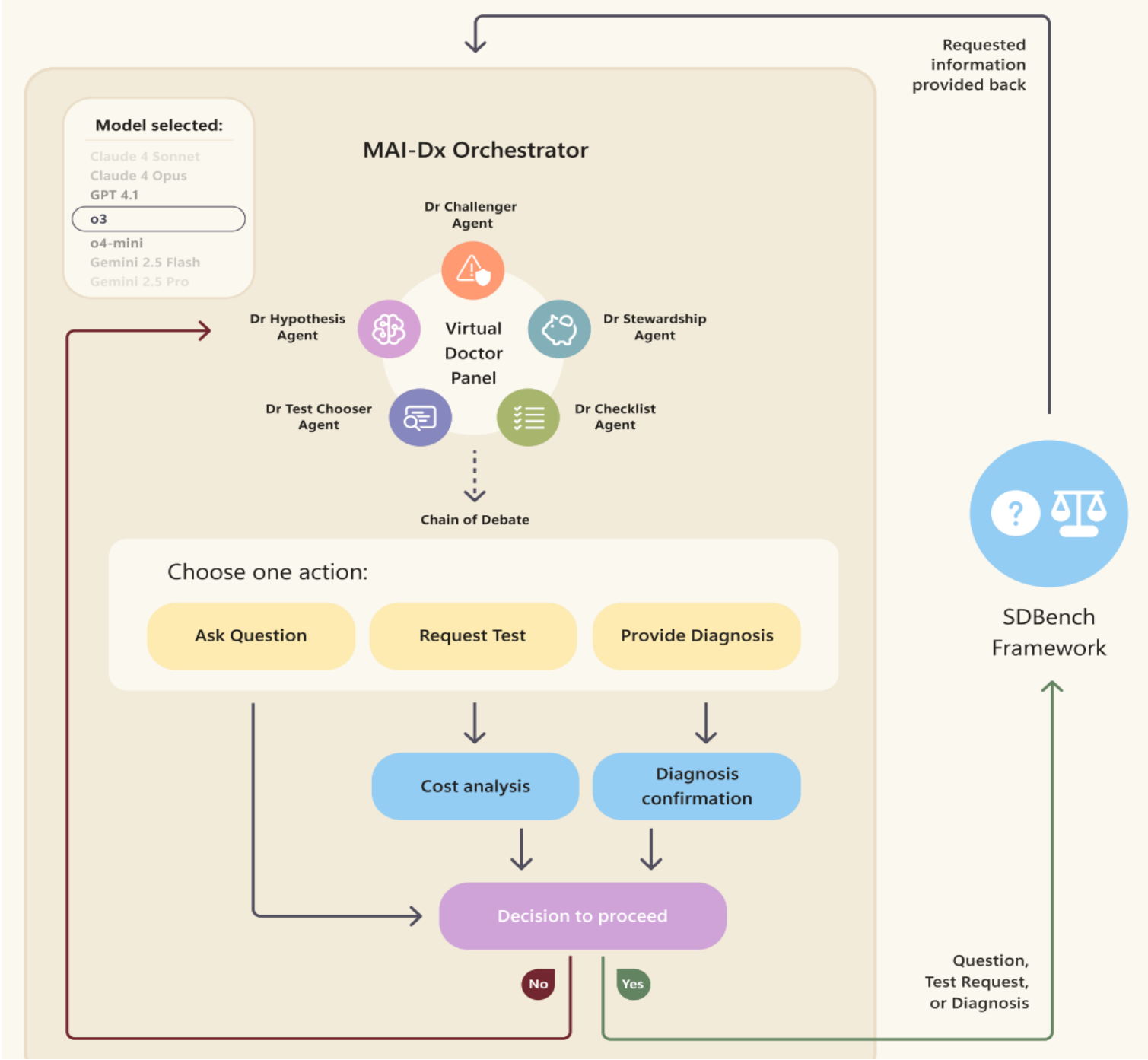
Study Scale

304 NEJM challenging cases 21 physicians, 12+ years experience



Nori et al. Sequential Diagnosis with Language Models. June 30, 2025

Medical AI Diagnostic Orchestrator



Top security and governance concerns about generative AI

Data oversharing and data leaks

80%

of leaders cited leakage of sensitive data as their main concern[1]

Controlled access to patient data

Identification of risky AI use

41%

of security leaders cited that the identification of risky users based on queries into AI was one of the top AI controls they want to implement[2]

Detecting PHI De-Identification

AI governance and risk visibility

84%

Want to feel more confident about managing and discovering data input into AI apps and tools[2]

Data and agent lineage tracking

1. First Annual Generative AI study: Business Rewards vs. Security Risks, Q3 2023, ISMG, N=400
2. Microsoft data security index 2024 report

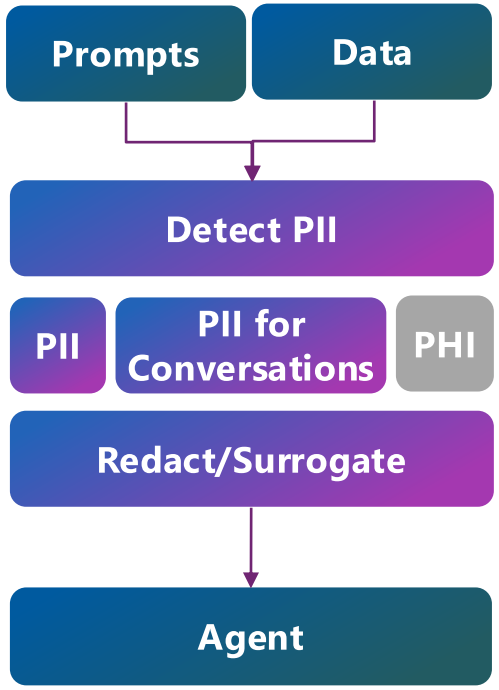
Controlled access to patient data

Microsoft Entra Agent ID (preview)

- Authentication
- Authorization
- Identity Protection
- Access Governance
- Visibility

Detecting PHI De-Identification

Azure AI Language

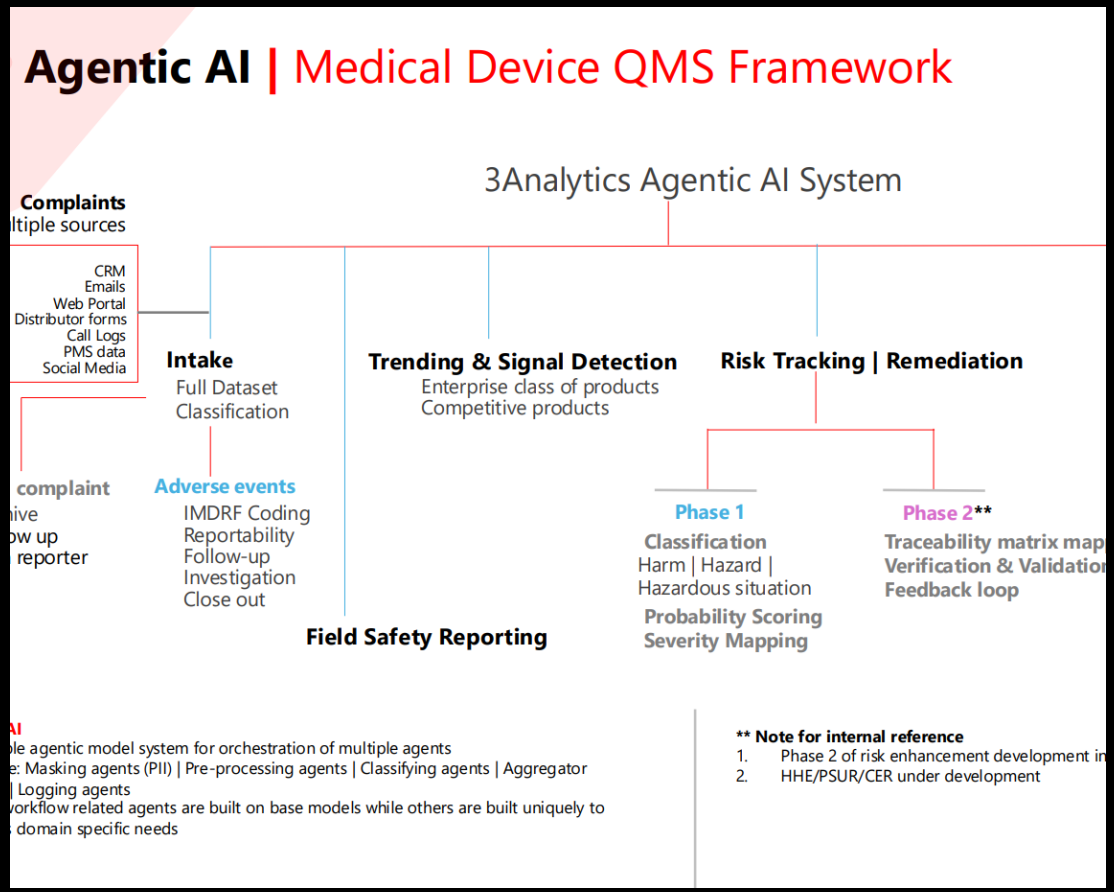


Data and agent lineage tracking

- BYO Evaluators & Monitoring
- Agent Admin Approval
- Agent Metadata Review
- Tenant & User Control
- Agent Removal & Cleanup
- Agents Inventory
- Agents Usage
- Data Access Logging in Purview

RWE and Agentic AI

Framework



Results

Improved Accuracy & Consistency
>90% accuracy and >95% consistency in HSHA extraction

Audit-Ready Outputs
Each extracted HSHA output includes justification text and confidence scores for SME review and audit readiness

Faster HSHA Extraction
95% of Effort Reduction in Complaint Analysis (from **~8–10 minutes to under 1 minute** per complaint)

Enhanced SME Productivity
~60% reduction in SME workload, as only borderline or low-confidence cases are escalated for validation

Scalable for Growth
The AI agentic pipeline processes thousands of complaints in batch mode without accuracy loss, enabling scalability of high-volume complaints

Foundation for Advanced Complaint Risk Scoring
Structured HSHA data enables future enhancements such as automated P1/P2/pHarm calculation, traceability to product requirements, and risk managem

Example Agent Use Case: Tumor Board Meeting

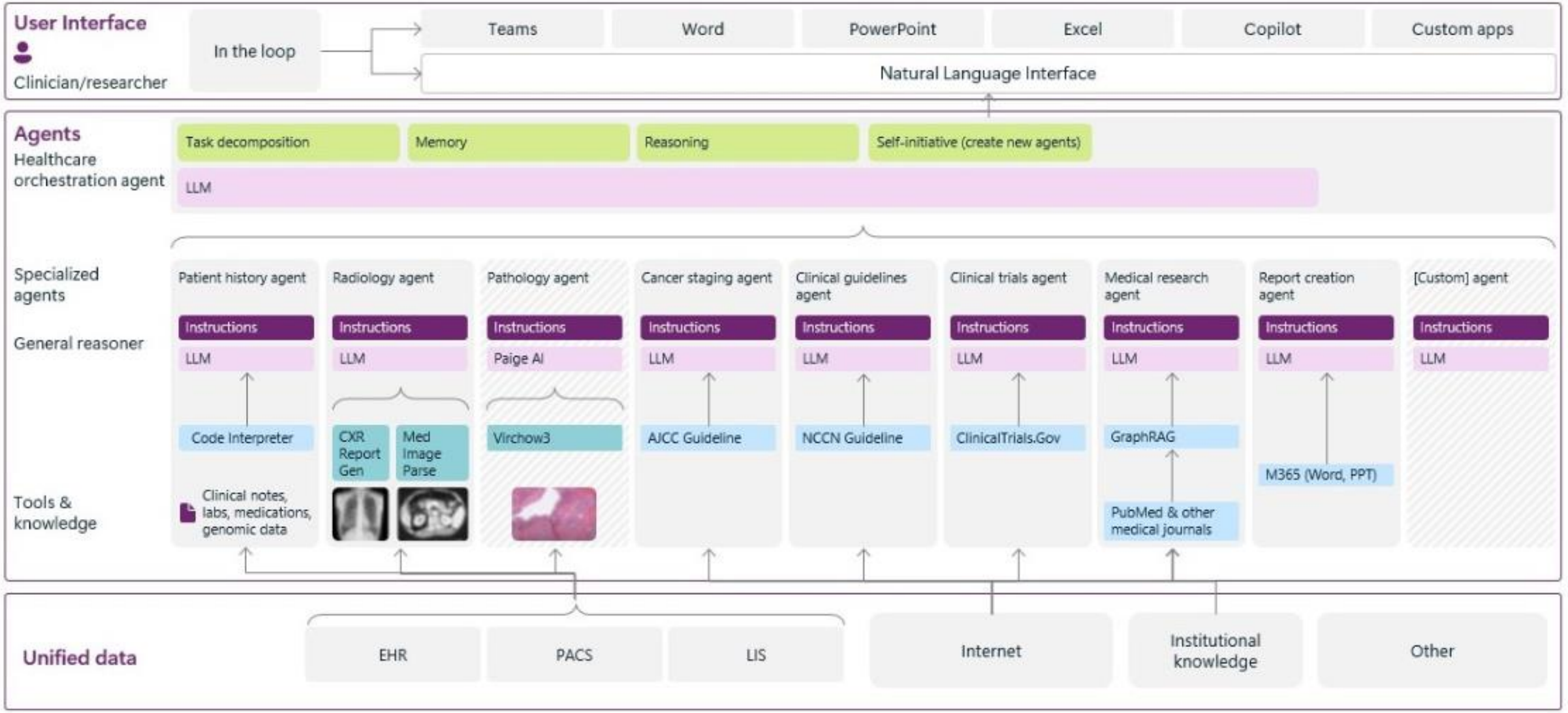
10+ specialists in a room.

Goal: Reason over domain specific data to **provide next steps for cancer treatment.**

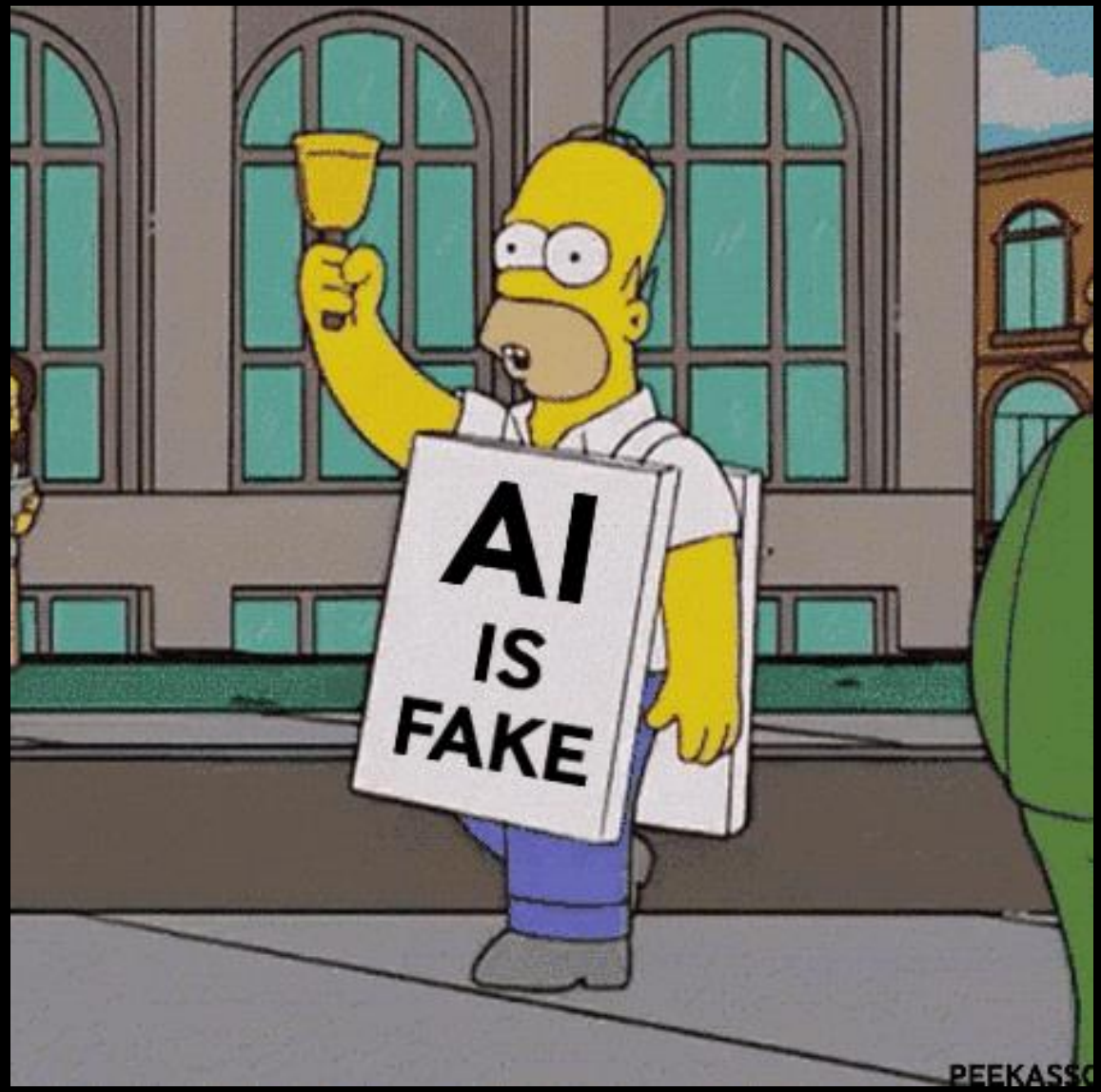
Demonstrated to **improve patient outcomes.**

Most moved to **virtual (Teams)** since Pandemic

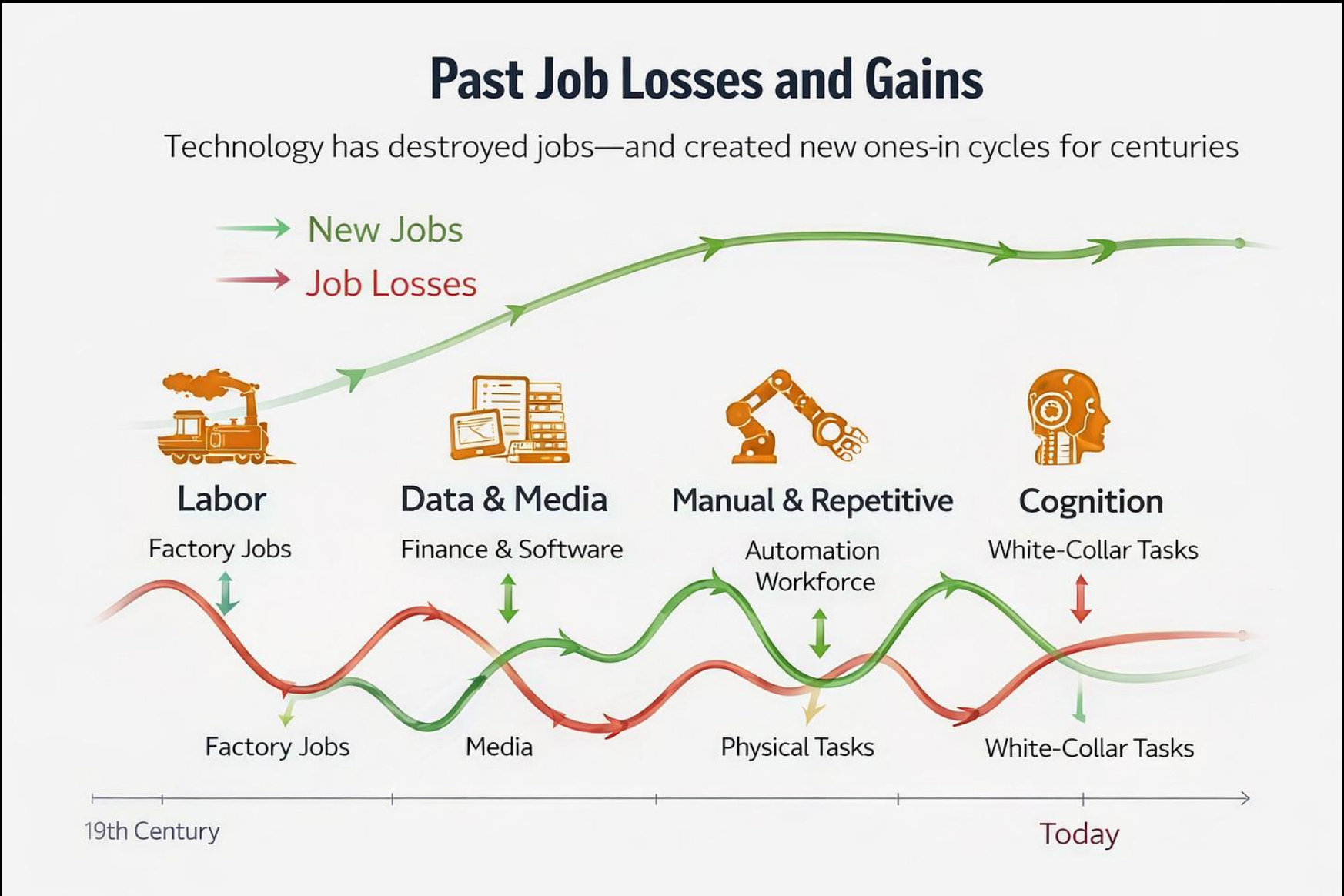
Healthcare Agent Orchestrator



Will AI Take Away my Job?

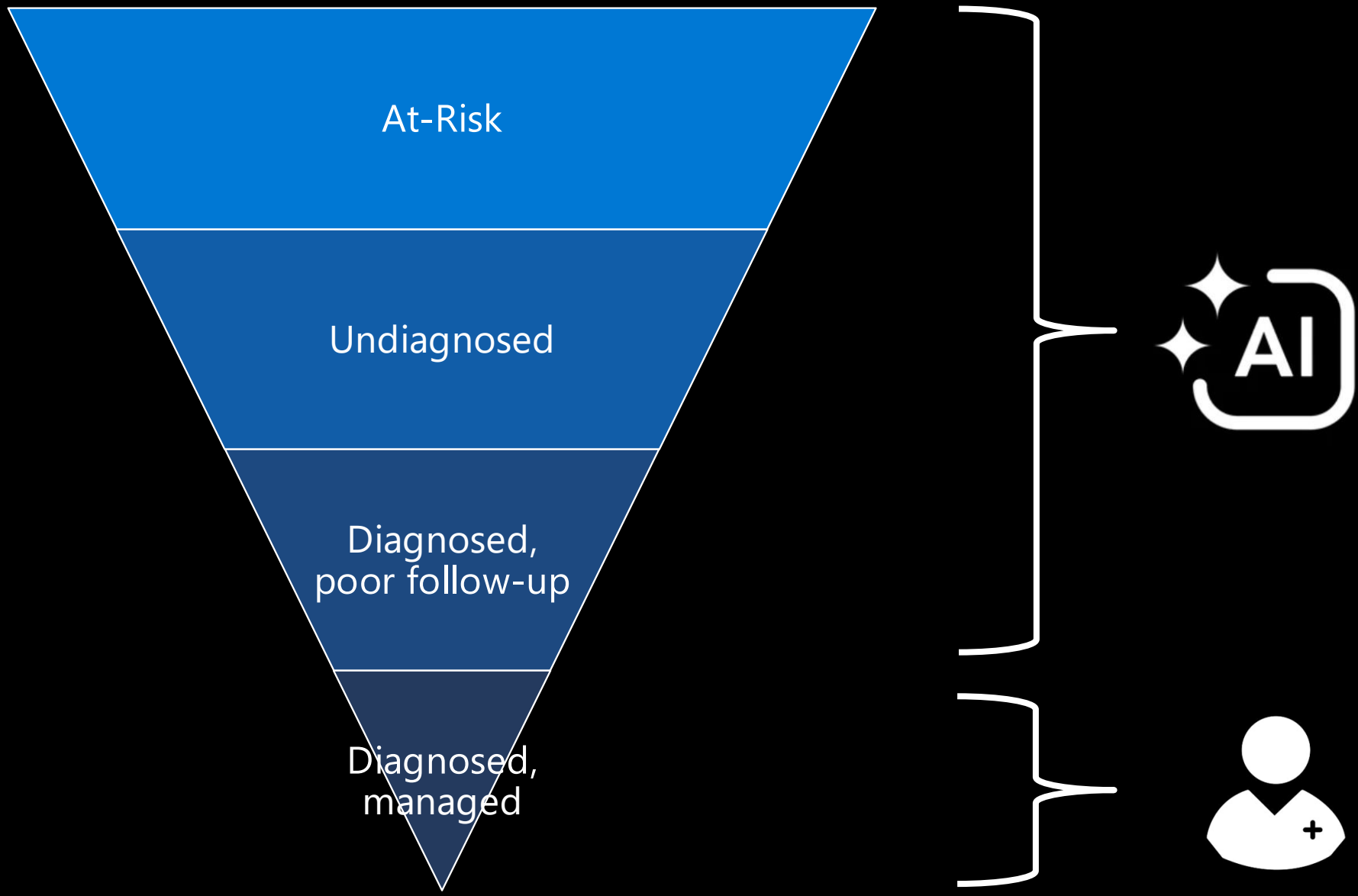


Fear of Job Loss



https://hoyemgeorge.substack.com/p/ai-jobs-and-the-jevons-paradox-why

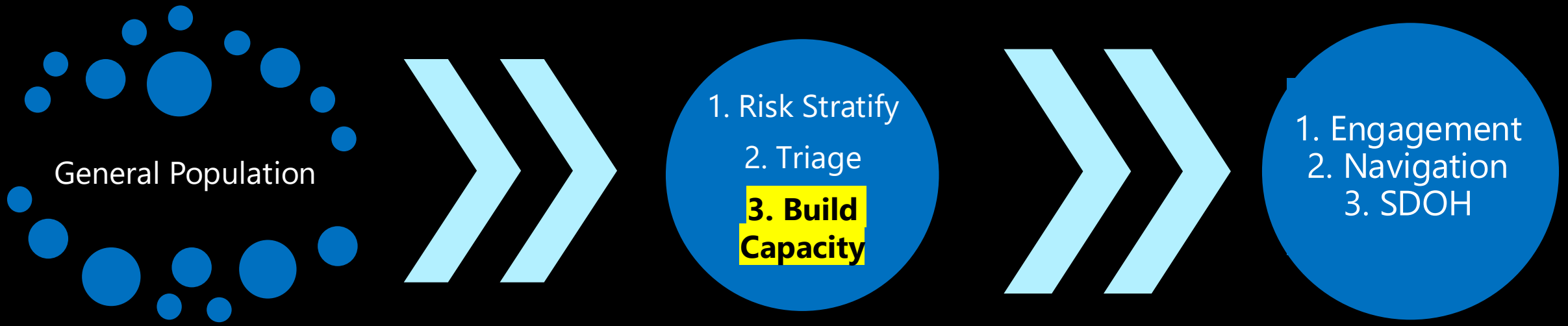
Entry Points to Care Delivery



Healthcare From The Eye

- DIABETES
- CARDIOVASCULAR DISEASES
- NEUROLOGICAL DISEASES
- EYE DISEASES

Population Health & AI



Identify Disease

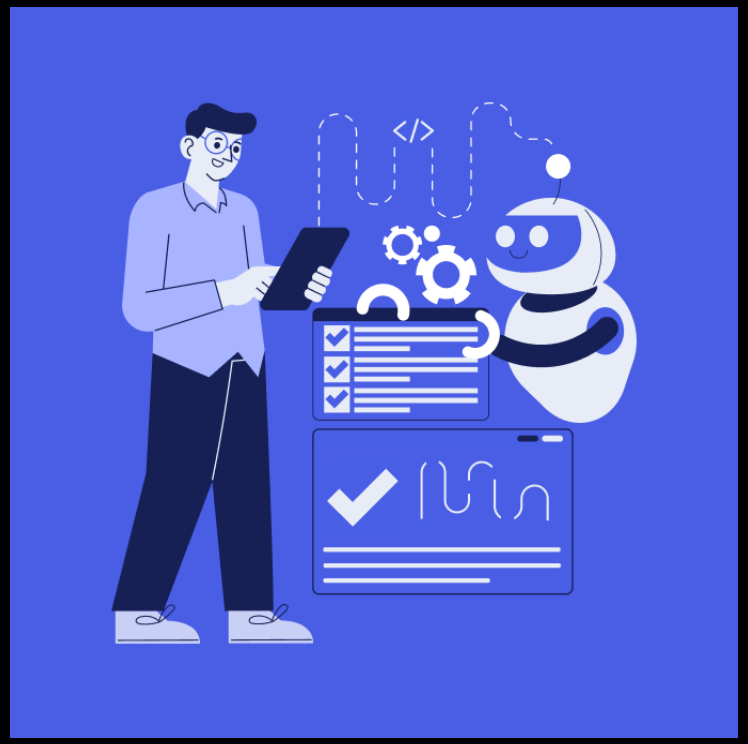
Manage Disease

Engage Patients

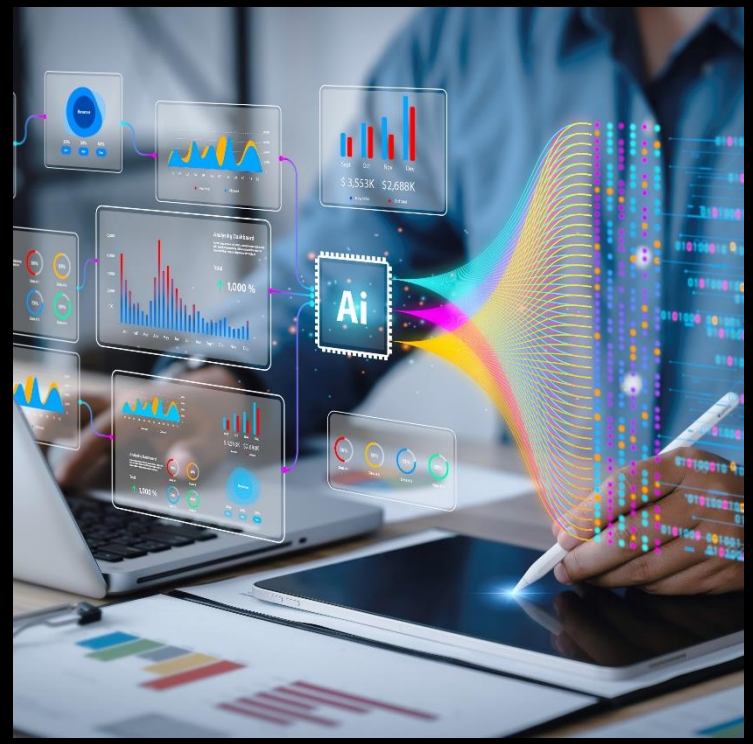
AI
Imparting us with knowledge seamlessly and efficiently

AI Skilling & Reskilling

Curriculum



Training



Jobs

Key Points

01 Pressing need for Responsible AI

02 CHAI provides framework

03 TRAIN provides tools

04 Agents need Governance

05 Society needs AI Upskilling

Microsoft

Thank You!