SI Metrology: Ensuring the Reliability of Super Intelligent Systems



Ram D. Sriram, Ph.D.

Senior Science Advisor

Information Technology Laboratory (ITL)

National Institute of Standards and Technology (NIST)

URL: http://www.nist.gov/itl/ssd/, sriram@nist.gov

2026-10-22

Art courtesy: N. Hanacek/NIST

Executive Order to Use SI

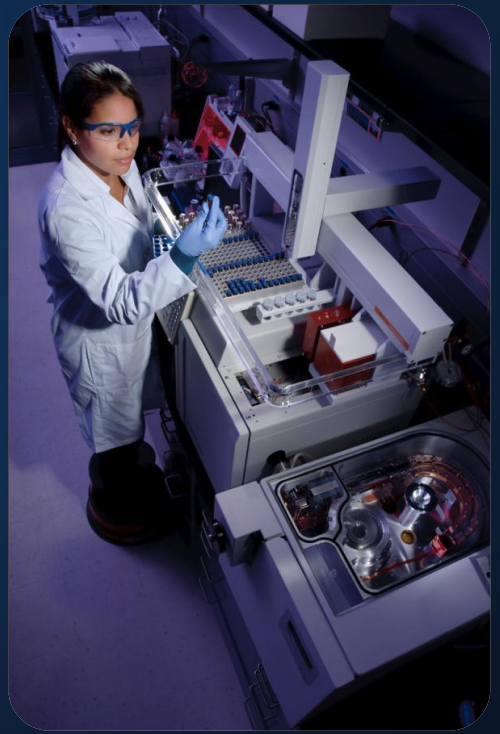
On Sept. 29, 2026, the President's Executive Order on Inaugurating the Era of Super Intelligence directed executive agencies to update its use of "artificial intelligence" to "super intelligence."

Outline

- NIST and SI (AI)
- SI (AI) Evolution
- SI Metrology and Governance
- Future Work
- Summary

NIST's Mission

To promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life.



©R. Rathe

Multiple SI Functions Exist at NIST Today

Information Technology Laboratory (ITL)	NIST SI/AI Consortium	Center for Advancing Innovation and Standards for Super Intelligence(CAISSI)	SI/AI Accelerators
• Leads developing measurements, testing, evaluations, and standards in SI (AI) • 40+ years of deep SI (AI) expertise & applied research • Stewards the SI (AI) Risk Management Framework	• 280+ industry, academic and non-profit orgs 200+ new applicants as of June 2026 • Expanded scope in May 2026 • Prioritize research in measurement science, innovation and adoption	• Work with NIST orgs to develop guidelines to measure and improve SI/AI security • Re-established from AISI in June 2025 • Evaluations and assessments of US and adversary SI (AI) capabilities	• Remove barriers to U.S. SI (AI) innovation and accelerate the application of AI tech • Adoption of SI/AI-driven autonomous agents for increased U.S. manufacturing productivity • Adoption of SI/AI-based agents for ultra-highspeed cyberthreat detection and remediation

NIST

Slides on NIST: Courtesy Craig Schlenoff

ITL's SI (AI) Work is Organized Across 4 Impact Areas

#1 Transforming how SI (AI) systems' trustworthiness is measured (TEVV)

- ❑ GenSI/AI Program
- ❑ ARIA (Assessing Risks and Impacts of SI/AI)
- ❑ Building Evaluation Probes into Agentic SI (AI)
- ❑ Dioptra

#2 Empowering industry and government to make informed decisions

- ❑ NIST SI/AI Risk Management Framework
- ❑ NIST SI (AI) Resource Center

#3 Positioning the U.S. to dominate in the SI (AI) technical standards arena

- ❑ "Zero Drafts" Pilot Project
- ❑ A Plan for Global Engagement on SI/AI Standards
- ❑ Technical Contributions to SI/AI Governance

#4 Enabling the U.S. to be at the forefront in applying SI (AI) to high-priority domains and sectors

- ❑ Development of cybersecurity standards, guidelines and best practices
- ❑ Accelerate adoption of SI/AI-driven agents for increased U.S. manufacturing productivity

SI (AI) Agent Standards Initiative

CAISSI, ITL, and other interagency partners are collaborating to advance trust, interoperability, and security for SI/AI agents.

Pillar 1: Facilitate industry standards

- Host technical convenings and conduct gap analyses to produce voluntary guidelines to inform industry-led standardization for SI/AI agents
- Maintain leadership on SI/AI agents at international standards bodies

Pillar 2: Foster community-led protocols

- Engage with SI (AI) ecosystem to identify and reduce barriers to interoperable agent protocols
- Invest in development and security of open source ecosystems including SI/AI agent protocol ecosystems

Pillar 3: Invest in research

- Conduct research on agent authentication and identity infrastructure to enable secure human-agent and multi-agent interactions
- Develop SOTA security evaluations to inform protocol development

NIST

Outline

- NIST and SI (AI)
- SI (AI) Evolution
- SI Metrology and Governance
- Future Work
- Summary

Theories of Human Intelligence

Multiple Intelligences

Howard Gardner: Intelligence consists of 8+ independent modular faculties (e.g., Logical-Mathematical, Spatial, Interpersonal) rather than a single score.

Triarchic Theory

Robert Sternberg: Focuses on "successful intelligence" through three core mechanisms: Analytical, Creative (novelty), and Practical (contextual adaptation).

CHC Taxonomy

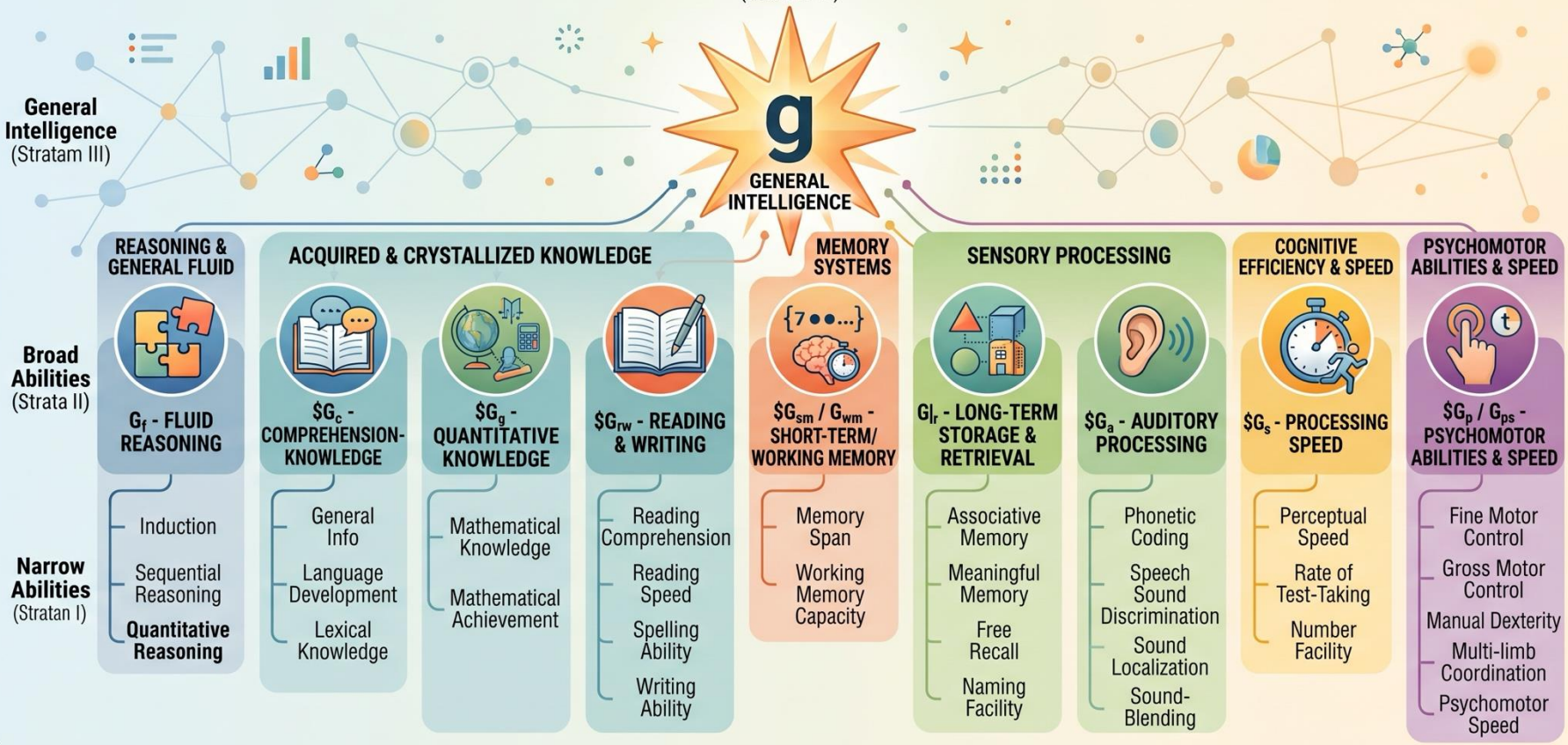
Cattell-Horn-Carroll: Hierarchical taxonomy with General Intelligence (*g*) at the top, Broad Cognitive Layer (*Gf*, *Gc*, G*sm*), and Narrow Specific Abilities.

Application to SI/AI: Provides modular benchmarks for domain-specific evaluations and contextual adaptation.

Application to SI/AI: Serves as a standard blueprint for multi-dimensional AI benchmark design.

CATTELL-HORN-CARROLL (CHC) THEORY OF INTELLIGENCE

(Stratum III)



Source: https://onlinelibrary.wiley.com/doi/epdf/10.1002/9781118660584.ese0431; Figure Gemini Generated

From Limited to General Intelligence

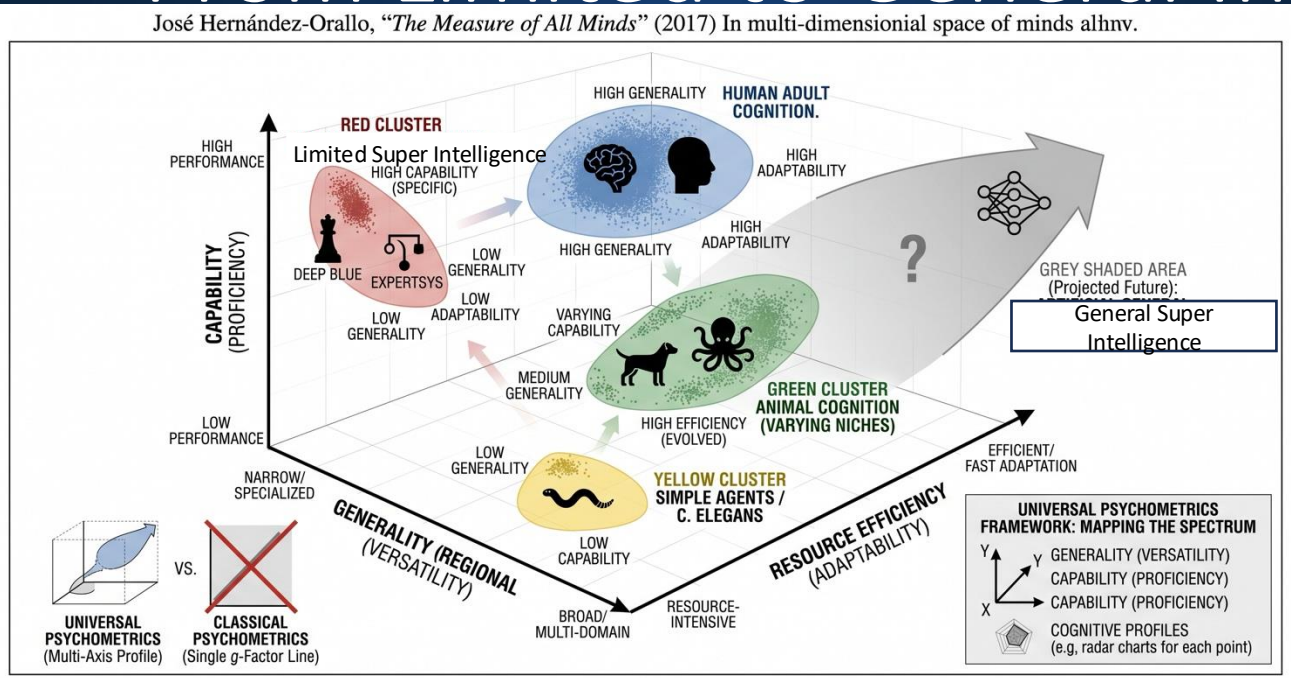
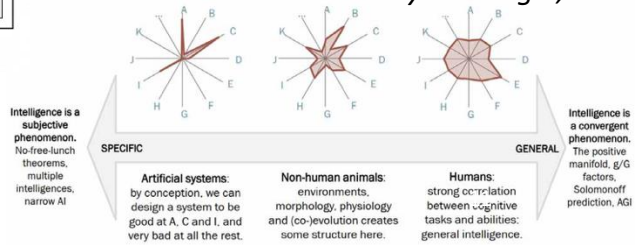


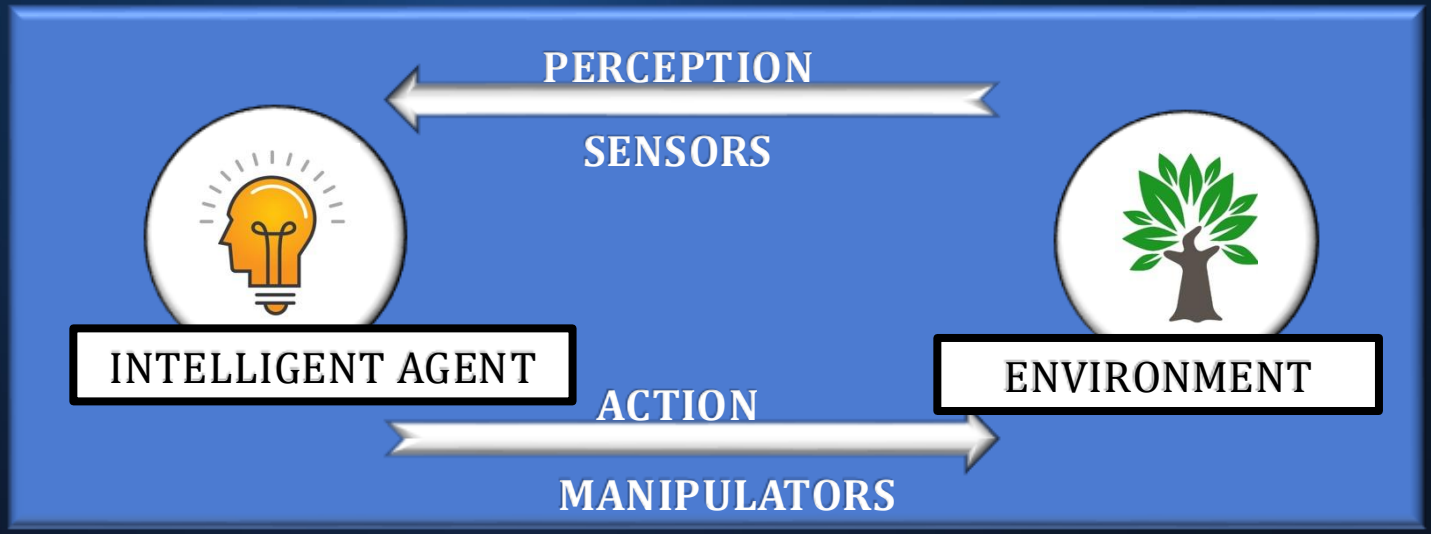
FIGURE 4.3: THE MULTI-DIMENSIONAL LANDSCAPE OF UNIVERSAL PSYCHOMETRICS.

Courtesy: Holzinger, et al.



Super Intelligence: A Simple View

Super intelligence is concerned with the development of computer programs that can equal or surpass human intelligence.



Three Tiers of SI Evolution

Limited SI (LSI)

Waves 1 - 3 (Prehistory to 2025)

Specialized for defined tasks in limited domains. Learns within strict boundaries.

Pre-1956 Foundations: Philosophies & Turing.

1957-80 Early Knowledge Systems: GPS, Mycin, Shakey.

1980-90 1st Wave: Knowledge-based Expert Systems

2000-2020 2nd Wave: Neural Networks and Deep Learning revolution.

2020- 3rd Wave: Symbiosis of Neural/Knowledge Nets & XAI.

General SI (GSI)

The 4th Wave (2025 - ??)

Possesses or surpasses human-level intellect across all broad cognitive categories.

- The Conscious Machine: Achieving a state where machine cognition mirrors human adaptability.
- Cross-Domain Transfer: Ability to seamlessly apply knowledge learned in one field to completely unrelated fields.
- Novel Reasoning: Can autonomously reason and plan long-term in highly unpredictable, novel scenarios.

Transcendent SI (TSI)

The Tsunami

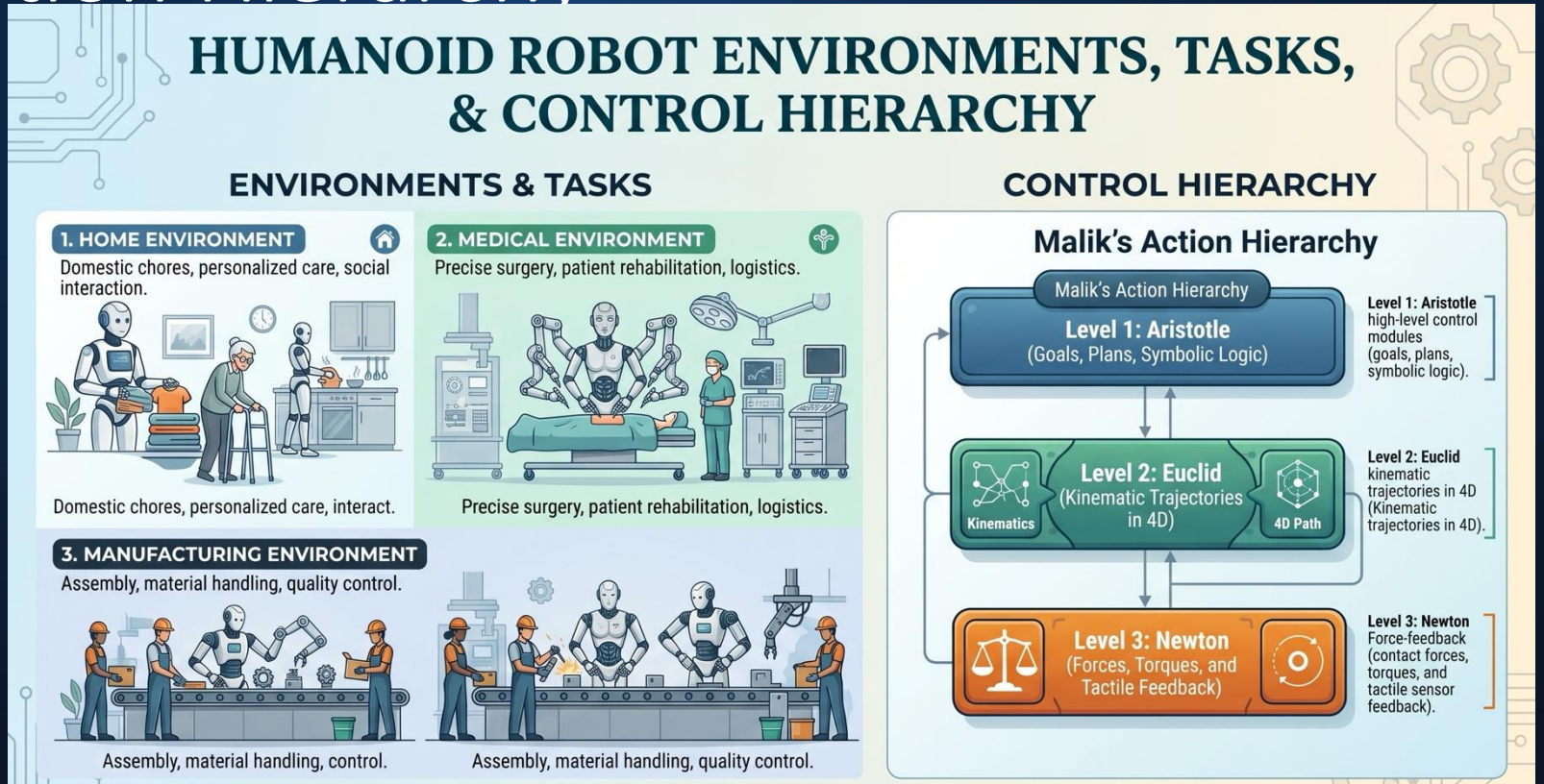
A hypothetical state surpassing the brightest human minds in every dimension.

- Recursive Improvement: AI designing better AI, triggering an intelligence explosion or "Tsunami".
- Autonomous Discovery: Capable of driving independent scientific discovery beyond human comprehension.
- Infinite Scale: Operates at cognitive speeds and capacities fundamentally impossible for biological brains.

Embodied/Physical Intelligence

Embodied LSI (Current State)	Embodied GSI (General Purpose Humanoid)	Embodied TSI
Industrial robotic arms or narrow bipedal walking algorithms. These systems execute predefined trajectory paths or narrow pick-and-place routines under highly structured factory conditions but fail when exposed to unexpected terrain or novel physical objects.	A humanoid robot capable of entering an unfamiliar human home, understanding natural language commands (e.g., "Clean up the spilled drink in the kitchen"), navigating unmapped physical obstacles, dynamically adjusting its grip force for fragile glassware, and physically adapting to unexpected floor slippery conditions without losing balance.	Physical platforms operating with superhuman reaction times, micro-meter kinematic precision, multi-agent physical swarm coordination, and hyper-efficient material construction capabilities.

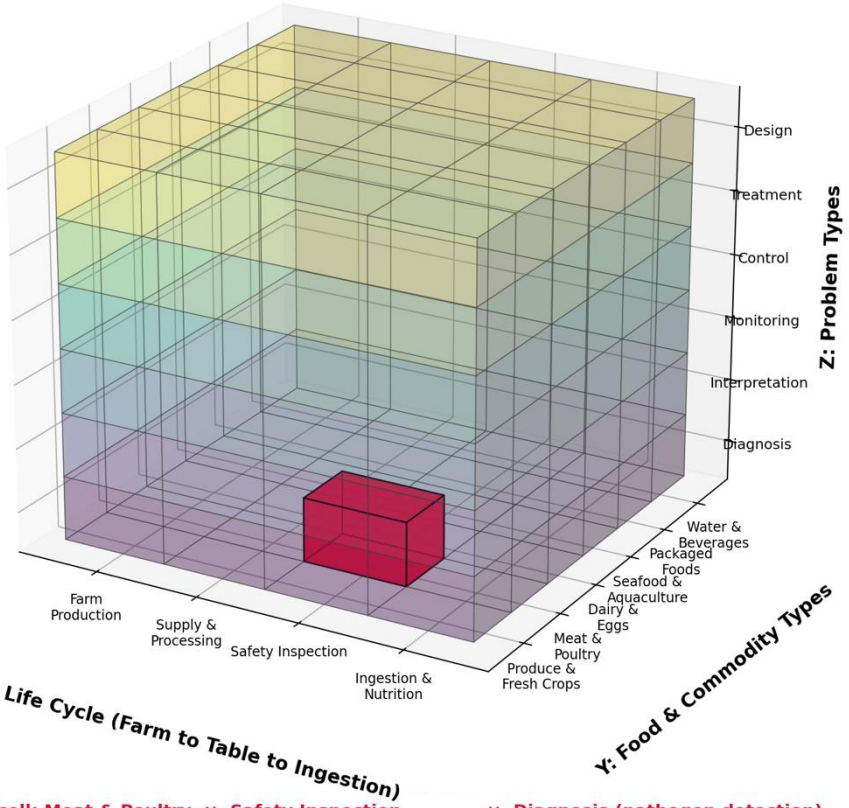
Humanoid Robots: Malik's Three Level Action Hierarchy



Gemini Generated

https://www.youtube.com/live/HvDU7Vk4pbc

Where Can SI (AI) Be Used? A Three-Dimensional Map of the Food System



Highlighted cell: Meat & Poultry × Safety Inspection × Diagnosis (pathogen detection)

With help from Claude AI

Outline

- NIST and SI (AI)
- SI (AI) Evolution
- SI Metrology and Governance
- Future Work
- Summary

SI Metrology and Governance

- Metrology for SI: Brief Introduction
- Characterizing Uncertainty in Deep Learning Models
- Determining Accuracy for Stem Cell Transplantation
- Evaluating Agentic SI (AI)
- Detecting Trojans
- Risk Management

The Case for SI Metrology

- Lord Kelvin “When you cannot measure something and express it in numbers, your knowledge is of a meager and unsatisfactory kind.”
- Popular adaption “If you cannot measure it, you cannot improve it.”
- A New Metric for Measuring Intelligence – Shane Legg (co-founder of DeepMind)
  - https://youtu.be/V6umr1OP8uo?si=bWOBfqqVI6pTcuYJ

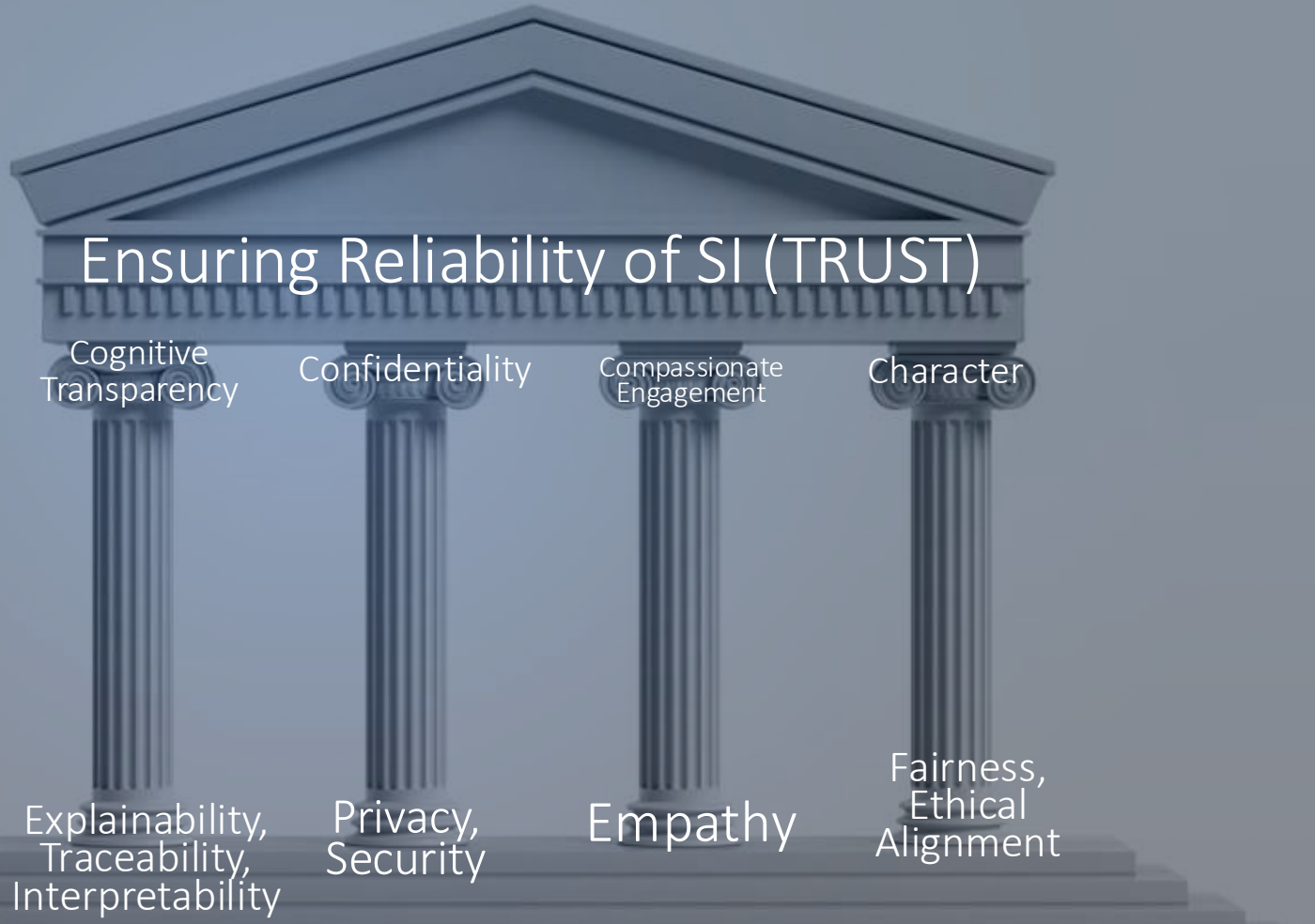
SI Metrology: The Science of Measurement (Features)

- Accuracy: These metrics generally depend on the domain and task involved, but may be further classified into: robustness which explores the resilience of systems against perturbations and adversarial attacks; hallucination rate which indicates how often model invents facts; conciseness in text retrieval tasks; semantic correctness which ensures systems understand underlying semantics; omission rate which is frequency with which SI fails to retain or surface critical information present in the source data (e.g., omitting a patient's allergy history during clinical summarization); groundedness with respect to how well the system does on previously proven results and supported by trusted sources (e.g., PubMed).

- Performance: These metrics are essential in assessing the runtime performance. Performance metrics can be further classified into memory efficiency, Floating point operations, SI (AI) model parameters needed.

NIST

Ensuring Reliability of SI (TRUST)



SI Metrology: The Science of Measurement (Features)

- Uncertainty Quantification: This metric would evaluate the confidence level of an AI's prediction (e.g., ability to deal with input that the system has not been trained on before). There may be different types of uncertainty quantifications for SI (AI) systems. For general guidelines about measurement uncertainty see https://www.nist.gov/itl/sed/topic-areas/measurement-uncertainty.

- Risk: There is a risk associated with any piece of software. Hence, it is important to develop measures to evaluate risk associated with use of SI (AI) tools. NIST has developed a framework called SI/AI RMF (see https://www.nist.gov/itl/ai-risk-management-framework) for aiding in managing risk.

Developing standardized methods for evaluating each of these are important

Human-in-the-Loop Governance

Evolution of Human Oversight

LSI (Oracle): Humans provide live ground-truth checks to calibrate accuracy and measure override rates.

GSI (Arbiter): Humans evaluate symbolic rule compliance to prevent over-trust or under-trust.

TSI (Auditor): Direct code auditing is impossible; humans shift *on the loop* to spot-audit SI critique teams.

With help from Gemini AI

Metrology Across the Spectrum: Human in the Loop

Tier	Core Metrology Target	HITL Human Role	Key HITL Interaction Metrics
LSI	Accuracy, Robustness, FLOPS (Efficiency), Uncertainty Quantification (UQ)	Ground-Truth Oracle	Correction Rate, Time-to-override
GSI	Dual-Layer Governance, Ontological Mapping, Garland Test	Symbolic Arbiter & Trust Calibrator	Rule-level Agreement, Calibrated Reliance
TSI	Multi-Agent Scaffolding, Trajectory Tracking	On-the-Loop Auditor	Spot-audit Coverage, Escalation Latency

NIST

With help from Gemini AI

Metrology for Embodied SI (AI)

Type of Intelligence	General Evaluation Metrics	Embodied/Physical AI Metrics
Limited Super Intelligence	Accuracy, Performance, Trust, UQ.	Focuses metrology on Level 3 ("Newton") force/torque loops, tactile skin integration, and multimodal 3D reconstruction accuracy (SAM 3D)
General Super Intelligence	Neuro-Symbolic Governance & Ontologies.	Validates how Level 1 ("Aristotle") symbolic plans constrain Level 2 ("Euclid") trajectories and Level 3 ("Newton") physical execution

NIST

With help from Gemini AI

Shane Legg and Marcus Hutter's UI Equation

The Universal Intelligence Equation

$$\Upsilon(\pi) := \sum_{\mu \in E} 2^{-K(\mu)} V_\mu^\pi$$

Variable Breakdown

- $\Upsilon(\pi)$ **(Universal Intelligence):** The overall intelligence score of an agent running policy/strategy π.
- $\sum_{\mu \in E}$ **(Sum over all environments):** Aggregates the agent's performance across all possible computable interactive environments E.
- V_μ^π **(Expected Value/Reward):** The total cumulative discounted reward agent π achieves when interacting with environment μ.
- $K(\mu)$ **(Kolmogorov Complexity):** The length of the shortest computer program that describes environment μ.
- $2^{-K(\mu)}$ **(Algorithmic Complexity Weight):** An algorithmic probability weight based on Solomonoff induction. Simpler environments (small $K(\mu)$) receive significantly higher weight, while overly complex, chaotic, or noisy environments (large $K(\mu)$) are exponentially down-weighted.

An agent's average ability to achieve goals across all possible environments, weighted inversely by how complex each environment is to describe.

Some Critiques:

1. Strictly "Incomputable"
2. Ignores computational Efficiency
3. Subjectivity of Turing Machine
4. Disconnected from Biology and Psychology

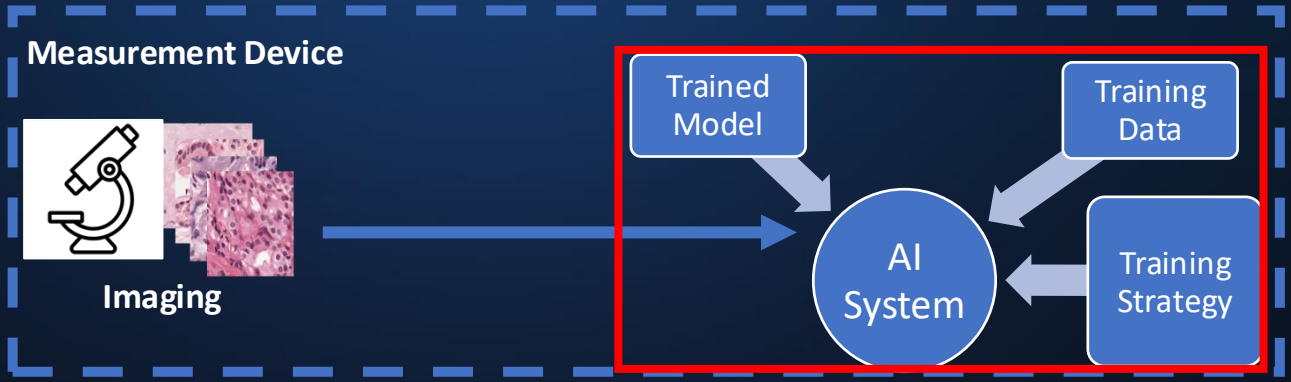
Some food for thought for DeepMind Institute: https://institute.deepmind.com/

SI Metrology and Governance

- Metrology for SI: Brief Introduction
- Characterizing Uncertainty in Deep Learning Models
- Determining Accuracy for Stem Cell Transplantation
- Evaluating Agentic SI (AI)
- Detecting Trojans
- Risk Management

Uncertainty Quantification

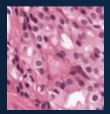
- **Foundational metrology research**
  - Transform AI analysis results into measurements (± uncertainty)
- Derive uncertainty & confidence metrics of AI systems
  - Analogy to physical metrology
- Associate uncertainty measurements with AI analysis results



Courtesy: Peter Bajcsy et al.

Why Are AI-Enabled Devices Unreliable?

- Trained AI Model – "visible" part of AI-Enabled System.
- Hidden parts – training data & training strategy.
- "Uncertainty Quantification" relative to AI system.

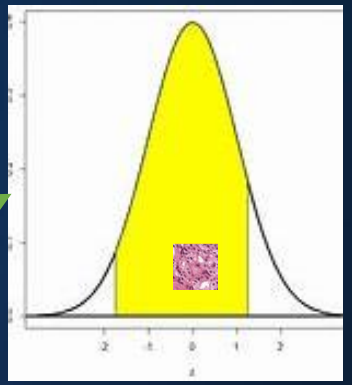


Sample

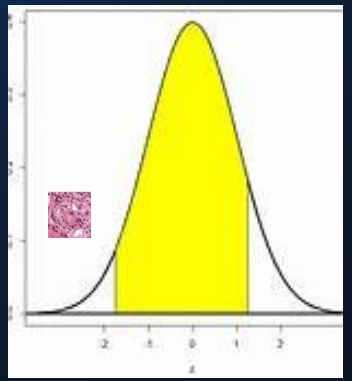
Trained AI Model

In Distribution

Out of Distribution



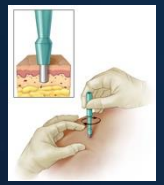
Reliable prediction



Unreliable prediction

Out-of-distribution (OOD) inputs are not flagged

Real World Example: Biopsy + Histological



Human in the Loop



Imaging

Pathologist

Reliable Cancer: Yes/No

Oncologist

- ↑ Unsustainable w/increasing data volumes.
- ↓ Human as a bottleneck vs. AI-enabled automation empowering human in the loop.



AI-Enabled Device



Imaging

Trained AI Model

Reliable Cancer: Yes/No

Oncologist

Out-of-distribution (OOD) inputs are not flagged

Objective of our Work



AI-Enabled Device



Imaging

Trained AI Model

Prediction

Uncertainty

Low Uncertainty

Oncologist

reliable

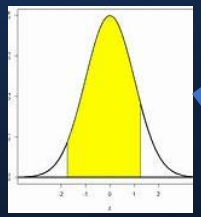
High Uncertainty

Pathologist

Uncertainty Quantification System

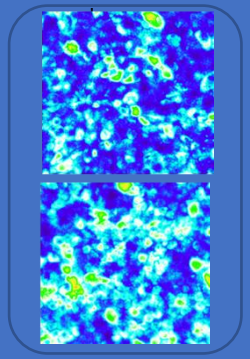
Training Data

Representation of In-Distribution Inside of the AI

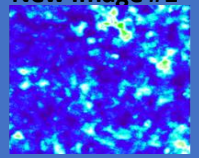


Key Measurement Ideas

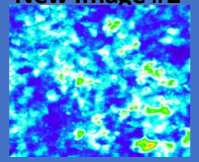
Training Data: Input Images



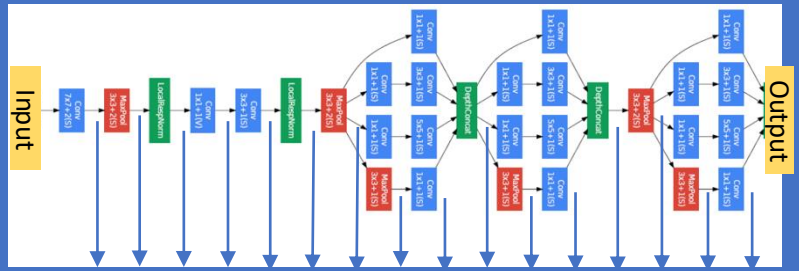
New Image #1



New Image #2



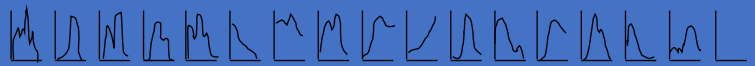
AI Architecture: GoogLeNet



Probe Nodes in AI Model

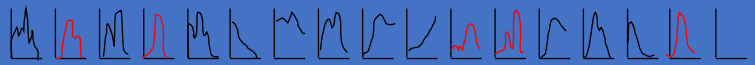


Well-aligned to the training data = reliable classification



Reliable Prediction

Not well-aligned to the training data = unreliable classification



Unreliable Prediction

Training Data: Labels

Label: High Potency

Label: Low Potency

SI Metrology and Governance

- Metrology for SI: Brief Introduction
- Characterizing Uncertainty in Deep Learning Models
- Determining Accuracy for Stem Cell Transplantation
- Evaluating Agentic SI (AI)
- Detecting Trojans
- Risk Management

Clinical Environment: Cell Therapy for Age–Related Macular Degeneration (AMD)

[Stem Cell Engineering of Retinal Pigment Epithelium]

Global cost is **$343 billion**



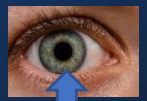
AMD blurred vision

11M in US

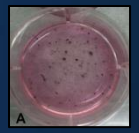


20/20 Vision









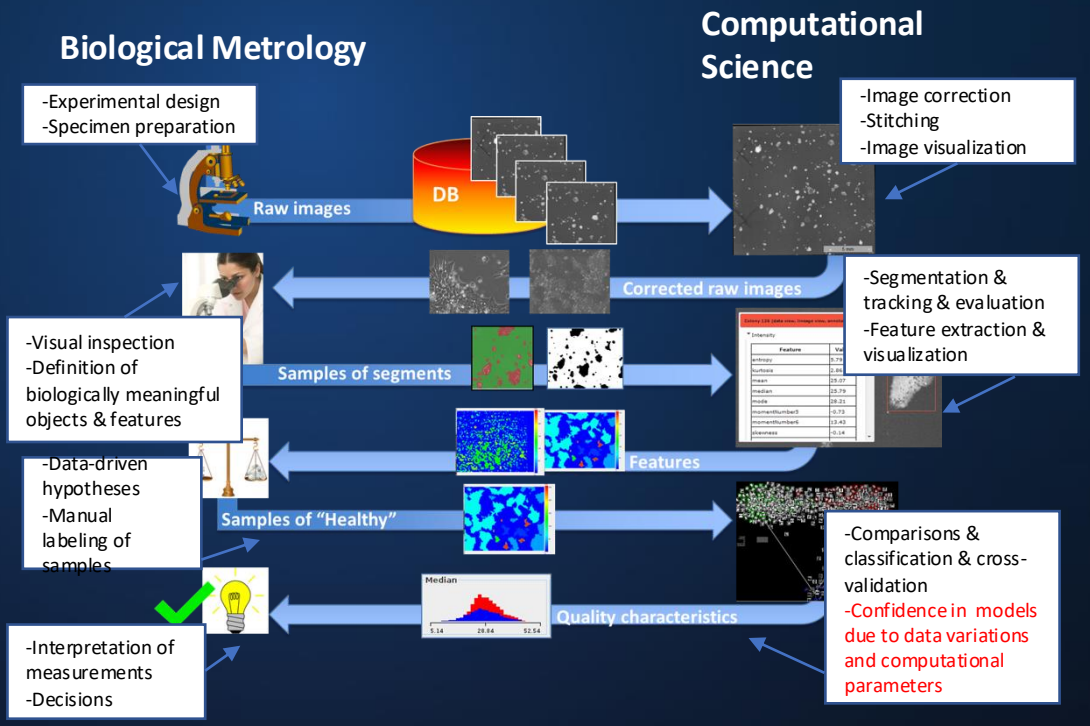
6 to 9 months



Scientific insights: Quality assurance criteria for cell implants

Courtesy: Peter Bajcsy

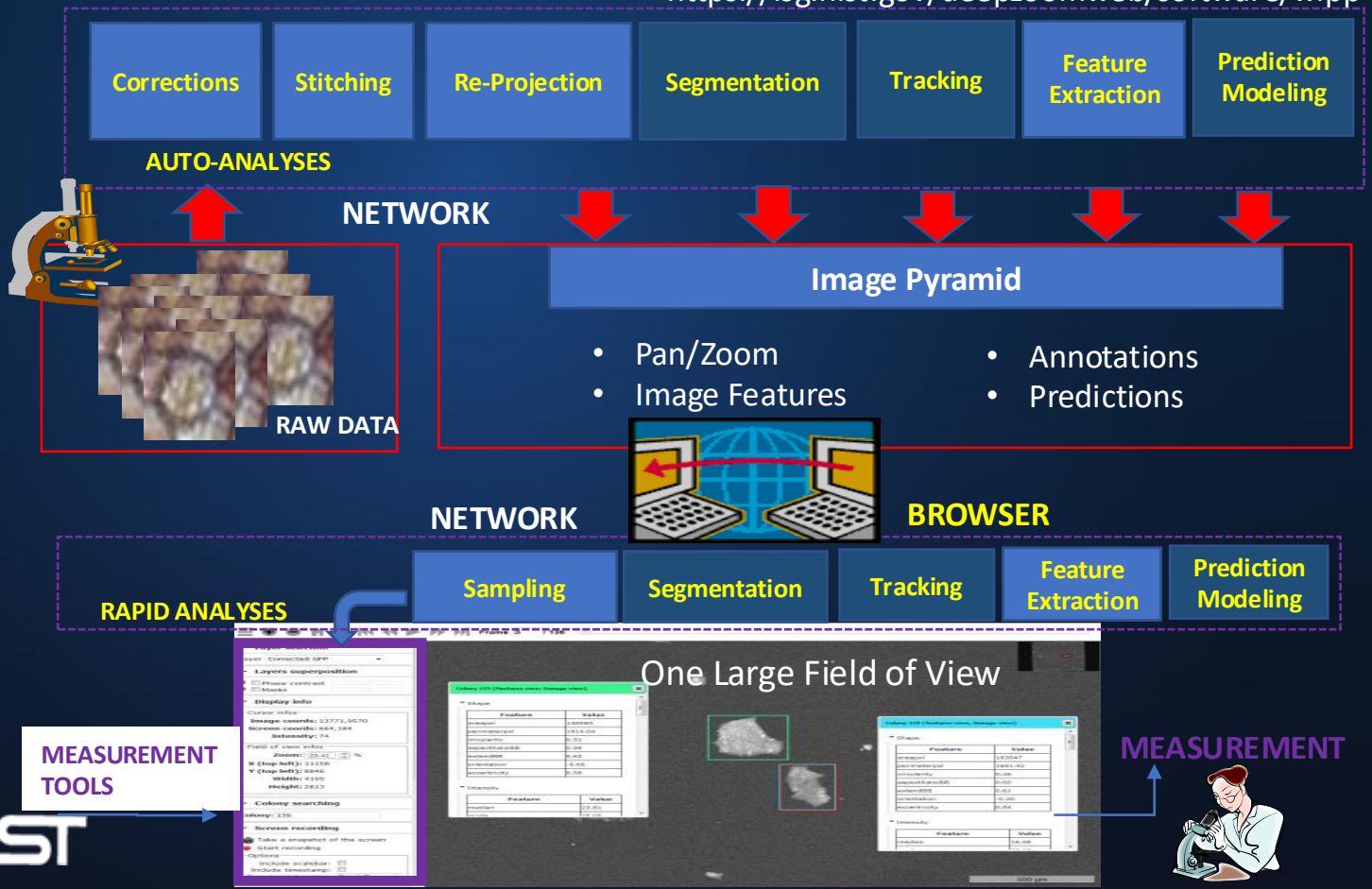
AI-based Quality Assurance of RPE Cell Implants for AMD Patients



Courtesy: Peter Bajcsy

IT/SI for Measurement Science

https://isg.nist.gov/deepzoomweb/software/wipp

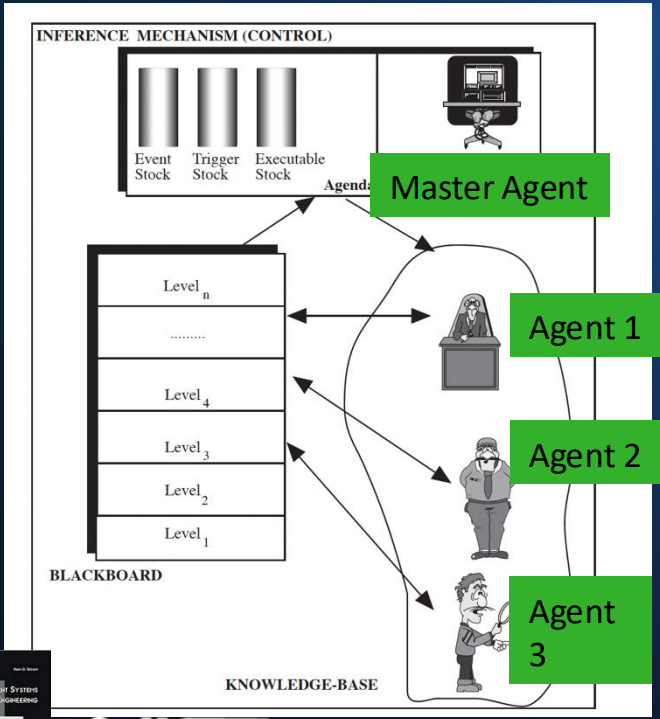


SI Metrology and Governance

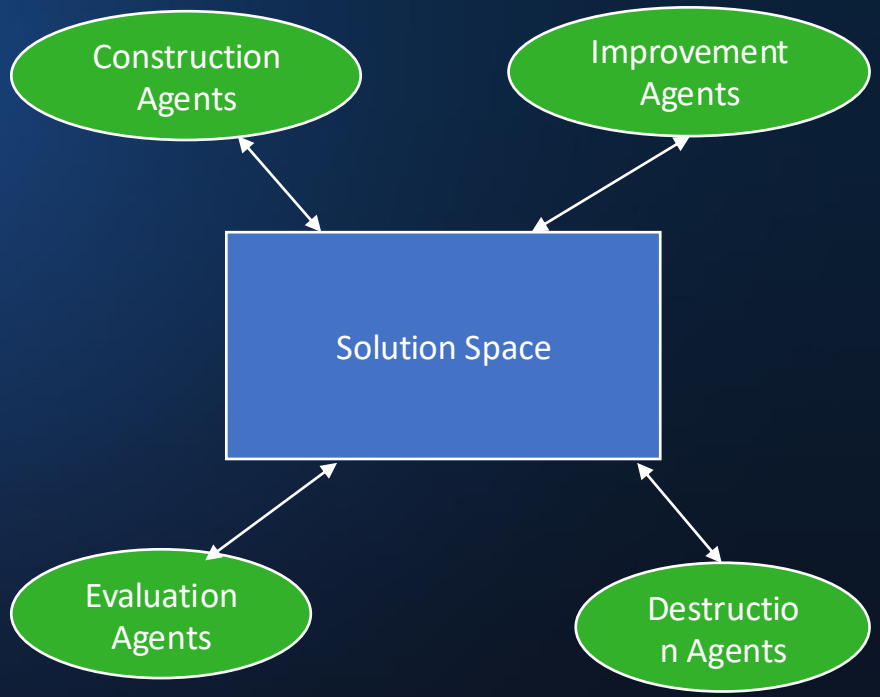
- Metrology for SI: Brief Introduction
- Characterizing Uncertainty in Deep Learning Models
- Determining Accuracy for Stem Cell Transplantation
- Evaluating Agentic SI (AI)
- Detecting Trojans
- Risk Management

Agent Frameworks of 1980s Revisited: First Wave

The Blackboard Framework



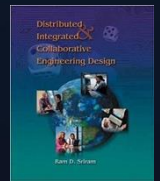
The A-Teams Approach





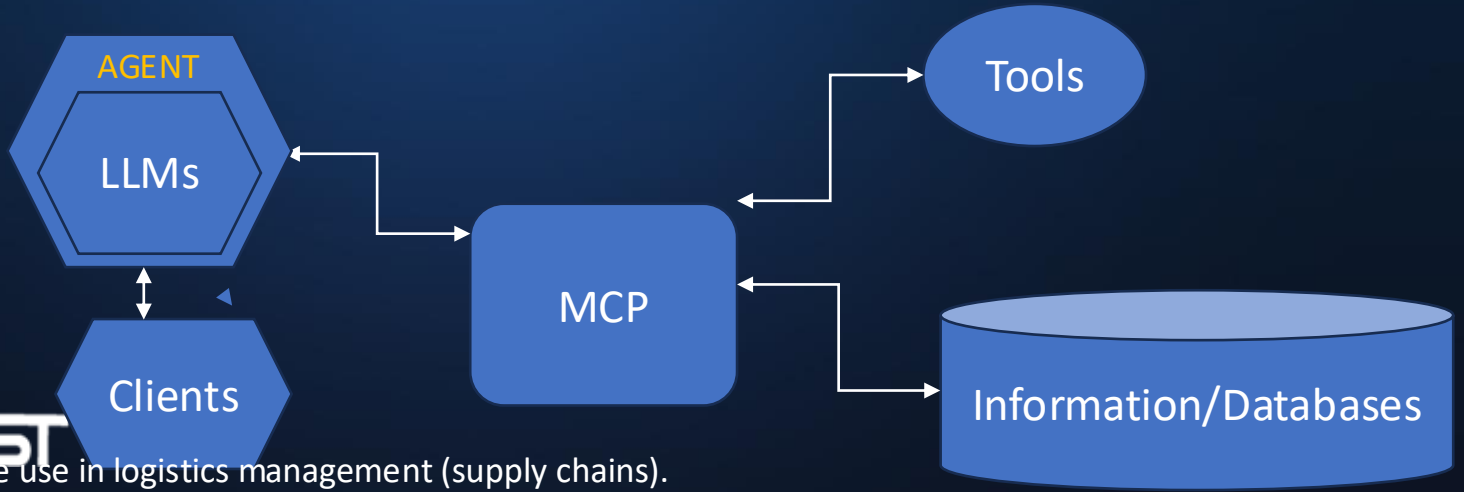
NIST

Sarosh Talukdar



Second Wave: Agentic SI with LLMs

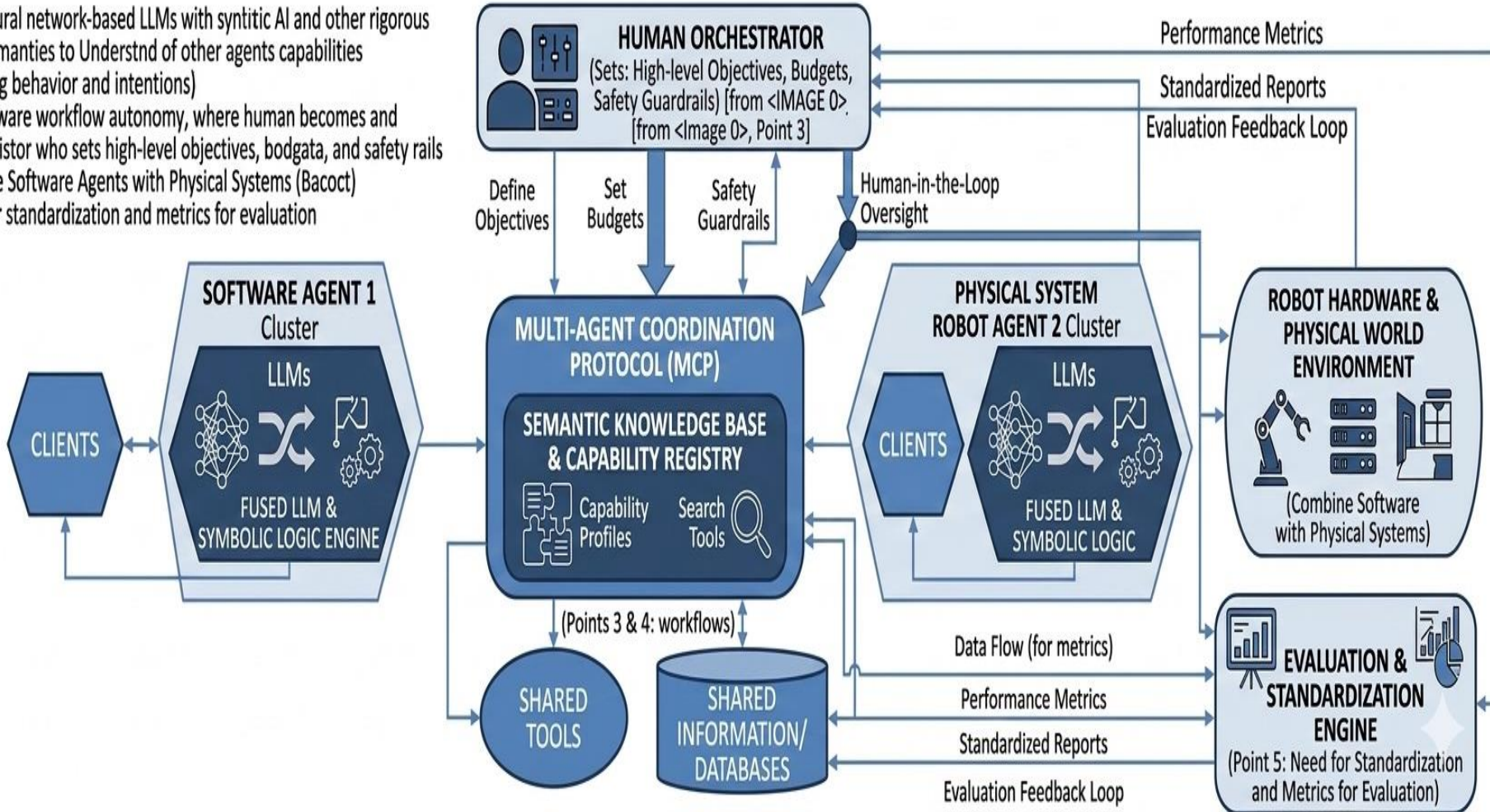
- Agents use large language models (LLMs) and other tools to understand context, analyze data, make decisions, learn from feedback, and take real-world actions. (Service Vs Model)
- For communication Agents use Model Context Protocol (MCP), which is an open source standard for connecting AI applications to external systems



Has found some use in logistics management (supply chains).

Agentic Architecture: Third Wave (A Possible Scenario)

- Fuse neural network-based LLMs with syntitic AI and other rigorous
- Have semanties to Understnd of other agents capabilities (Inducing behavior and intentions)
- Shift to ware workflow autonomy, where human becomes and orchestristor who sets high-level objectives, bodgata, and safety rails
- Combine Software Agents with Physical Systems (Bacoct)
- Need for standardization and metrics for evaluation

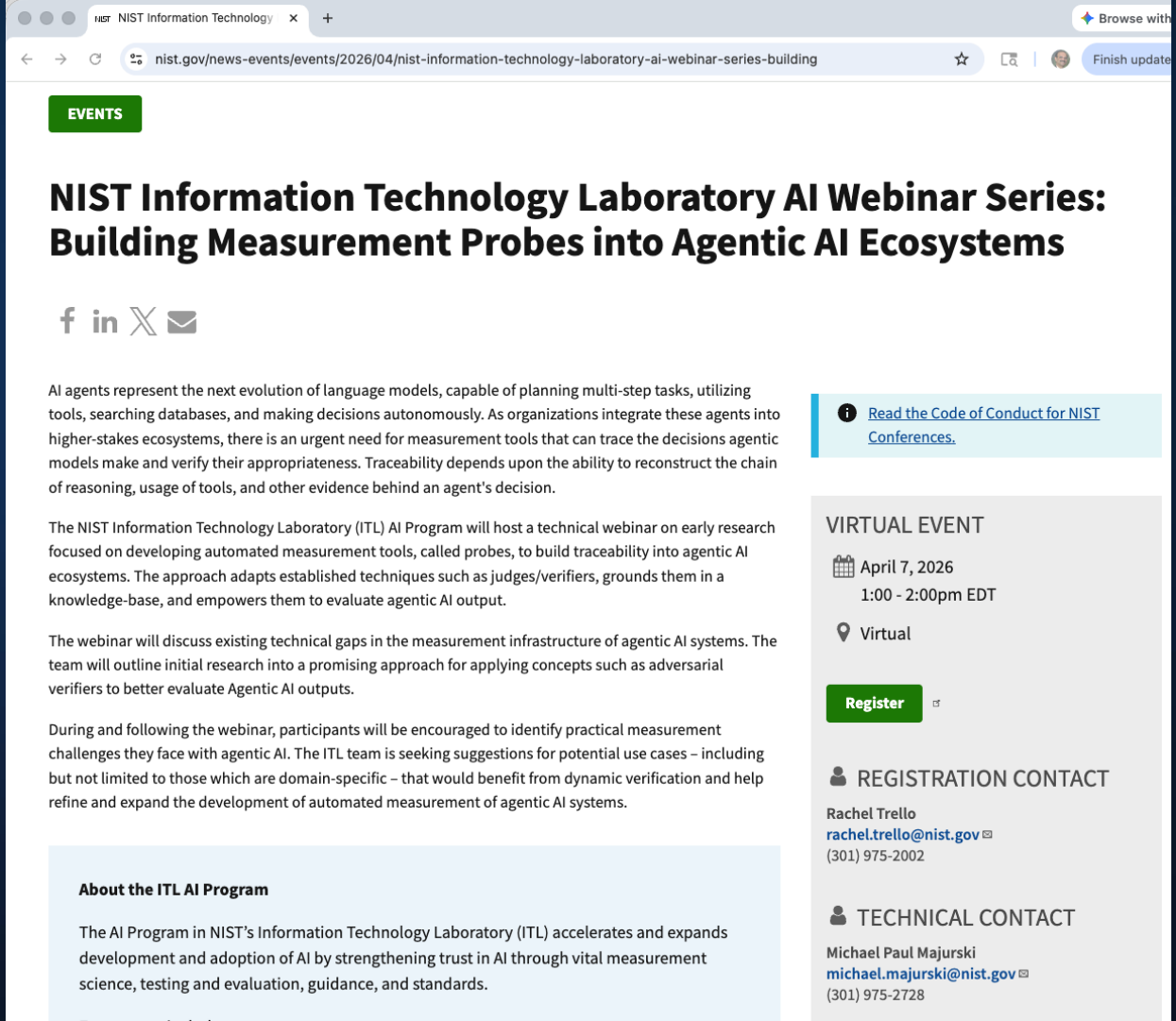


Gemini

Agentic SI (AI) Evaluation

A platform for measuring and experimenting with LLM-based AI Systems focusing on Agentic AI (Tim, Derek, Ben, Alden, Michael, Guillaume, Philippe)

- Testbed for LLM dynamic context construction methodology evaluation



EVENTS

NIST Information Technology Laboratory AI Webinar Series: Building Measurement Probes into Agentic AI Ecosystems

AI agents represent the next evolution of language models, capable of planning multi-step tasks, utilizing tools, searching databases, and making decisions autonomously. As organizations integrate these agents into higher-stakes ecosystems, there is an urgent need for measurement tools that can trace the decisions agentic models make and verify their appropriateness. Traceability depends upon the ability to reconstruct the chain of reasoning, usage of tools, and other evidence behind an agent's decision.

The NIST Information Technology Laboratory (ITL) AI Program will host a technical webinar on early research focused on developing automated measurement tools, called probes, to build traceability into agentic AI ecosystems. The approach adapts established techniques such as judges/verifiers, grounds them in a knowledge-base, and empowers them to evaluate agentic AI output.

The webinar will discuss existing technical gaps in the measurement infrastructure of agentic AI systems. The team will outline initial research into a promising approach for applying concepts such as adversarial verifiers to better evaluate Agentic AI outputs.

During and following the webinar, participants will be encouraged to identify practical measurement challenges they face with agentic AI. The ITL team is seeking suggestions for potential use cases – including but not limited to those which are domain-specific – that would benefit from dynamic verification and help refine and expand the development of automated measurement of agentic AI systems.

About the ITL AI Program

The AI Program in NIST's Information Technology Laboratory (ITL) accelerates and expands development and adoption of AI by strengthening trust in AI through vital measurement science, testing and evaluation, guidance, and standards.

Read the Code of Conduct for NIST Conferences.

VIRTUAL EVENT

April 7, 2026
1:00 - 2:00pm EDT

Virtual

Register

REGISTRATION CONTACT

Rachel Trello
rachel.trello@nist.gov
(301) 975-2002

TECHNICAL CONTACT

Michael Paul Majurski
michael.majurski@nist.gov
(301) 975-2728

Emerging Trends in AI Agent Evaluation

Felix Chen[1,*] Stephan Rabanser[1,*] Peter Kirgis[1] Nitya Nadgir[1]
Rishi Bommasani[2] Magda Dubois[3] Elham Tabassi[4] Cozmin Ududec[3]
Anka Reuel[2] Sanmi Koyejo[2] Sayash Kapoor[1,5,*] Arvind Narayanan[1,6,*]

[1]Princeton University [2]Stanford University [3]UK AI Security Institute
[4]The Brookings Institution [5]UC Berkeley [6]Institute for Advanced Study

[*]Corresponding authors: {felix_chen, rabanser, arvindn}@princeton.edu, sayashk@berkeley.edu

October 19, 2026

Abstract

AI models are increasingly deployed as agents: rather than responding to a single prompt, they plan over many steps, call external tools, and act within software and real-world environments. This shift challenges the existing outcome-focused benchmarking paradigm for evaluating AI models. Evaluating these agent systems must now encompass an array of scaffolds, tools, and environments that present a far more complex challenge for evaluators today. First, this shift has necessitated many improvements to measurement methodology, scientific rigor, and evaluation infrastructure. Second, the deployment of agents in real settings has expanded the scope of evaluation from the lab to the field. Third, the accompanying changes within the broader evaluator ecosystem hold new implications for those interested in contributing to this work. We survey 33 shifts, practices, and questions driven by the AI agent paradigm across the above three dimensions—the "How," "What," and "Who" of AI agent evaluation. Our goal is to help consolidate the emerging discipline of agent evaluation, equipped with an evolving methodological toolkit (Part 1), an orienting research agenda (Part 2), and an updated portrait of the evaluation landscape (Part 3).

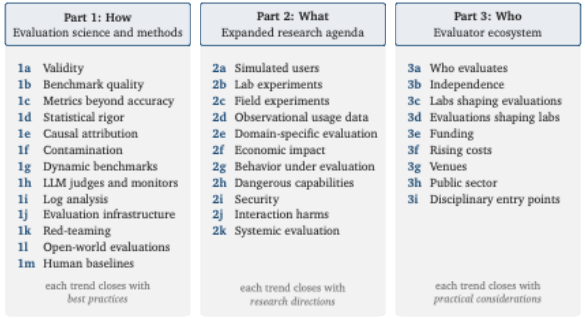


Figure 1: Overview of the 33 trends surveyed in this paper.

SI Metrology and Governance

- Metrology for SI: Brief Introduction
- Characterizing Uncertainty in Deep Learning Models
- Determining Accuracy for Stem Cell Transplantation
- Evaluating Agentic AI
- Detecting Trojans
- Risk Management

TrojAI

Problem:

- Adversarial AI attacks insert trojan triggers that alter the behavior of the system.
- Modern AI advances are characterized by vast, crowdsourced datasets that are impractical to clean or monitor.
- Developers may ingest trojan triggers through the process of transfer learning.



Goal: Identify trojan triggers prior to widespread deployment

Courtesy: Michael Majurski , Derek Juba, Tim Blattner, et al.

[2602.07152] Trojans in Artifi
Browse with AI
arxiv.org/abs/2602.07152
Finish update

Cornell University

We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate

arXiv > cs > arXiv:2602.07152

Search... All fields Search
Help | Advanced Search

Computer Science > Cryptography and Security

[Submitted on 6 Feb 2026 (v1), last revised 27 Feb 2026 (this version, v2)]

Trojans in Artificial Intelligence (TrojAI) Final Report

Kristopher W. Reese, Taylor Kulp-McDowall, Michael Majurski, Tim Blattner, Derek Juba, Peter Bajcsy, Antonio Cardone, Philippe Dessauw, Alden Dima, Anthony J. Kearsley, Melinda Kleczynski, Joel Vasanth, Walid Keyrouz, Chace Ashcraft, Neil Fendley, Ted Staley, Trevor Stout, Josh Carney, Greg Canal, Will Redman, Aurora Schmidt, Cameron Hickert, William Paul, Jared Markowitz, Nathan Drenkow, David Shriver, Marissa Connor, Keltin Grimes, Marco Christiani, Hayden Moore, Jordan Widjaja, Kasimir Gabert, Uma Balakrishnan, Satyanadh Gundimada, John Jacobellis, Sandya Lakkur, Vitus Leung, Jon Roose, Casey Battaglino, Farinaz Koushanfar, Greg Fields, Xihe Gu, Yaman Jandali, Xinqiao Zhang, Tara Javidi, Akash Vartak, Tim Oates, Ben Erichson, Michael Mahoney, Rauf Izmailov, Xiangyu Zhang, Guangyu Shen, Siyuan Cheng, Shiqing Ma, XiaoFeng Wang, Haixu Tang, Di Tang, Xiaoyi Chen, Zihao Wang, Rui Zhu, Susmit Jha, Xiao Lin, Manoj Acharya, Weichao Zhou, Feisi Fu, Panagiota Kiourti, Chenyu Wang, Zijian Guo, H M Sabbir Ahmad, Wenchao Li, Chao Chen

The Intelligence Advanced Research Projects Activity (IARPA) launched the TrojAI program to confront an emerging vulnerability in modern artificial intelligence: the threat of AI Trojans. These AI trojans are malicious, hidden backdoors intentionally embedded within an AI model that can cause a system to fail in unexpected ways, or allow a malicious actor to hijack the AI model at will. This multi-year initiative helped to map out the complex nature of the threat, pioneered foundational detection methods, and identified unsolved challenges that require ongoing attention by the burgeoning AI security field. This report synthesizes the program's key findings, including methodologies for detection through weight analysis and trigger inversion, as well as approaches for mitigating Trojan risks in deployed models. Comprehensive test and evaluation results highlight detector performance, sensitivity, and the prevalence of "natural" Trojans. The report concludes with lessons learned and recommendations for advancing AI security research.

Subjects: **Cryptography and Security (cs.CR)**; Artificial Intelligence (cs.AI); Machine Learning (cs.LG)
Cite as: arXiv:2602.07152 **[cs.CR]**
(or arXiv:2602.07152v2 **[cs.CR]** for this version)
https://doi.org/10.48550/arXiv.2602.07152

Submission history

From: Taylor Kulp-McDowall [view email]
[v1] Fri, 6 Feb 2026 19:52:14 UTC (14,527 KB)
[v2] Fri, 27 Feb 2026 21:52:03 UTC (14,525 KB)

Access Paper:
View PDF
view license

Current browse context:
cs.CR
< prev | next >
new | recent | 2026-02
Change to browse by:
cs
cs.AI
cs.LG

References & Citations
NASA ADS
Google Scholar
Semantic Scholar

Export BibTeX Citation

Bookmark

SI Metrology and Governance

- Metrology for SI: Brief Introduction
- Characterizing Uncertainty in Deep Learning Models
- Determining Accuracy for Stem Cell Transplantation
- Identifying Metrics for Healthcare Chatbots
- Evaluating Agentic AI
- Detecting Trojans
- Risk Management

NIST AI Risk Management Framework and Resources





NIST AI RMF: A voluntary resource for organizations designing, developing, deploying, or using AI systems to manage AI risks and promote effective and reliable AI that promotes public trust.

This will be updated soon as per President's Executive Order

Outline

- NIST and SI (AI)
- SI (AI) Evolution
- SI Metrology and Governance
- Future Work
- Summary

Next Steps and Strategic Recommendations (Soft SI)

Governance Horizon	Primary Objective	Key Action Items
LSI Deployments	Operationalize Trust Metrics	Standardize Evaluation Cards; establish concrete benchmarks for Uncertainty Quantification (UQ) and human correction/override rates.
GSI Systems	Prevent Semantics and Logic Drift	Formalize top-level ontologies; establish neuro-symbolic testing protocols to validate symbolic goal constraints on sub-symbolic models; develop metrics for Garland Test.
TSI / Agentic Networks	Autonomous Multi-Agent Oversight	Develop guidelines for certifying automated multi-agent auditing networks ("A-Teams") and define explicit Human-on-the-Loop escalation latencies and coverage ratios.

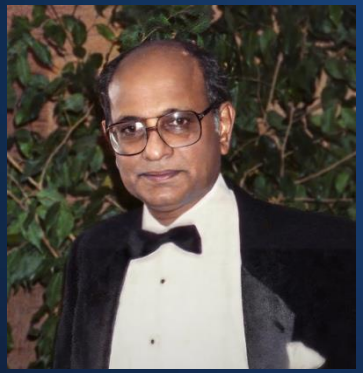
Summary

- Important to develop benchmarks, metrics and standards for evaluating SI (AI) systems
- Metrics and standards will need to be expanded with new concepts such as agentic technologies
- Standards introduced at the right time will lead to innovation
- Testing algorithms will increase trust in SI (AI) [also may help in curtailing unforeseen behavior]
- Benchmark datasets are need
- Learning programs need to explain reasoning
- Public-Private partnerships may help accelerate progress

NIST

Dedication

- This talk is dedicated to my mentors



Prof D. V. Reddy
(August 04, 1932 – December 20, 2025)



Prof Steven J. Fenves
(June 06, 1931 – December 23, 2025)

Acknowledgments and Disclaimer

- This talk is based on the work by several people at NIST. Credit is provided in appropriate slides. My thanks to all those who provided the slides.

- Commercial equipment and software, many of which are either registered or trademarked, are identified in order to adequately specify certain procedures. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.