UNIVERSITY OF CALIFORNIA

Systemwide Academic Congress

What the Future Holds: A UC Congress on the Impact and Promise of Artificial Intelligence

Wed, Feb 28 - Thurs, Feb 29, 2024

UCLA

REGISTER NOW

UNIVERSITY OF CALIFORNIA

bitly

Trustworthy AI:
Whose Trust Needs to be Earned and How

Ida Sim, MD, PhD

- Professor of Medicine and Computational Precision Health
- Chief Research Informatics Officer
- Co-Director, UCSF UC Berkeley Joint Program in Computational Precision Health

February 7, 2024

Disclosures

- **Vivli**, Co-Founder, Board of Directors, & Consultant
- **Open mHealth**, Co-Founder
- **The Commons Project Foundation**, General Assembly Member
- **98point6**, past Medical Advisory Board, shareholder
- **Myia**, shareholder

Views expressed are my own and not the views of UCSF

Take-Home Points

Trust in AI is **earned** from a person or community; trust is earned by the AI being **worthy of trust** by that person or community

Trustworthiness is best achieved by **continuing demonstration of robustness and reliability**

- *Algorithmic* transparency, interpretability, and explainability and are not sufficient to earn trust from patients, clinicians, and the public

AI Vigilance methods, organization, funding, and sustainability are crucial for achieving Trustworthy AI

Outline

- **Definitions: AI, Machine Learning, Large Language Models**
- **Trust and Trustworthiness**
- **Robust/Reliable AI**
- **Conclusion**

Artificial Intelligence (AI)

Ability of a machine to perform tasks (and behave) like an intelligent being

Machine Learning (ML)

computer algorithms that find and apply patterns in (huge amounts of) data

https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/

Large Language Models (LLMs)

aka generative AI

- **Hallucinating:** generated sentences may or may not be true
- **Stochastic:** parrots next word with random probability, generating sentences
- **Parrot:** has heard a lot of words and can “parrot” them back based on word patterns

Large Language Models (LLMs)

aka generative AI

- **Hallucinating:** generated sentences may or may not be true
- **Stochastic:** parrots next word with random probability, generating sentences
- **Parrot:** has heard a lot of words and can "parrot" them back based on word patterns

Microsoft ⟷ OpenAI — GPT4

Google — Gemini

Meta — Llama 2

Trust and Trustworthiness

Definitions

From Oxford Languages Dictionary

Trust

“firm belief in the reliability, truth, ability, or strength of someone or something”

Trustworthiness

“the ability to be relied on as honest or truthful”

Definitions

From Oxford Languages Dictionary

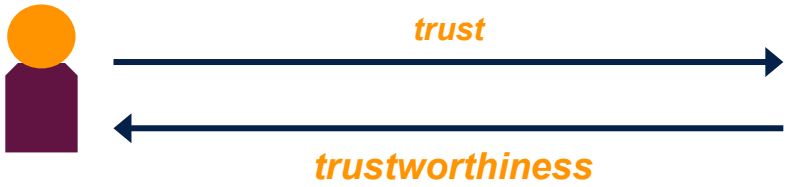
Trust

"firm belief in the reliability, truth, ability, or strength of someone or something"

Trustworthiness

"the ability to be relied on as honest or truthful"

- Patients
- Clinicians
- Public



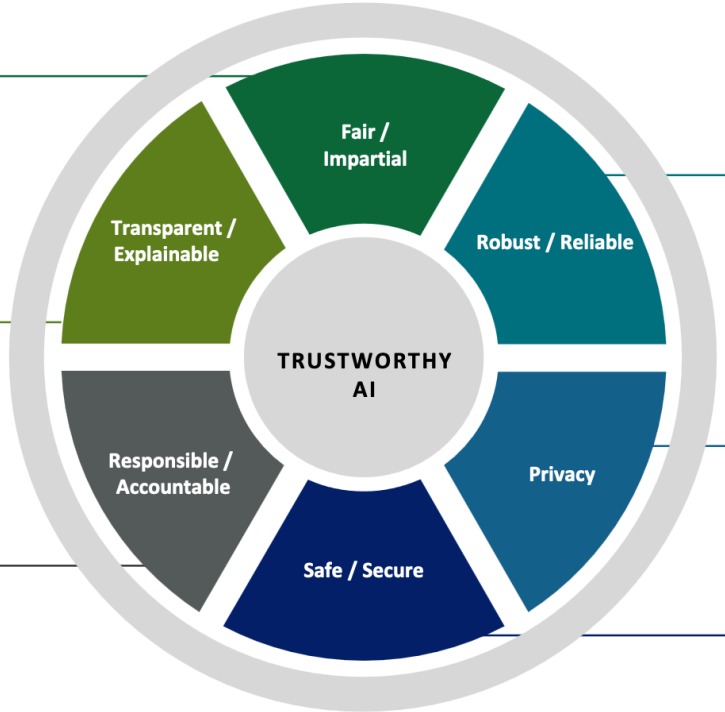
& non-LLM AI

HHS Principles of Trustworthy AI

Fair / Impartial
AI applications should include checks from internal and external stakeholders to help ensure equitable application across all participants

Transparent / Explainable
All relevant individuals should understand how their data is being used and how AI systems make decisions; algorithms, attributes, and correlations should be open to inspection

Responsible / Accountable
Policies should outline governance and who is held responsible for all aspects of the AI solution (e.g., initiation, development, outputs, decommissioning)



Robust / Reliable
AI systems should have the ability to learn from humans and other systems and produce accurate and reliable outputs consistent with the original design

Privacy
Individual, group, or entity privacy should be respected, and their data should not be used beyond its intended and stated use; data used has been approved by the data owner or steward

Safe / Secure
AI systems should be protected from risks (including Cyber) that may directly or indirectly cause physical and/or digital harm to any individual, group, or entity

TAI principles are not mutually exclusive, and tradeoffs often exist when applying them.

15

https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf

Algorithmic Transparency: Useful, Not Sufficient

- "All relevant individuals should understand how…their data is being used"
  - To earn trust: policy transparency and communications for how patient data and clinician data (e.g., EHR practice patterns) are used for any purpose, not just AI
- "All relevant individuals should understand how… AI systems make decisions"
  - To earn trust: decision-making transparency and communications around system decisions, not just AI systems
    - *Systems* (people, organizations) decide on allocation of resources in the real world; algorithms (AI and otherwise) are tools that support and implement decision-making by systems
- Algorithmic transparency is needed for robust/reliable AI but *does not by itself* lead to trustworthiness

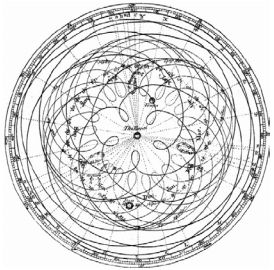
Algorithmic Inspectability: Useful, Not Sufficient

- **“algorithms, attributes, and correlations should be open to inspection”**
- Some LLMs are proprietary (e.g., GPT4), some are open source (e.g., Llama 2)
    - Who needs to dissect the parrot?
    - Why? What are the consequences of inspection?
    - How does inspection by itself give AI “the ability to be relied on as honest or truthful”?

- We don’t conduct inspection of statistical models (e.g., logistic regression) that are deployed in our health systems
- Algorithmic inspectability needed for robust/reliable AI but does not by itself lead to trustworthiness

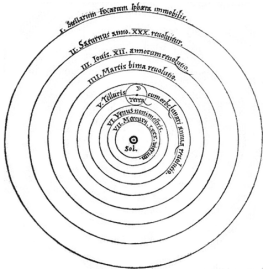
Algorithmic Explainability: Not Sufficient, May be Misleading

- Explainability not a reasonable expectation of generative AI
  - LLMs and Machine Learning models are stochastic “black boxes”
  - Rational logical explainability are available from different AI technologies (e.g., belief networks, knowledge graphs)
- Models can be explanatory but wrong



Ptolomaic model:
Sun revolves around the Earth

Both models are highly predictive and explanatory about the observed motion of celestial bodies



Copernican model:
Earth revolves around the Sun

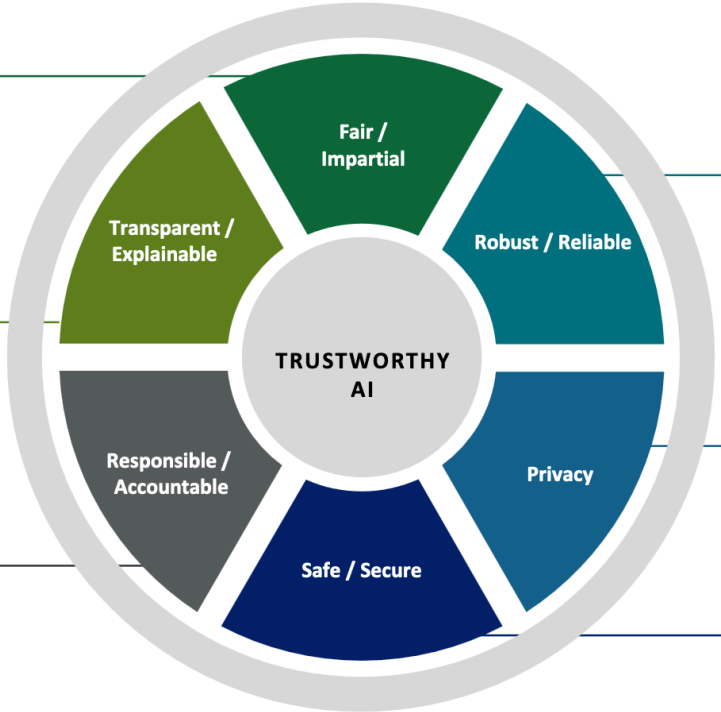
- Algorithmic explainability does not by itself lead to (justified) trustworthiness

HHS Principles of Trustworthy AI

Fair / Impartial
AI applications should include checks from internal and external stakeholders to help ensure equitable application across all participants

Transparent / Explainable
All relevant individuals should understand how their data is being used and how AI systems make decisions; algorithms, attributes, and correlations should be open to inspection

Responsible / Accountable
Policies should outline governance and who is held responsible for all aspects of the AI solution (e.g., initiation, development, outputs, decommissioning)



Robust / Reliable
AI systems should have the ability to learn from humans and other systems and produce accurate and reliable outputs consistent with the original design

Privacy
Individual, group, or entity privacy should be respected, and their data should not be used beyond its intended and stated use; data used has been approved by the data owner or steward

Safe / Secure
AI systems should be protected from risks (including Cyber) that may directly or indirectly cause physical and/or digital harm to any individual, group, or entity

TAI principles are not mutually exclusive, and tradeoffs often exist when applying them.

https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf

Robust/Reliable AI

Software is fundamentally different from Pharmaceuticals

Fixed molecular entity that doesn't change after FDA approval

Microsoft ↔ OpenAI — GPT4

Google — Gemini

Meta — Llama 2

Software performance continually changes

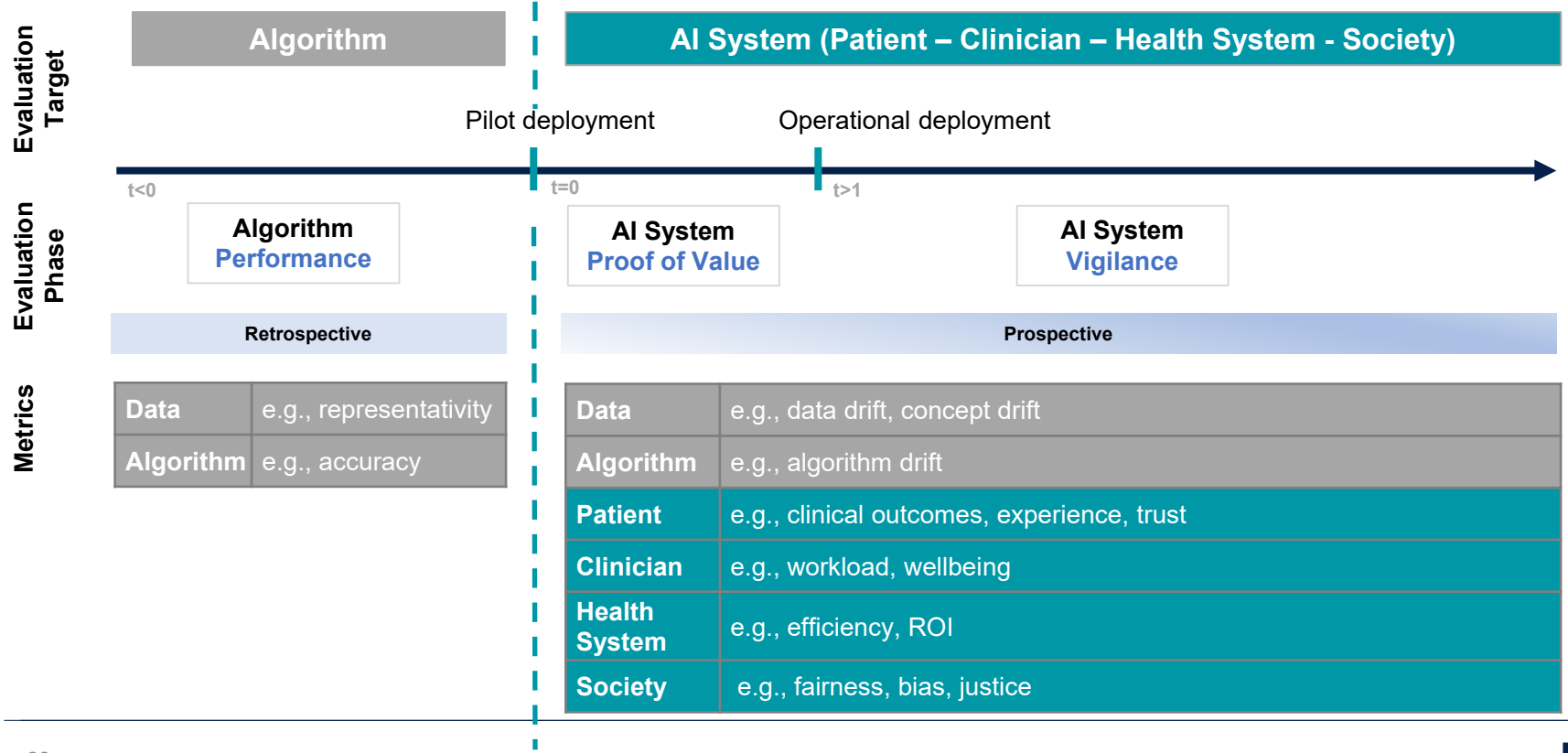
- Software is continually upgraded
- **Concept drift:** definitions change, e.g., long COVID
- **Data drift:** change in frequency, distribution, relationship of variables
- **Algorithm drift:** use case no longer aligned, e.g., change in payment incentives

Post-market pharmacovigilance looks out for adverse events

Post-deployment AI vigilance looks out for overall performance drift

UCSF

Continuing Demonstration of Robustness and Reliability



Take-Home Points

Trust in AI is **earned** from a person or community; trust is earned by the AI being **worthy of trust** by that person or community

Trustworthiness is best achieved by **continuing demonstration of robustness and reliability**

- *Algorithmic* transparency, interpretability, and explainability are not sufficient to earn trust from patients, clinicians, and the public

AI Vigilance methods, organization, funding, and sustainability are crucial for achieving Trustworthy AI

Acknowledgements

- **Sara Murray, MD** Chief Health AI Officer
- **Ki Lai** Chief Data Officer
- **Mandy Terrill** Associate Chief Informatics Officer-Research
- **Romain Pirracchio, MD** AI Vigilance

…and their teams, and many others

University of California
San Francisco

ida.sim at ucsf.edu

Towards a Model of AI Governance at University of California Health

UC Center Sacramento Speaker Series

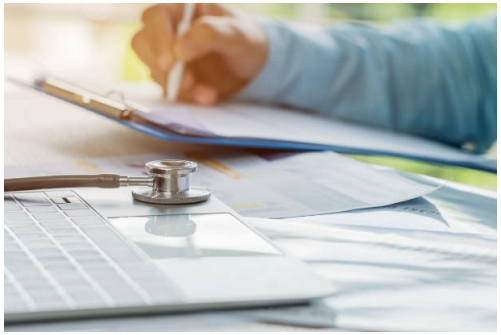
Cora Han, University of California Health Chief Health Data Officer

February 7, 2024

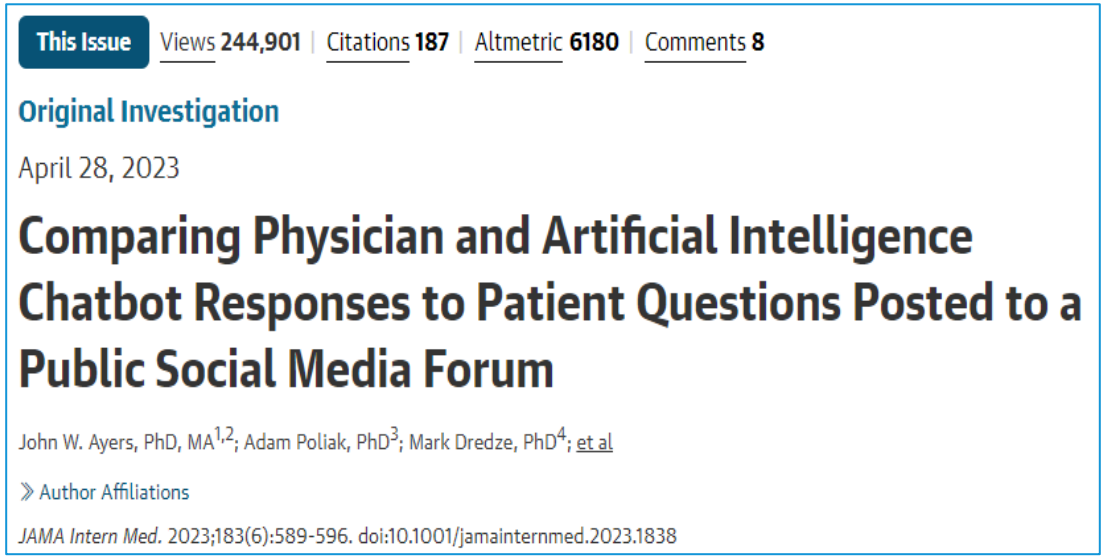
UNIVERSITY OF CALIFORNIA
HEALTH

AI in Healthcare is Not New

- Improving hospital administration
- Clinical decision support
- Population health management
- Payment management



Generative AI – Accelerated Interest and Excitement



This Issue | Views 244,901 | Citations 187 | Altmetric 6180 | Comments 8

Original Investigation

April 28, 2023

Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum

John W. Ayers, PhD, MA[1,2]; Adam Poliak, PhD[3]; Mark Dredze, PhD[4]; et al

» Author Affiliations

JAMA Intern Med. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838

Researchers compared written responses from physicians and ChatGPT to real-world health questions and found that a panel of licensed healthcare professionals preferred ChatGPT's responses 79% of the time, rating ChatGPT's responses as higher quality and more empathetic.

AI Will be Transformative But Presents Risks

Validity and Reliability	Safety	Accountability and Transparency	Security and Resiliency
Explainability and Interpretability	Privacy	Fairness	Workforce and Labor Impacts

Guardrails are Essential

Value of AI Governance

•Builds trust with end users and those impacted – patients, providers, administrators, community

•Enables vetting and authorization of AI tools more quickly and in a transparent, replicable manner

•Reduces the risk of unexpected harm and reputational damage

•Ensures compliance with existing end evolving laws and regulations

•Promotes safe and ethical innovation ecosystem

AI Governance – A Highly Active Space

Health Care Artificial Intelligence Code of Conduct

CHAI
Coalition for Health AI

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Office of Governor
GAVIN NEWSOM

"Governor Newsom Signs Executive Order to Prepare California for the Progress of Artificial Intelligence"

UNIVERSITY OF CALIFORNIA
HEALTH

President Biden's Executive Order on AI – Directives for HHS

- Establish an HHS AI Task Force Charged with Developing a Strategic Plan for Responsible Use of AI in the Health Sector
- Develop an AI Assurance Policy
- Ensure Compliance with Nondiscrimination Laws
- Create an AI Safety Program
- Prepare a Strategy for Regulating Use of AI in Drug Development
- Issue Grants and Awards

Governor Newsom's Executive Order on AI

"For decades, California has been a global leader in education, innovation, research, development, talent, entrepreneurship, and new technologies. As these technologies continue to grow and develop, California has established itself as the world leader in GenAI innovation with 35 of the world's top 50 AI companies and a quarter of all AI patents, conference papers, and companies globally."

-Press Release, Sept. 6, 2023

The Executive Order contains directives to state agencies and departments aimed at studying and deploying GenAI ethically and responsibly throughout state government.

AI Governance – NIST AI Risk Management Framework

- Characteristics of trustworthy AI systems
  - Valid and reliable
  - Safe
  - Secure and resilient
  - Accountable and transparent
  - Explainable and interpretable
  - Privacy-enhanced
  - Fair with harmful bias managed

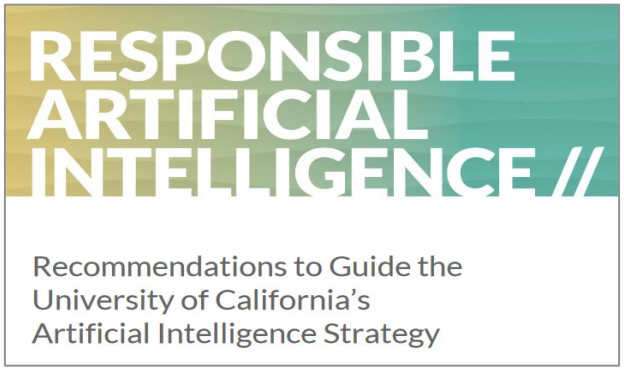
Deeper Dive into Healthcare AI Governance

Much work on principles and frameworks for responsible AI, but less on practical steps for operationalizing them.

CHAI (“Coalition for Health AI”)

- Build consensus around ways to measure trustworthy characteristics of AI systems
- Develop considerations and evaluation criteria for each stage of AI lifecycle
- Tailor to what makes healthcare different

UC Systemwide AI Council



- Health
- Policing
- Human Resources
- Student Experience

Health AI Governance Forum

Goals

- Share expertise and AI governance resources at each health location
- Surface concerns and difficult use cases
- Help cut through the AI governance noise
- Develop guidance
- Align UC AI efforts

A Few Takeaways for Developing AI Governance

- Important to address risk over the AI Lifecycle
- Requires multidisciplinary approach
- Generative AI amplifies existing risk and presents new risks
- Must keep apace with developing laws and regulations
- Developing AI governance will be an ongoing process

What's Ahead? Reading the Tea Leaves

UNIVERSITY OF CALIFORNIA
HEALTH

Thank you!

cora.han@ucop.edu

UNIVERSITY OF CALIFORNIA
HEALTH