APPLIED RESEARCH LABORATORY FOR INTELLIGENCE AND SECURITY

"SHOULD YOU RELY ON THAT AI?"

A New Look at Policy, Standards, and Requirements Specification

28 January 2021

Panelists

Lt Gen (ret) Ed Cardon

Former Director of the United States Army Office of Business Transformation and former Commander of the Second United States Army/United States Army Cyber Command; Professor of the Practice, Applied Research Laboratory for Intelligence and Security, University of Maryland

Dr. Chad Bieber

Director, Test and Evaluation, Project Maven, Johns Hopkins University Applied Physics Laboratory

Dr. Jane Pinelis

Chief, Test and Evaluation of AI/ML, Joint Artificial Intelligence Center, Former T&E lead for Project Maven

Prof. Michael Horowitz

Richard Perry Professor of Political Science, Director, Perry World House, University of Pennsylvania

Prof. Ben Shneiderman

Distinguished Professor of Computer Science and University of Maryland Institute for Advanced Computer Studies (UMIACS); Founding Director, Human Computer Interaction Lab; Affiliate, Institute for Systems Research and College of Information Studies, University of Maryland

Moderator: Dr. Craig Lawrence, Director, Systems Research, Applied Research Laboratory for Intelligence and Security; Visiting Research Scientist, Institute for Systems Research, Clark School of Engineering, University of Maryland

APPLIED RESEARCH LABORATORY FOR INTELLIGENCE AND SECURITY

UNIVERSITY OF MARYLAND

Panelists

Lt Gen (ret) Ed Cardon

Former Director of the United States Army Office of Business Transformation and former Commander of the Second United States Army/United States Army Cyber Command; Professor of the Practice, Applied Research Laboratory for Intelligence and Security, University of Maryland

Panelists

Dr. Chad Bieber

Director, Test and Evaluation, Project Maven, Johns Hopkins University Applied Physics Laboratory

Panelists

Dr. Jane Pinelis

Chief, Test and Evaluation of AI/ML, Joint Artificial Intelligence Center

Panelists

Prof. Michael Horowitz

Richard Perry Professor of Political Science, Director, Perry World House, University of Pennsylvania

Policy Challenges Surrounding AI Testing

Michael C. Horowitz

University of Pennsylvania

January 28, 2021

Bottom Line Up Front

- Advances in AI have national security applications across a range of arenas, from the back office to the battlefield – need a way to test and validate AI-enabled systems
- Questions exist both about ***how*** to effectively test AI systems and the standards for those tests compared to non-AI systems
- Critical challenge: navigating between the risk of a trust gap and the risk of automation bias in policymaker perspectives on AI

The Stakes

- Effective AI testing and evaluation standards for the national security realm is important for multiple reasons:
  - To generate trust necessary for AI adoption
  - To reduce the risk of AI backsliding
  - To decrease the potential for accidents with AI-enabled systems

Key Dilemma: Designing AI testing policymakers can understand

What is Getting Tested?

- Systems with continual learning

AND/OR

- Systems without continual learning

AI Testing Standards Compared to Other Systems

- Same standards as non-AI systems
- Lower standards than non-AI systems
- Higher standards than non-AI systems

Should testing standards depend on the area of application, specifics of the machine learning approach, or both?

Trust, Confidence, and AI (1)

Trust Gap

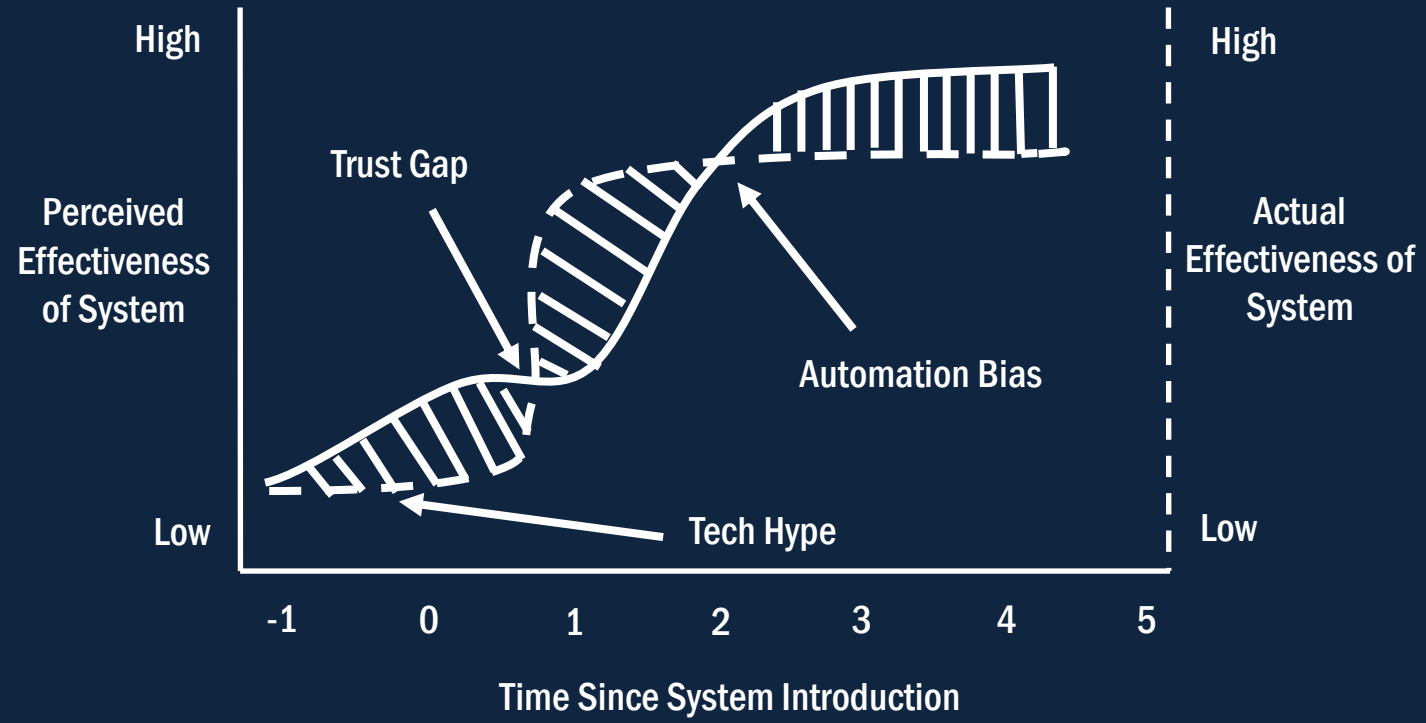
- Inability to trust machines to do work of people
- Unwillingness to deploy or properly use systems
- Example: Ground Tactical Air Controllers

V

Automation Bias

- Delegation of cognitive judgment to machine – trusting too much
- Failure to question algorithms if they make mistakes
- Example: Air France Crash
- Example: Patriot Missile fratricide

Trust, Confidence, and AI (2)



PERRY WORLD HOUSE
UNIVERSITY OF PENNSYLVANIA
Penn

Reducing the Risk of AI Backsliding

Backsliding refers to when accidents during adoption processes -> backlash against broader technology adoption

- AI is uniquely vulnerable to backsliding, as past AI winters show
- Reducing the risk:
  - Aligning expectations about AI with technological reality
  - Emphasizing the role of the *human*
  - Modernizing infrastructure

PERRY WORLD HOUSE
UNIVERSITY OF PENNSYLVANIA
Penn

Conclusion

- Effective testing and evaluation standards are critical to AI adoption in national security, and preventing AI backsliding
- Testing standards should depend on the type of AI application, and the degree of confidence in the AI method
- Need to navigate between the risk of trust gaps and automation bias through testing

Panelists

Prof. Ben Shneiderman

Distinguished Professor of Computer Science and University of Maryland Institute for Advanced Computer Studies (UMIACS); Founding Director, Human Computer Interaction Lab; Affiliate, Institute for Systems Research and College of Information Studies, University of Maryland

UMd ARLIS Workshop: "Should You Rely on that AI?"
January 28, 2021

Panel: A New Look at Policy, Standards, and Requirements Specification

Ben Shneiderman **@benbendc**

Founding Director (1983-2000), Human-Computer Interaction Lab
Professor, Department of Computer Science

Member, National Academy of Engineering

Photo: BK Adams

UNIVERSITY OF MARYLAND

PETER WALL INSTITUTE FOR ADVANCED STUDIES
THE UNIVERSITY OF BRITISH COLUMBIA VANCOUVER

Interdisciplinary research community
- Computer Science & Info Studies
- Psych, Socio, Educ, Jour & MITH

hcil.umd.edu
vimeo.com/72440805

Designing the User Interface

Design Theories

Direct manipulation

Menus, speech, search

Social Media

Information Visualization

www.cs.umd.edu/hcil/DTUI6

Sixth Edition: 2016

Web links

The **University of Maryland, College Park** (often referred to as the **University of Maryland**, **Maryland**, **UM**, **UMD**, **UMCP**, or **College Park**) is a public research university[10] located in the city of College Park in Prince George's County, Maryland, approximately 4 miles (6.4 km) from the northeast border of Washington, D.C. Founded in 1856, the university is the flagship institution of the University System of Maryland. With a fall 2010 enrollment of more than 37,000 students, over 100 undergraduate majors, and 120 graduate programs,

Tiny touchscreen keyboards

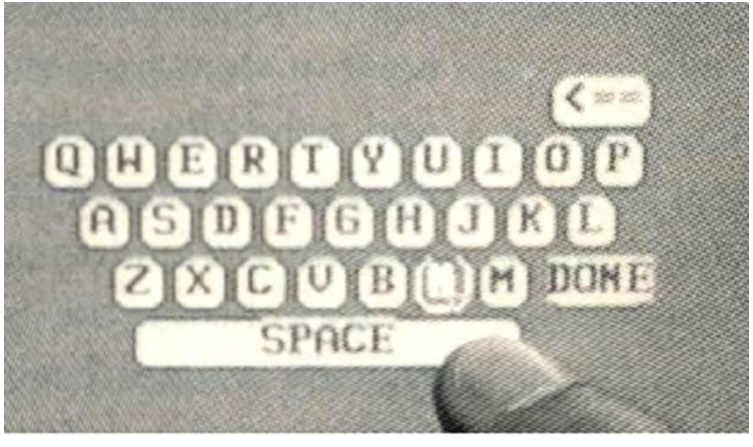
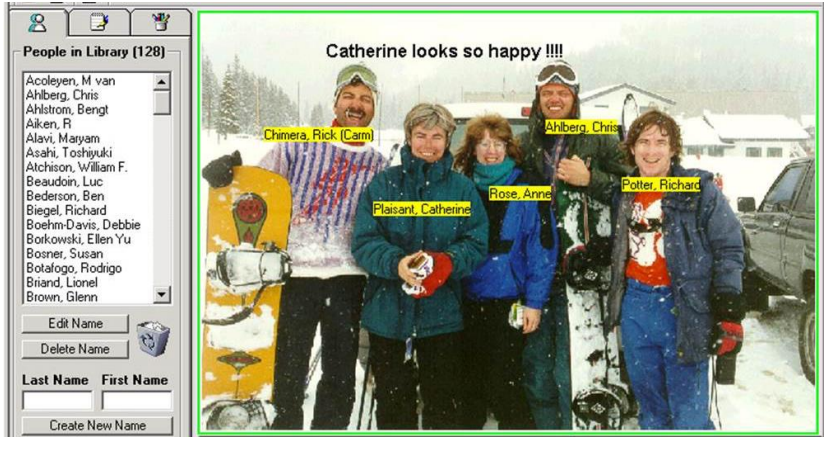
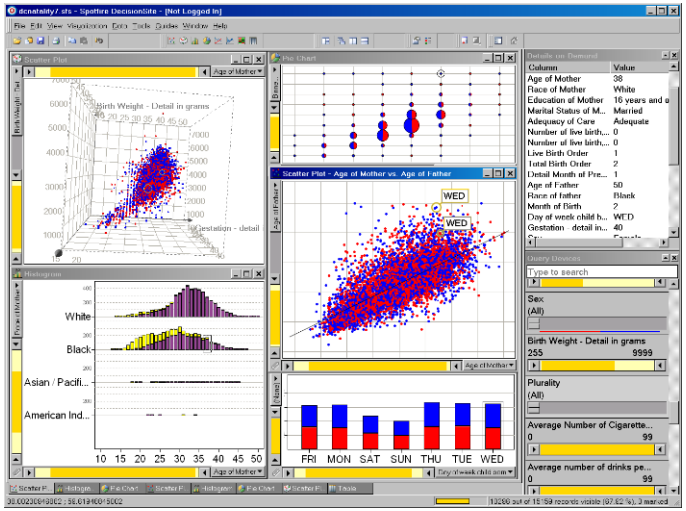


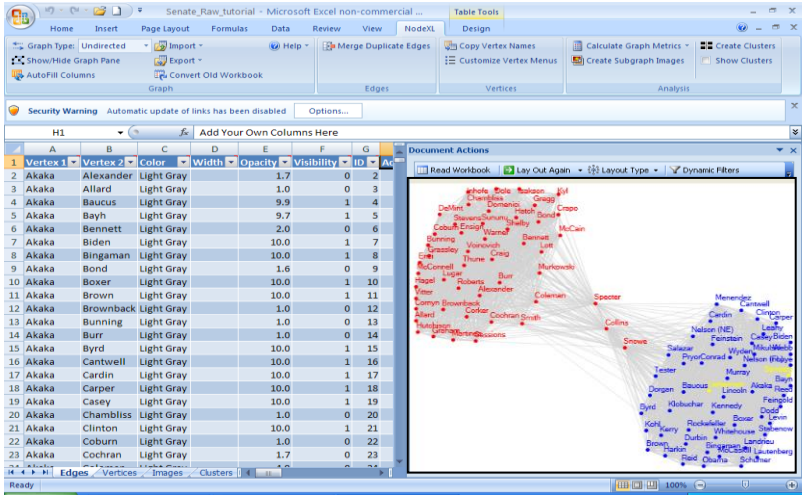
Photo tagging



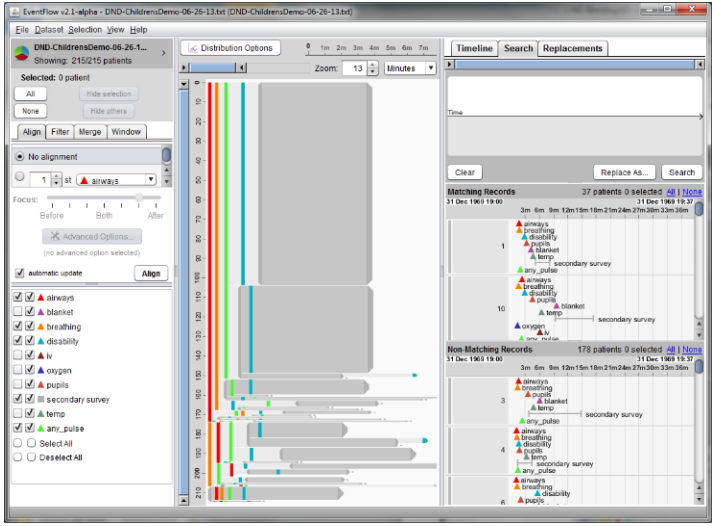


Spotfire

Treemaps
FinViz



NodeXL



EventFlow


What is Human-Centered AI?

Human-Centered AI

Amplify, Augment, Enhance & Empower People

Human-Centered AI

Amplify, Augment, Enhance & Empower People

→ 1000-fold improvements in capabilities

Information	Photography
Search	Navigation
Email & Text	Business Formation

Human-Centered AI

Amplify, Augment, Enhance & Empower People

→ 1000-fold improvements in capabilities

Information	Photography
Search	Navigation
Email & Text	Business Formation

→ RST: Reliable, Safe & Trustworthy

Human-Centered AI

Amplify, Augment, Enhance & Empower People

→ 1000-fold improvements in capabilities

Information	Photography
Search	Navigation
Email & Text	Business Formation

→ RST: Reliable, Safe & Trustworthy

→ Human self-efficacy, creativity & responsibility

Human-Centered AI

Amplify, Augment, Enhance & Empower People

→ 1000-fold improvements in capabilities

Information	Photography
Search	Navigation
Email & Text	Business Formation

→ RST: Reliable, Safe & Trustworthy

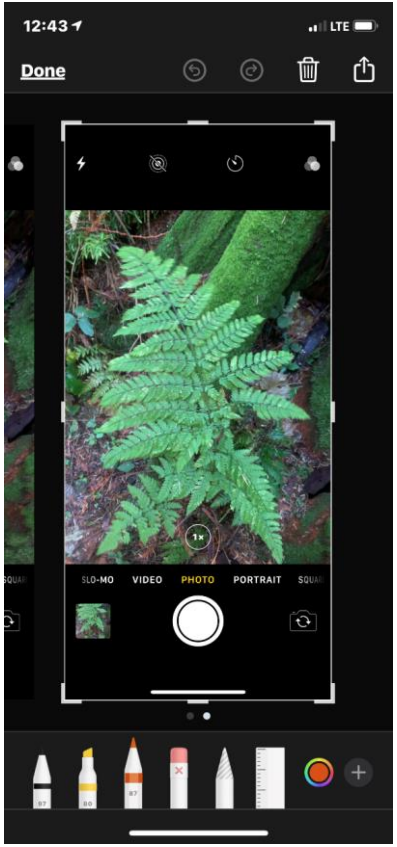
→ Human self-efficacy, creativity & responsibility

→ Human values, rights & dignity

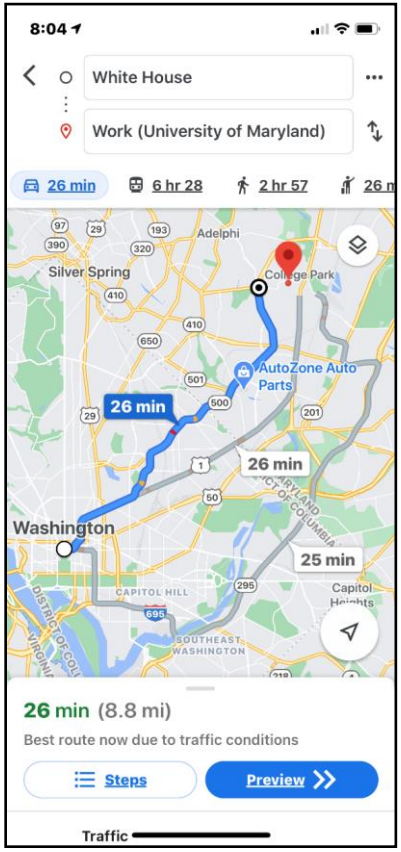
IJHCI (2020) **https://www.tandfonline.com/doi/full/10.1080/10447318.2020.1741118**

Supertools

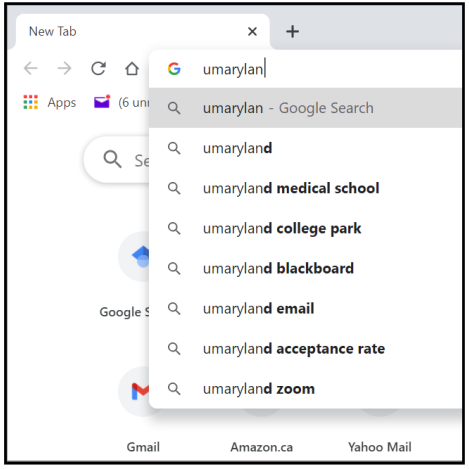
Digital Camera Controls



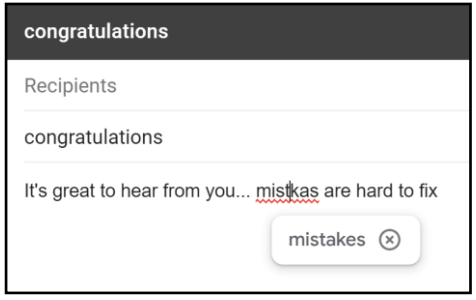
Navigation Choices



Texting Autocompletion



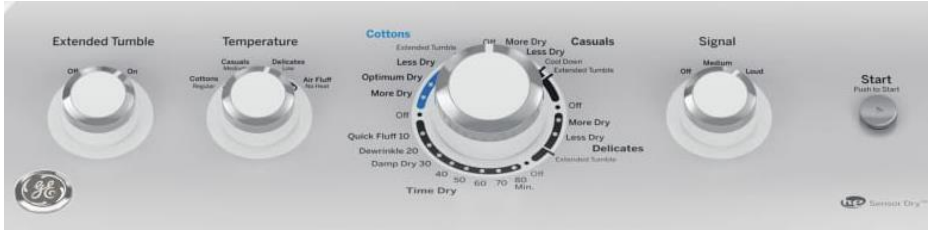
Spelling correction



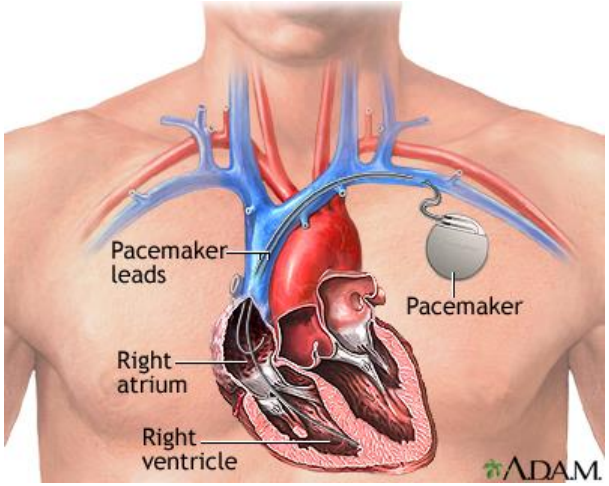
Active Appliances

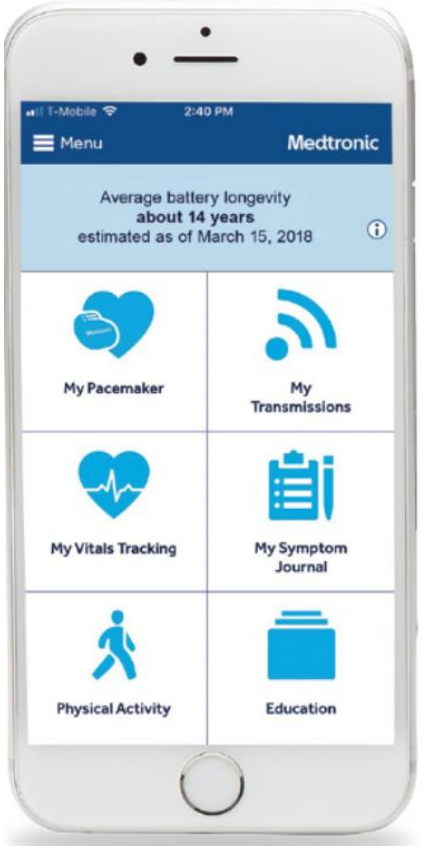
Coffee maker, Rice cooker, Blender

Dishwasher, Clothes Washer/Dryer



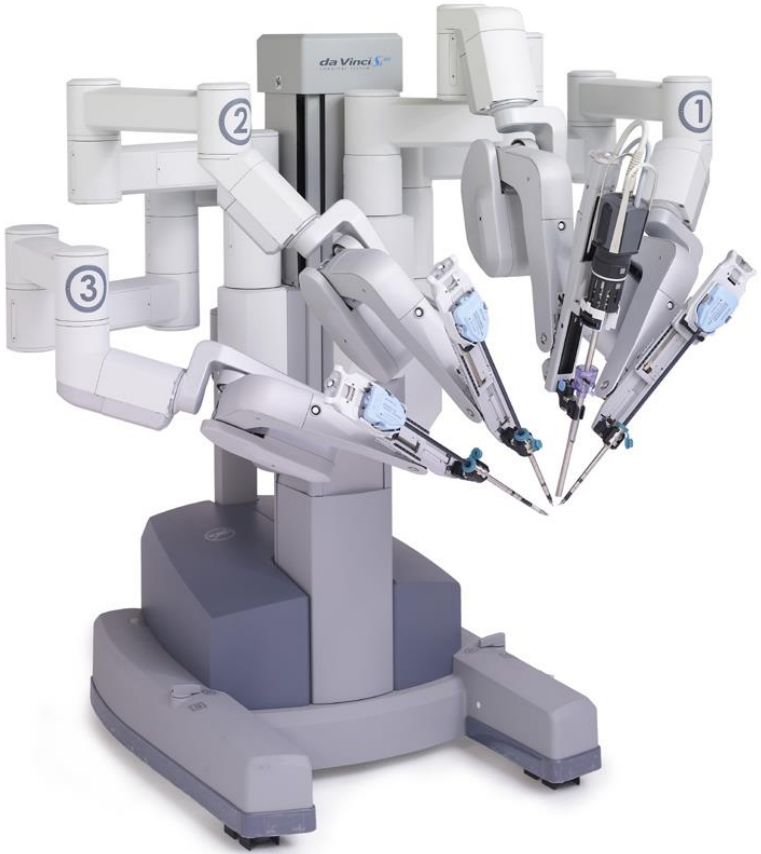
Implanted Cardiac Pacemakers





NASA Mars Rovers are Tele-Operated

DaVinci Tele-Operated Surgery

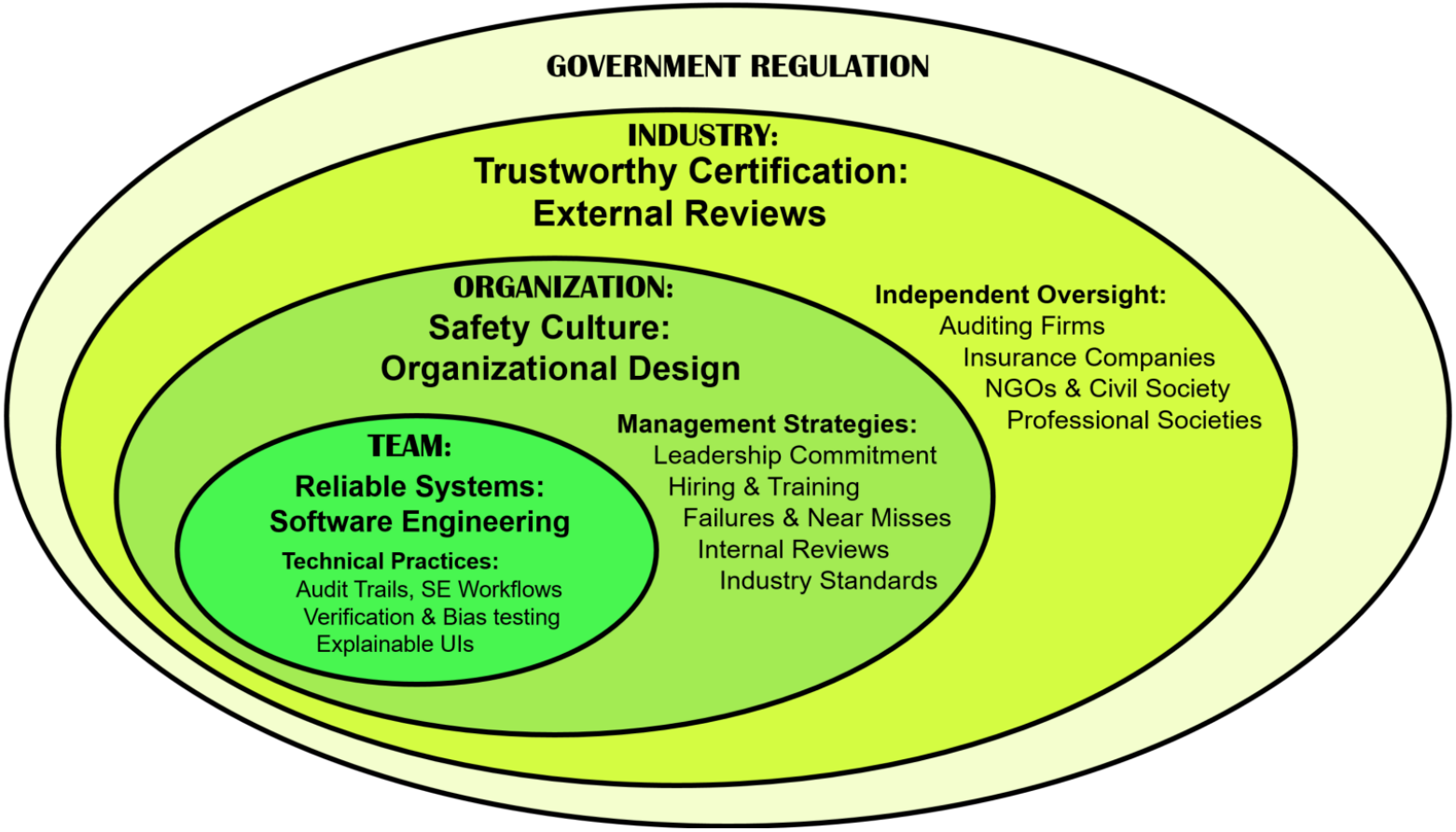


“Robots don’t perform surgery. Your surgeon performs surgery with da Vinci by using instruments that he or she guides via a console.”

https://www.davincisurgery.com/

Bloomberg Terminal

Governance Structures for Human-Centered AI



HCAI Attributes that Are Candidates for Assessment

General virtues of the system itself

- **Trustworthy**: Can users trust the system to perform correctly?
- **Responsible/Humane**: Has the system been designed, developed, and tested in a responsible way?
- **Ethical Design**: Were stakeholders involved in the design?
- **Ethical Data**: Was the data collected in an ethical manner?
- **Ethical Use**: Will the system's outcome be used in an ethical manner?
- **Well-being/Benevolence**: Does the system support human health, comfort, and values?
- **Secure**: How vulnerable is the system to attack?
- **Private**: Does the system protect a person's identity and data?

Performs well in practice

- **Robust/Agile**: Does the system perform well when inputs change?
- **Reliable/Dependable**: Does the system do the right thing?
- **Available**: Is the system running when needed?
- **Resilient/Adaptive**: Can the system recover from disruptions?
- **Testable/Verifiable/Validatable/Certifiable:** Can be tested to verify adherence to requirements?
- **Safe**: Does the system have a history of safe use?

Clarity to stakeholders

- **Accurate**: Does the system deliver correct results on test cases and real world cases?
- **Fair/Unbiased**: Are the system outcomes unbiased?
- **Accountable/Liable**: Who or what is responsible for the system's outcome?
- **Transparent/Open**: Is it clear to an external observer how the system's outcome was produced?
- **Interpretable/Explainable/Intelligible/Explicable:** Can the explain why an outcome has occurred?
- **Usable**: Can a human use it easily?

Enables independent oversight

- **Auditable**: Can the system be audited by others for retrospective forensic analysis of failures?
- **Trackable:** Does the system display status and next steps so human intervention is possible?
- **Traceable:** Is the system designed to allow tracing back from an outcome to the root cause?
- **Redressable**: Is there a process for those harmed to request review and compensation?
- **Insurable**: Does the design permit insurance companies to offer policies?
- **Recorded**: Does the system record activity for retrospective forensic review?
- **Open**: Is code and data publicly available for others to review?
- **Certified**: Have certification bodies reviewed and approved the system?

Complies with accepted practices

- **Compliant with standards**: Does the system comply with relevant standards, e.g. IEEE P7000 series?
- **Compliant with accepted software engineering workflows:** Was a trusted process used?


Summary

Human-Centered AI

Amplify, Augment, Enhance & Empower People

→ 1000-fold improvements in capabilities

Information	Photography
Search	Navigation
Email & Text	Business Formation

→ RST: Reliable, Safe & Trustworthy

→ Human self-efficacy, creativity & responsibility

→ Human values, rights & dignity

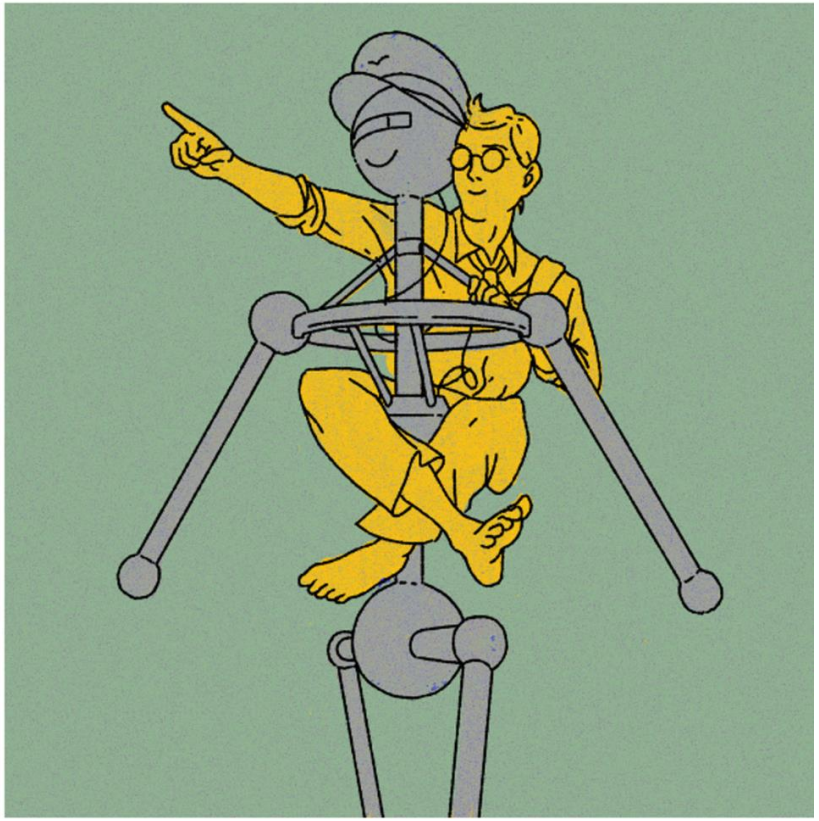
IJHCI (2020) **https://www.tandfonline.com/doi/full/10.1080/10447318.2020.1741118**

A Case for Cooperation Between Machines and Humans

A computer scientist argues that the quest for fully automated robots is misguided, perhaps even dangerous. His decades of warnings are gaining more attention.

By John Markoff

May 21, 2020 Updated 3:09 p.m. ET



https://www.nytimes.com/2020/05/21/technology/ben-shneiderman-automation-humans.html

Human-Centered Artificial Intelligence: Reliable, safe & trustworthy, *International Journal of Human-Computer Interaction 36,* 6 (March 2020). https://doi.org/10.1080/10447318.2020.1741118

Design lessons from AI's two grand goals: Human emulation and useful applications, *IEEE Transactions on Technology & Society 1,* 2 (June 2020). https://ieeexplore.ieee.org/document/9088114

Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy Human-Centered AI systems, *ACM Trans. on Interactive Intelligent Systems 10,* 4 (Oct 2020). https://dl.acm.org/doi/10.1145/3419764

Human-Centered Artificial Intelligence: Three fresh ideas, *AIS Trans. on Human-Computer Interaction 12,* 3 (Oct 2020). https://aisel.aisnet.org/thci/vol12/iss3/1/

Summary & resources: https://hcil.umd.edu/human-centered-ai/


APPLIED RESEARCH LABORATORY FOR
INTELLIGENCE
AND SECURITY
UNIVERSITY OF MARYLAND

Q&A