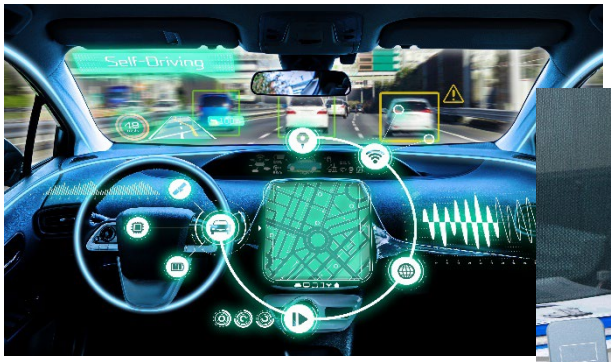
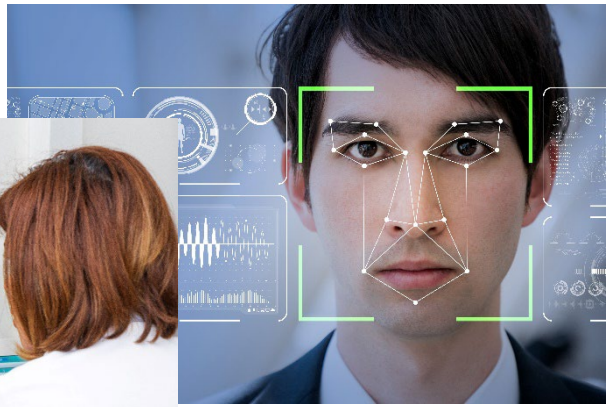
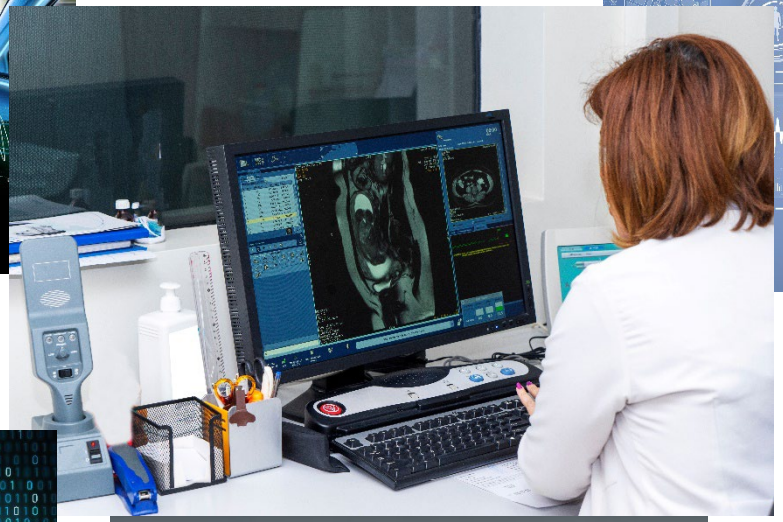
CONSIDERATIONS IN DESIGN & EVALUATION OF 'INTELLIGENT' DECISION AIDS

Emilie Roth
Roth Cognitive Engineering

(Promise of) Application of AI is Ubiquitous







However…

Challenges Remain

- While AI systems may perform well in the lab – they can fail when introduced in the field
- Automation 'bias' can often lead to worse performance than an individual working on their own
- Design to support Human-AI ***joint performance*** will be critical for success

55° Stanford, CA >
Wed, Nov 11, 2020

Newsweek

U.S. | World | Business | Tech & Science | Culture | Newsgeek | Sports | Health | The Debat

TECH & SCIENCE

Google's AI Health Screening Tool Claimed 90 Percent Accuracy, but Failed to Deliver in Real World Tests

BY JASON MURDOCK ON 4/28/20 AT 11:53 AM EDT

Google's medical AI was super accurate in a lab. Real life was a different story.

If AI is really going to make a difference to patients we need to know how it works when real humans get their hands on it, in real situations.

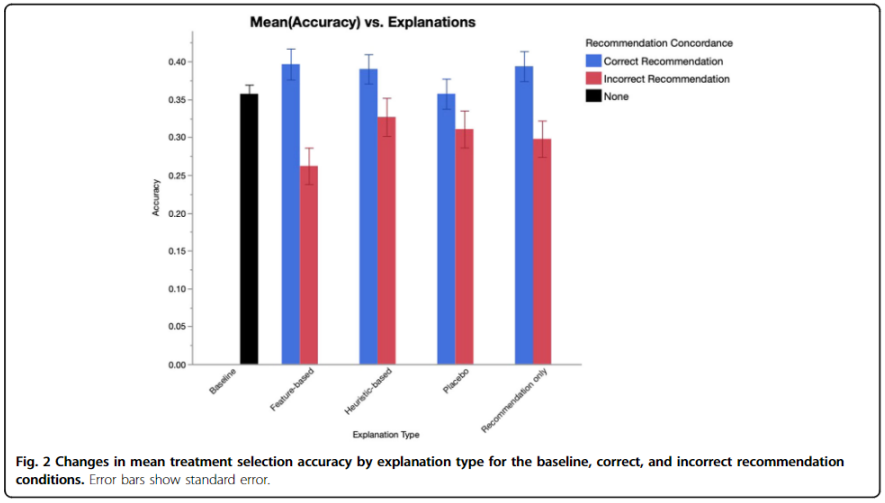


Fig. 2 Changes in mean treatment selection accuracy by explanation type for the baseline, correct, and incorrect recommendation conditions. Error bars show standard error.

Jacobs et al. (2021) How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry*.

Talk Outline

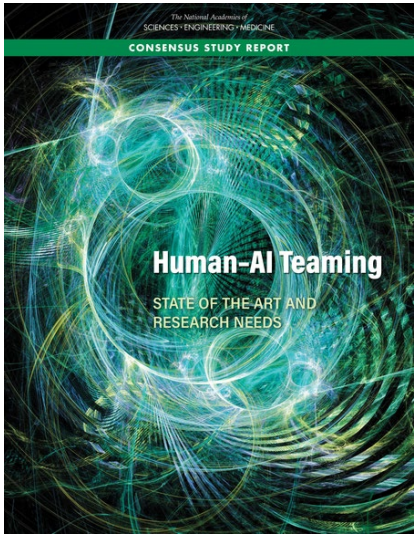
- Findings of recent National Academies Consensus Study on Human-AI Teaming

- Implications for design and evaluation of new forms of 'Intelligent' technology

Consensus Study Report

Committee Members:

- MICA R. ENDSLEY, (Chair) SA Technologies
- BARRETT S. CALDWELL, Purdue University
- ERIN K. CHIOU, Arizona State University
- NANCY J. COOKE, Arizona State University
- MARY L. CUMMINGS, Duke University
- CLEOTILDE GONZALEZ, Carnegie Mellon University
- JOHN D. LEE, University of Wisconsin-Madison
- NATHAN J. MCNEESE, Clemson University
- CHRISTOPHER MILLER, Smart Information Flow Technologies
- EMILIE ROTH, Roth Cognitive Engineering
- WILLIAM B. ROUSE, NAE Georgetown University



https://www.nap.edu/download/26355

Why Human-AI Teaming?

Human-AI team is defined as "one or more people and one or more AI systems requiring collaboration and coordination to achieve successful task completion"

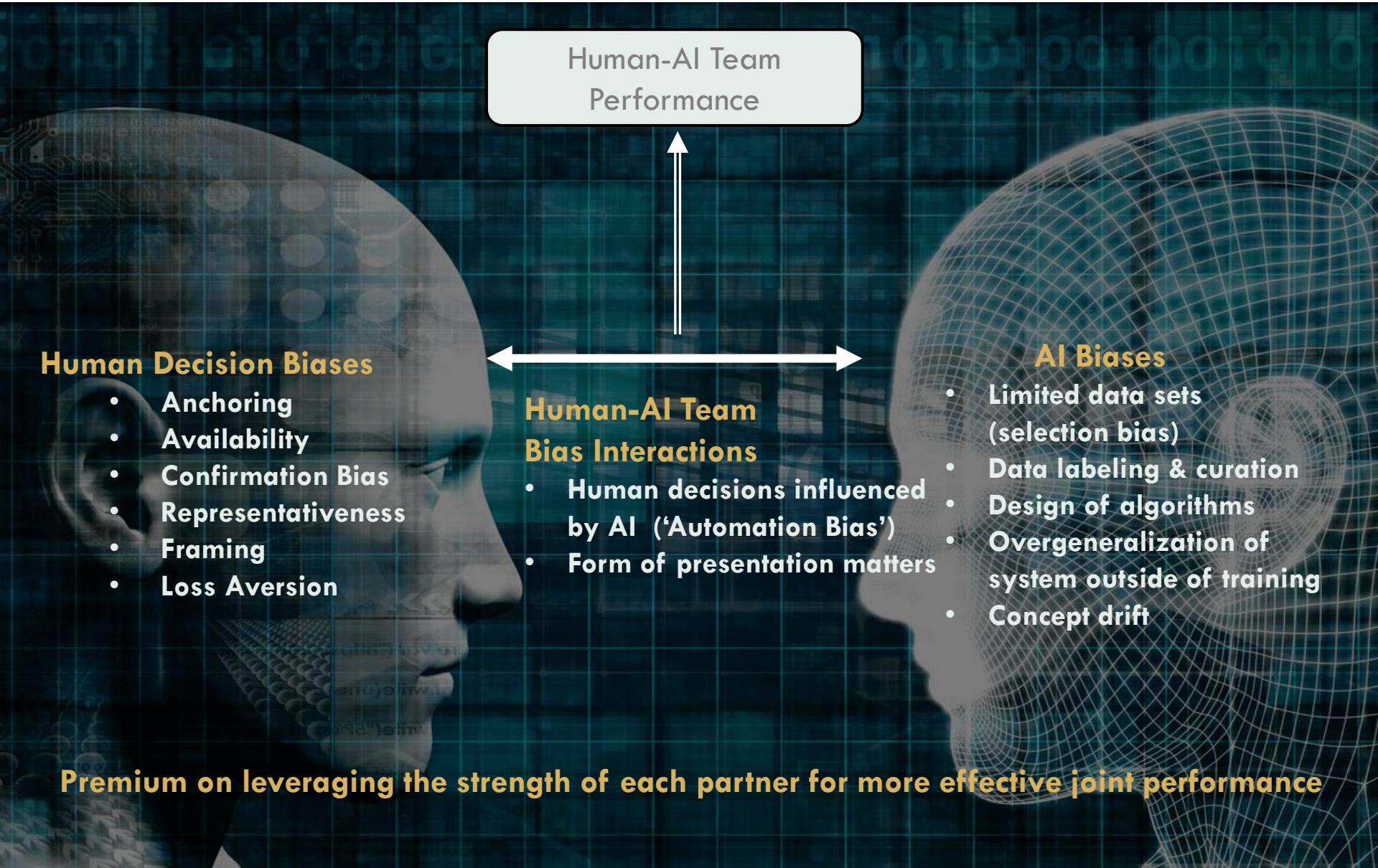
- Qualifies as a team
  - Common goals
  - Specific roles
  - Interdependence
- Teaming provides
  - Mutual support and back-up
  - Adapt to changing demands
- Does not imply humans and AI are equivalent
  - Functionality
  - Capabilities
  - Authority
  - Responsibilities

- Research on Human-Human Teams provides a starting point
  - Teamwork skills
  - Team SA
  - Team training
- Research on Human-Automation Interaction and Autonomy relevant

Human-Centered AI

- Designing an AI system to work well as a teammate increases human-centeredness
- Augments human capabilities and raises performance beyond that of either entity

Human-AI Team Bias



Fostering Effective Teaming

Human-AI Team Performance

Teamwork Coordination

- **Ensure goal alignment**
- **Support collaboration (including anticipation and back-up)**

Shared Mental Models

(Self)

Model of Human
(Capabilities, Limits
Behaviors, Functionality)

Goals, Functions, Tasks, Plans
State, Modes

Sufficiency of World SA
Ability to Perform Tasks
Actions, Outputs & Effectiveness
Impact of Actions on Others

Projected Actions

SA of Human Teammate *(or self)*

Model of the World
(Terrain, Environment, Enemy, Friendly, Civilian, Systems, Methods of Operation)

Current State
Confidence in Information

Impact on Goals
Conformance with Goals

Projected Actions/State

SA of World

Model of AI
(Capabilities, Limits
Behaviors, Functionality)

Goals, Functions, Tasks, Plans
State, Modes

Sufficiency of World SA
Ability to Perform Tasks
Actions, Outputs & Effectiveness
Impact of Actions on Others

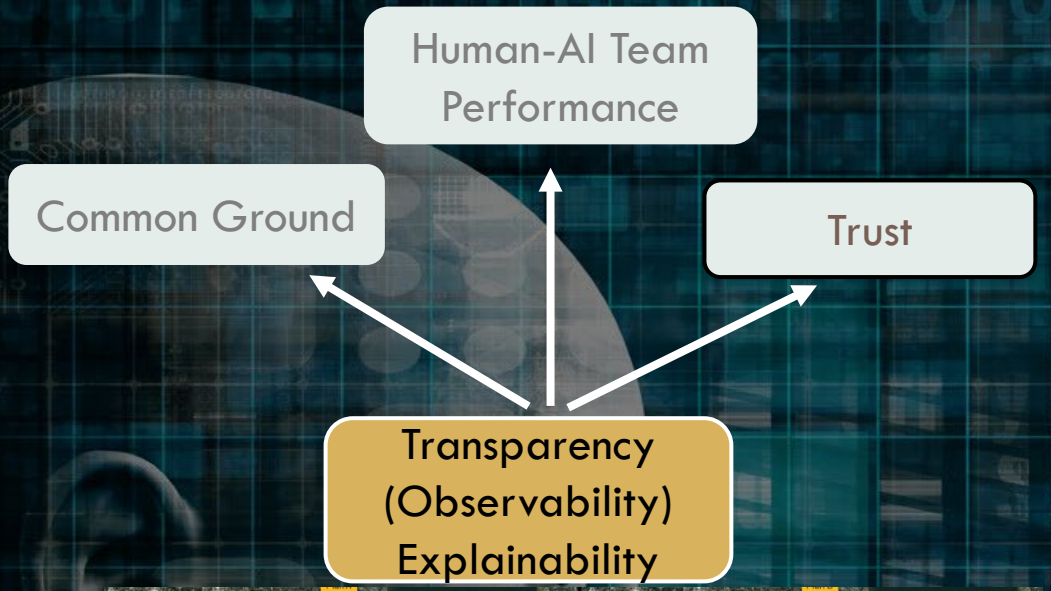
Projected Actions

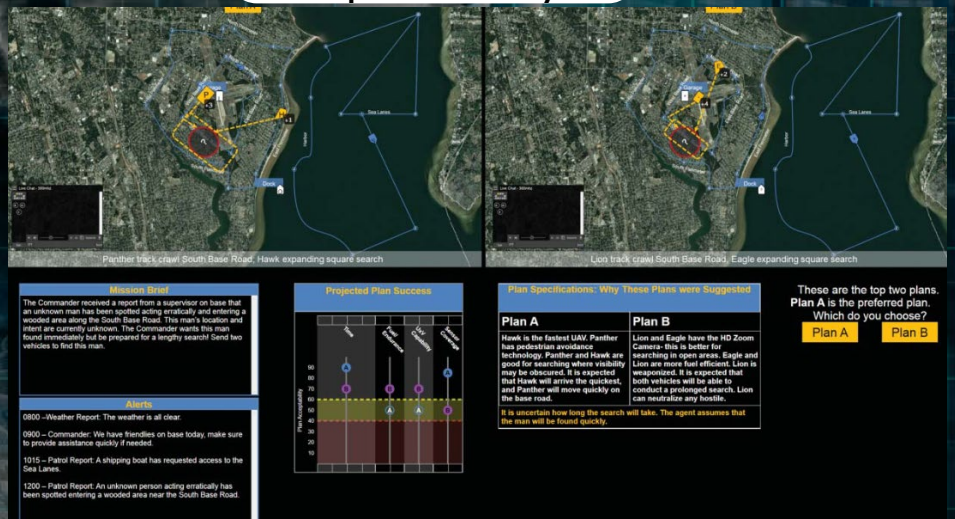
SA of AI Teammate *(or self)*

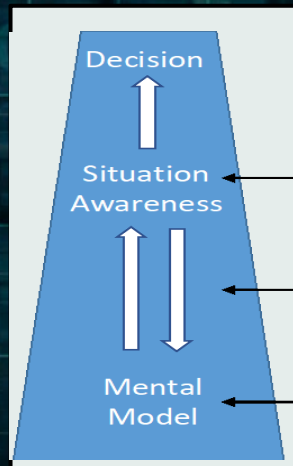
(Self)

Shared Situation Awareness ('Common Ground')

AI Transparency & Explainabilty







Display Transparency

Explainability

Training

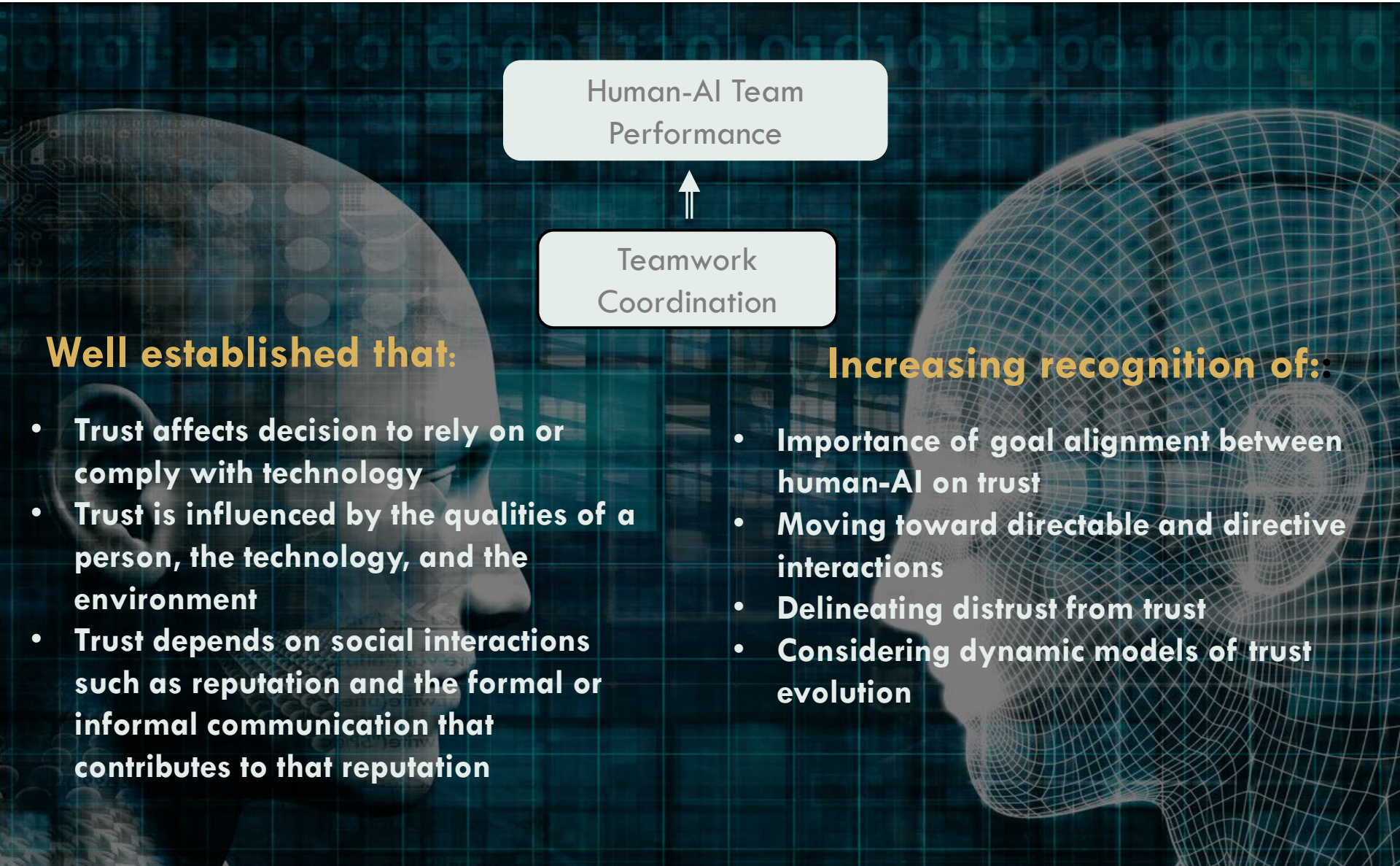
Display transparency

- Provides a real-time understanding of the actions of the AI system:
  - Goals
  - Progress
  - Constraints/Affordances
  - Projection to future
  - Potential limitations

Explainability

- Provides information in a backward-looking manner on logic, factors, or reasoning

Human-AI Team Trust



Well established that:

- Trust affects decision to rely on or comply with technology
- Trust is influenced by the qualities of a person, the technology, and the environment
- Trust depends on social interactions such as reputation and the formal or informal communication that contributes to that reputation

Increasing recognition of::

- Importance of goal alignment between human-AI on trust
- Moving toward directable and directive interactions
- Delineating distrust from trust
- Considering dynamic models of trust evolution

Human-System Integration (HSI) Considerations

HSI incorporates human-centered analyses, models, and evaluations throughout the development and implementation lifecycle so as to mitigate the risk of downstream system failure.

Issues Raised and Research Needs

Human-AI Team Design and Testing Methods

Human-AI Interaction Design for Effective Joint Performance

Human-AI Team Development Teams

AI System Lifecycle Testing and Auditability

AI Cyber Vulnerabilities

Human-Systems Integration for Agile Software Development

Implications for Design and Evaluation of New 'Intelligent' Technologies

People

Context of Work

Technology

A Cognitive Engineering Perspective

Importance of:

- Understanding the ***context of work***:
  - Complications that can challenge performance of human or AI agent.
- Designing systems to optimize the ***joint*** Human-AI team performance.
  - Success often leverages the strength of people on the scene (*Collaborative Automation / Shared Autonomy)*
  - Technology needs to be observable, understandable, and directable.
- Evaluating the ***joint*** Human-AI Team
  - For more resilient performance

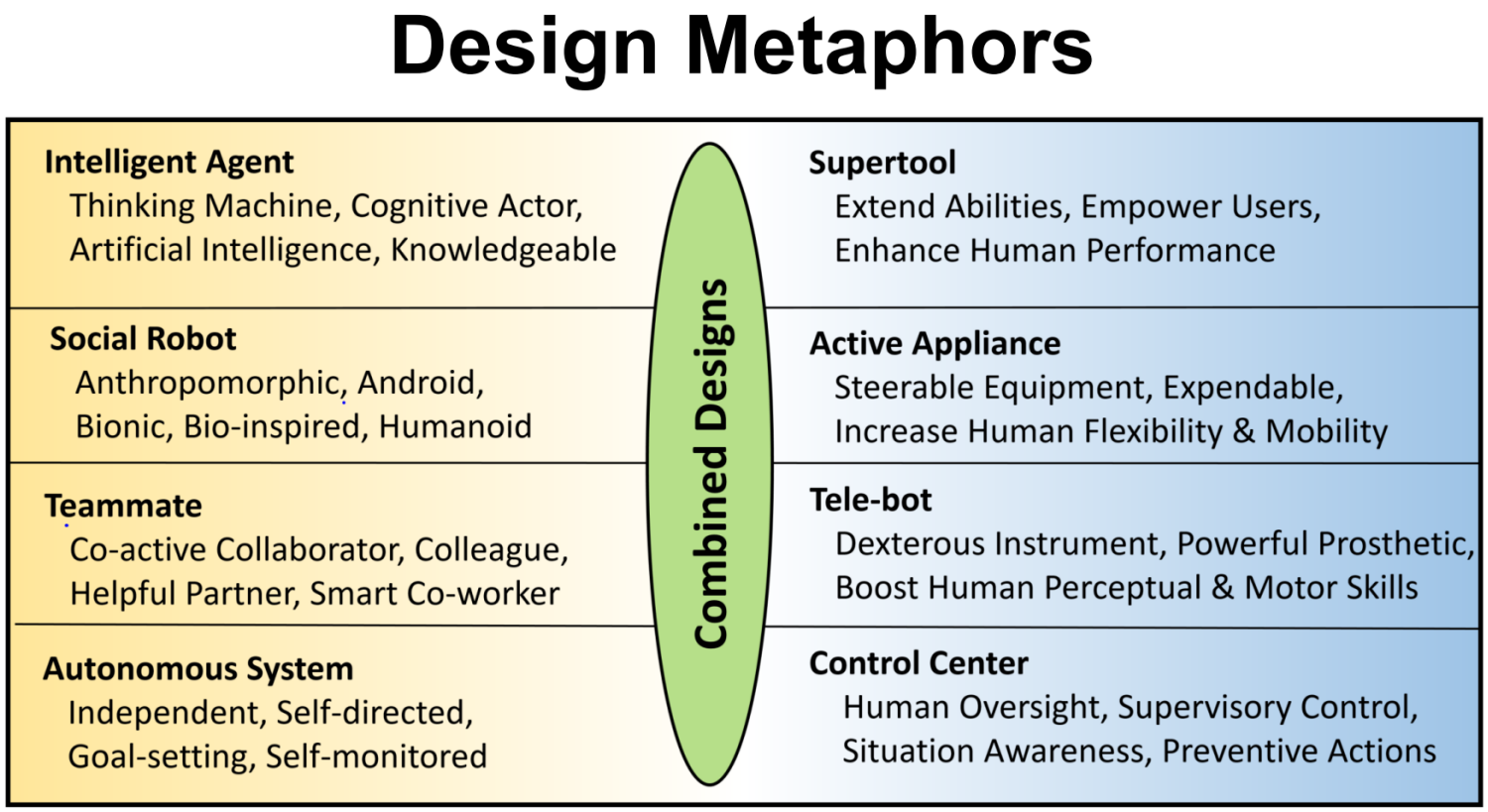
Evaluating the Joint Human-AI Team

- Include a range of scenarios
  - Straightforward 'textbook' cases
  - 'Edge Cases' at the boundaries of the capabilities of the AI system
- Employ multiple evaluation measures:
  - Measure of objective joint performance
  - Measures of Trust
  - User Mental Models of how the AI system works
  - User evaluations of the AI system via post-study questionnaires.

	Positive	Negative
AI/XAI System	(1) How the System works: Parts, connections, functions, relationships, control logic	(3) How the system Fails: Breakdowns, limitations
The User/Learner	(2) How to make the System work: Detecting anomalies, appreciating the System's responsiveness, performing workarounds and adaptations	(4) How the User/Learner gets confused: The kinds of errors the User made, or other Users might make

The mental model matrix
(Klein, Borders, Hoffman, & Mueller, 2021)

Human-AI Teaming is One of Many Metaphors



Ben Shneiderman new book 'Human-Centered AI' (2022)

Conclusions

- AI systems should be designed to support the needs of people who will have the ultimate responsibility for the outcomes.
- An important measure of success is the joint human-AI team performance.
- This requires making AI systems better 'team players':
  - Observable
  - Understandable
  - Directable
- The National Academies Consensus Study Report presented 57 inter-related research objectives to meet this vision
  - Near, Mid and Far Term

Relevant References

Chiou EK, Lee JD. Trusting Automation: Designing for Responsivity and Resilience. Human Factors. April 2021. doi:10.1177/00187208211009995

Committee on Human-Systems Integration Research Topics (2021). *Human-AI Teaming: State of the art and research needs*. Washington DC: The National Academies Press.

Klein, G., Borders, J., Hoffman, R. R. and Mueller, S. T. (2021). 'A method for evaluating users' understanding of XAI Systems: The Mental Model Matrix. Technical Report, DARPA Explainable AI Program.

Roth, E. M., Sushereba, C., Militello, L. G., Diiulio, J., Ernst, K. (2019). Function allocation considerations in the era of human autonomy teaming. *Journal of Cognitive Engineering and Decision Making*, 13, 199-220.

Roth, E. M., DePass, B., Harter, J., Scott, R., Wampler, J. (2018). Beyond levels of automation: Developing More Detailed Guidance for Human Automation Interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 150-154.

Roth, E. M., DePass, B., Scott, R., Truxler, R., Smith, S. and Wampler, J. (2017). Designing collaborative planning systems: Putting Joint Cognitive Systems Principles to Practice. In P. J. Smith and R. R. Hoffman (Eds). *Cognitive Systems Engineering: The Future for a Changing World*. Boca Raton: Taylor & Francis, CRC Press. (247-268).

Shneiderman, B. (2022) Human-Centered AI. New York: Oxford University Press.

Thank you

Emilie Roth
Roth Cognitive Engineering
emroth@rothsite.com