2022 AAAI Spring Symposium
Putting AI in the Critical Loop: Assured Trust and Autonomy in Human-Machine Teams
21-23 March 2022

Deferring Decisions: Effects of Varying Interaction Structures on Human-AI Performance

22 March 2022

Erin K. Chiou, Ph.D.
Assistant Professor of Human Systems Engineering
Arizona State University

erin.chiou@asu.edu
@ErinChiou

1

Acknowledgments

- Study team members and co-authors: Pouria Salehi, Mickey Mancenido, David Mosallanezhad, Myke Cohen, and Aksheshkumar Shah
- This material is based on work supported by the U.S. Department of Homeland Security [17STQAC00001-04-00, 17STQAC00001-05-00] and the Air Force Office of Scientific Research [FA9550- 18-1-0067]. The views and conclusions contained in this presentation should not be interpreted as representing the official policies, either expressed or implied, of the research sponsor

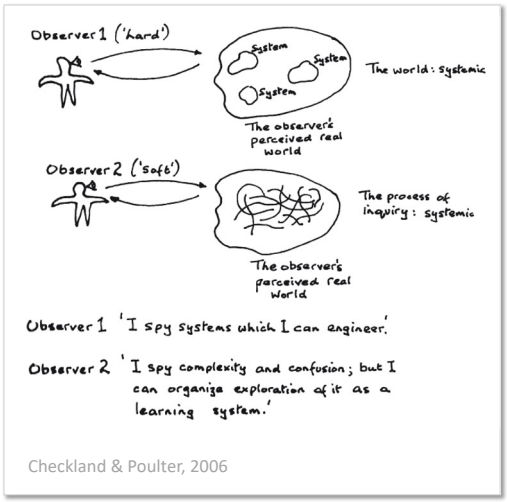
2

2

Control strategies in designing human-AI systems

- Hard and formal controls enforce behaviors of a system to cause a particular outcome
- Soft and informal controls influence behaviors of a system to achieve a particular outcome

(Beniger, 1986; Rosenblat & Stark, 2016; Björkdahl & Holmén, 2019)

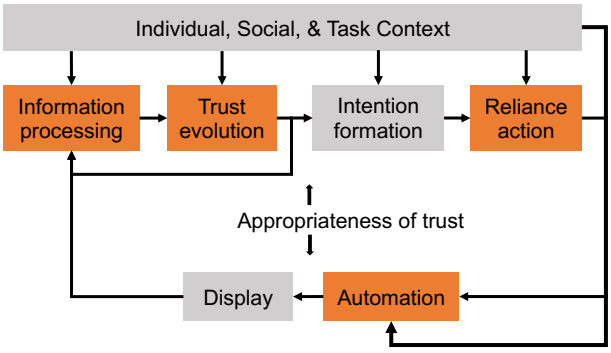


Checkland & Poulter, 2006

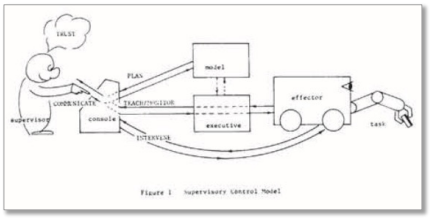
3

3

Soft control: Trust in supervisory control automation



The information processing view of trust in automation adapted from Lee & See (2004)



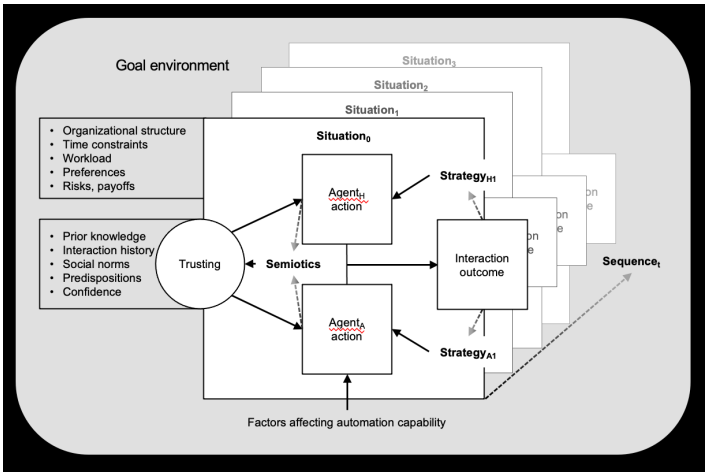
Sheridan (1975) schematic of supervisory control, evolving from direct teleoperated control

4

4

Trusting increasingly autonomous systems

Chiou & Lee, 2021. Trusting automation: Designing for responsivity and resilience. *Human Factors*. https://doi.org/10/gjvcr2



A relational framing of trust summarized by the four concepts:
Situation
Semiotics
Sequence
Strategy

Cited in:
National Academies of Sciences, Engineering, and Medicine, Board on Human Systems Integration. *2021. Human-AI Teaming: State of the Art and Research Needs*. Washington, DC: The National Academies Press.

5

5

Concept of interaction structures

- Many studies of trust focus on supervisory control structures
- However, new AI-enabled capabilities are poised to generate more interactive situations, and different resulting decision structures
- Previously, we studied human-agent interaction from the perspective of cooperative control structures, looking at the effects of negotiated exchange and reciprocal exchange in a joint hospital resource management task (Chiou & Lee, 2015; 2016; 2019)
- We applied this same lens of thinking-through the nuances of interaction structures to address stakeholder concerns with respect to operator trust in a new identity management system being piloted by TSA (2019)

6

Decision Deferral: Rate of Deferral

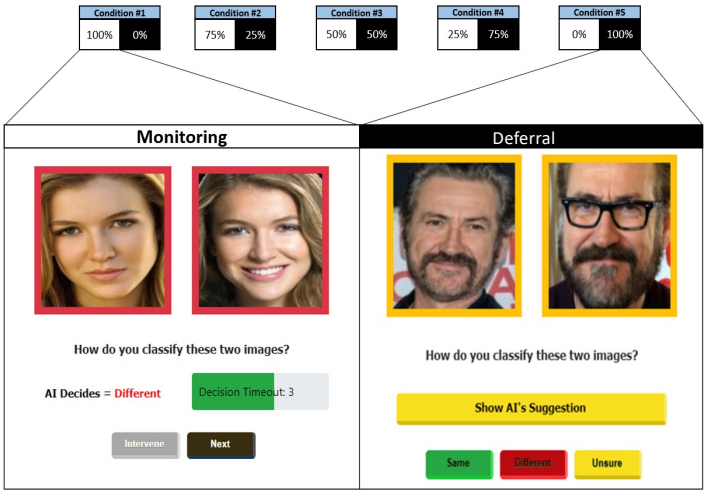
- Inspired by a framework for human-machine decision-making in which a machine defers a decision to a human, given some threshold of uncertainty (Madras, Pitassi, & Zemel, 2018)
- We designed a face-matching task for human-machine joint security screening to investigate the effects of varying deferral rates and interaction structures on trust and performance
- Five deferral rate conditions were tested between-subjects at 0, 25, 50, 75, and 100% deferral, from fully monitoring to fully manual
- Our stakeholders were interested in the following:
  - What deferral rates are most effective for joint system performance?
  - How do deferral rates affect other performance metrics including workload and trust in the AI?

7

7

Decision Deferral: Rate of Deferral

- From stakeholder input, AI would not be deployed without human monitoring, so two interaction structures were operationalized: Monitoring and Deferral
- An open-source AI that employs 1-to-1 face verification Siamese network with 95% accuracy
- AI's input accuracy from decision deferral framework = 52%
- VGG-Face Dataset (Parkhi et al., 2014) with 2,622 celebrity identities



8

8

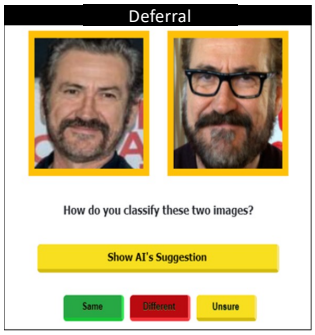
Decision Deferral: Rate of Deferral

- *N* = 96 from Amazon Mechanical Turk (Mturk) , U.S.-based, mostly men with 4-year college degrees
- Several data quality safeguards were employed (Mancenido et al., 2021)
- Initial results showed that higher deferral rates (75% and 100% hard deferral) are associated with higher sensitivity and lower workload, and lower throughput and lower trust in the AI-enabled system (i.e., more monitoring more perceived workload)
- The most effective deferral rates for increasing throughput and maintaining trust in the AI were the 25% and 50% deferral conditions, but at a cost of lower sensitivity and higher workload
- …Not surprising results if you have read the supervisory control automation literature from the past half-century

9

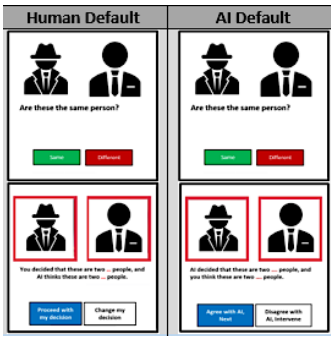
9

Deferral Interaction Structures



Study 1: Seeking Advice vs. Monitoring

- Participants have the option to "Show AI's Suggestion" to help determine their judgement
- However, it is their choice whether to consult the AI before making their judgement



Study 2: Checking Agreement vs. Monitoring

- Participants first make a judgment and then are presented with both their determination and the judgement made by the AI
- The judgment is either framed as their decision or as the decision of the AI

10

10

Decision Deferral: Defaults and Difficulty

- In a second study, we focused on the 25% and 75% deferral conditions
- We looked at whether framing the default decision as either originating from the human or the AI affected decisions to rely on the AI and performance in the task (Johnson & Goldstein, 2003)
- After making a deferral decision, participants would be presented with a second screen that showed either:
  - "You decided that these are different [the same] people [person]" with the buttons: PROCEED WITH MY DECISION | CHANGE MY DECISION
  - "The AI decided…." with the buttons: AGREE WITH THE AI | DISAGREE WITH THE AI

11

11

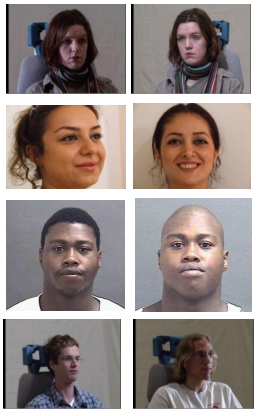
Decision Deferral: Defaults and Difficulty

(a)

(b)

(c)

(d)



- Sample Pairs including (a) easy match, (b) easy mismatch, (c) difficult match, (d) difficult mismatch
- We conducted a separate study to account for task difficulty; the AI's resulting input accuracy in this study on the selected image pairs was 66% on average
- The face database used in this study included images from a variety of sources including VGG-Face database version 1 (Parkhi, Vedaldi, and Zisserman 2014), Multi-PIE database (Gross et al. 2010), Iranian Face database (Bastanfard, Nik, and Dehshibi 2007), and MORPH academic database (FaceAgingGroup 2007)

12

12

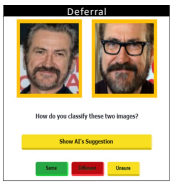
Decision Deferral: Defaults and Difficulty

- *n* = 86 participants from Mturk, sample skewed young (30s), white, counterbalanced across men and women, with 4-year college degrees
- Generally, a higher monitoring rate (lower deferral rate) led to:
  - higher joint performance in terms of accuracy and sensitivity, but also higher disagreements with the initial decision (and correct refusals)
  - faster completion times, lower incorrect decisions, misses, and incorrect single-step decisions
- Operators in the human default condition were less accurate than the AI default and agreed more with the AI even when the AI was wrong …an argument for ~~more automation~~ caution in framing deferral interaction structures?
- There were no differences between groups regarding perceived workload and trust (or any of our control variables like automation complacency)

13

13

Deferral Interaction Structures

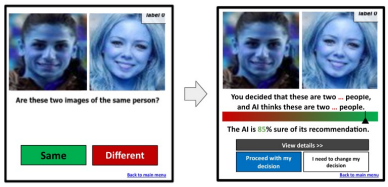


Study 1: Rate of Referral—Seeking Advice

- Higher deferral rates are associated with higher sensitivity and lower workload, and lower throughput and lower trust in the AI-enabled system

Study 2: Defaults and Difficulty—Checking Agreement

- Measures of perceived task load and trust were not significantly affected
- This suggests that when the task is difficult, differences in perceived trust and task load are negligible because people cannot distinguish errors in AI performance



Ongoing Study: Multisource AI Scorecard Table (MAST)—Checking Validity

- Participants make a judgment and then are presented with both their determination and the judgement made by the AI
- Participants are also provided information regarding the confidence of the AI and other details supporting its judgment

14

14

Advancing trust theory and measurement

- Trust was measured in these studies, but few meaningful differences between conditions
- New relational framing that factors in the goal environment – will the perceived purpose of the AI come to matter more than perceptions of process and performance?
  - Does the agent share my goals?
  - Do I understand how the agent is helping me reach my goals?
  - Is the agent good at helping me achieve my goals?
- Sensitivity of the trust questionnaire for our task context?
  - Using a widely-cited instrument developed via word elicitation and factor analysis with English majors
  - Not directly based on the purpose, process, performance

15

15

Trust and the "Multisource AI Scorecard Table" (MAST)

(Blasch, Sung, & Nguyen 2020)

- Based on analytic tradecraft standard ICD 203
- How effective is MAST at predicting trust perceptions and behaviors with AI-enabled decision support systems?
- Evaluating MAST against validated trust and message credibility instruments
  Jian, Bisantz & Drury 2000; Chancey et al., 2017; Appelman & Sundar, 2015
- Participants will receive a description of an AI-enabled system and then be asked to rate the system.
- Dimension reduction (multiple correspondence analysis) and structural equation modeling to assess the relationships between items

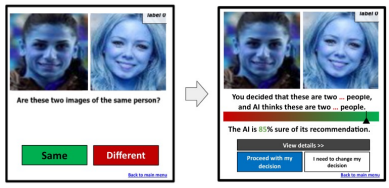
Nine criteria rated 0-3:
Sourcing
Uncertainty
Distinguishing
Analysis of Alternatives
Customer Relevance
Logical Argumentation
Consistency
Accuracy
Visualization

16

16

Trust and the "Multisource AI Scorecard Table"

"Readit" is an NLP-based text summarization system used in an intelligence analyst task



"Facewise" is a CNN-based face ID verification system used in transportation security screening

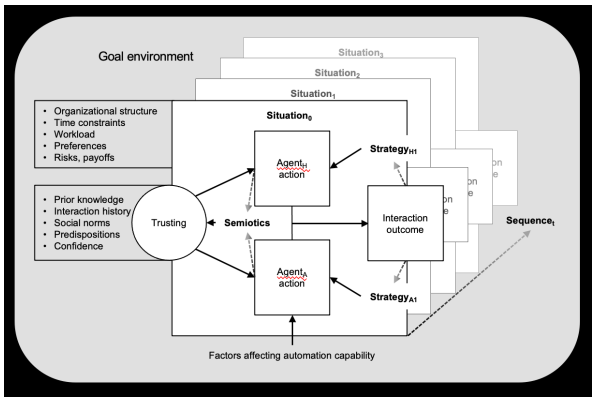
- Evaluate differences between ratings, trust perceptions, and behaviors
- Testing two different AI-enabled decision support systems to assess the generalizability of MAST
- The project is funded by the Center for Accelerating Operational Efficiency (U.S. DHS) with stakeholders

  DHS Office of Intelligence & Analysis
  TSA Human Performance Branch
  US Naval Research Laboratory
  Air Force Research Laboratory and the
  National Institute of Standards and Technology

17

17

Embrace complexity and interactivity to support trust and autonomy in human-AI teams



Human systems engineering areas of expertise:
Trust construct
Trust measures
Trust outcomes
Exchange structures
Work system accountability
Resilience engineering

Methods:
Microworld design
Wizard-of-oz
Mixed methods field research
Collaborative research

erin.chiou@asu.edu
@ErinChiou

18

18