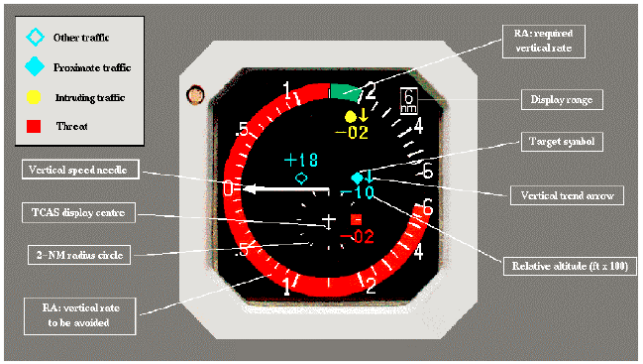
Trust in and dependence on imperfect automation

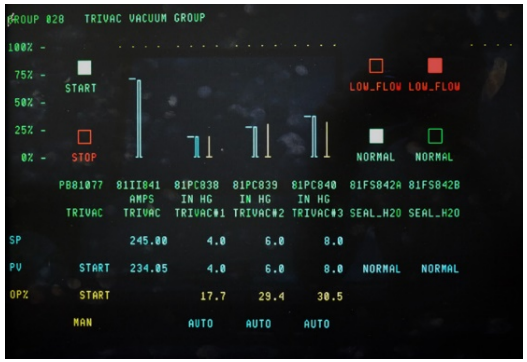
Jessie Yang, PhD
Industrial and Operations Engineering
University of Michigan, Ann Arbor

Automation dependence and performance| Motivation

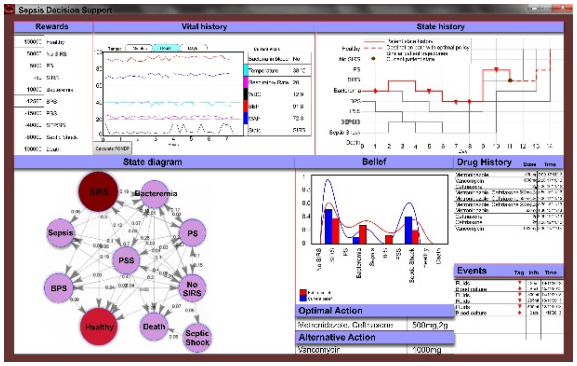
- Use of automation to assist human performance



Aircraft collision avoidance system



Alarm management



Clinical decision support system

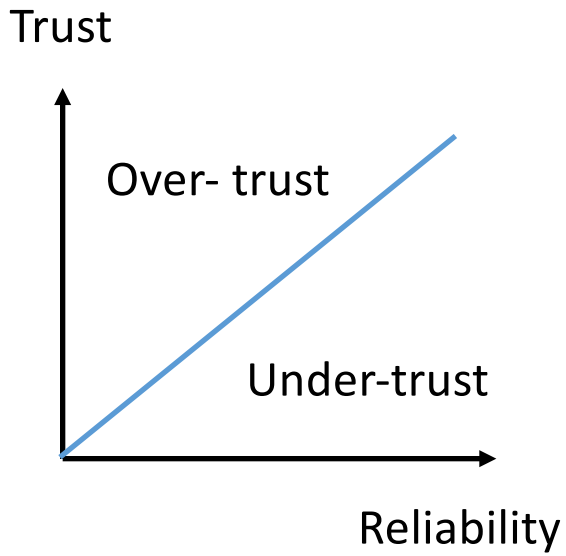
- Ideally – performance gain; In reality, not always..

Grounding of the Royal Majesty, 1995

Crash of Korean air flight 801, 2001

Automation dependence | Introduction

- Inappropriate use of automation[1,2]
- Two major reasons
  - Automation is sometimes imperfect[3,5]
  - Trust-reliability miscalibration[4,5]



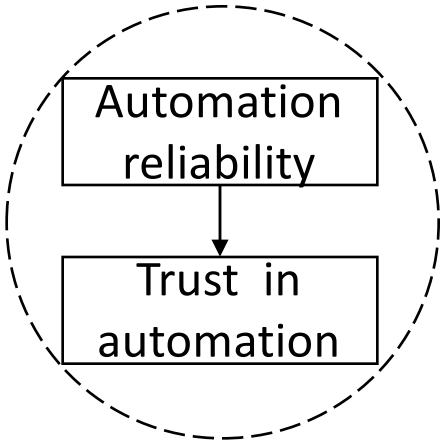
- Trust in automation - belief, intention, attitude[2], behavior
- Trust is an attitude, usage/dependence is a behavior

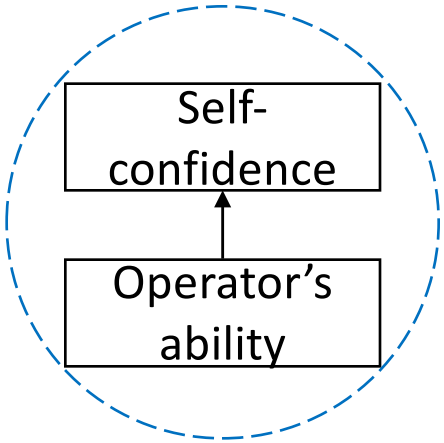
1. Barnes & Evans III, 2010; 2. Lee & See, 2004; 3. Wickens & Dixon, 2002; 4. Sheridan & Parasuraman, 2005; 5. Wickens, Levinthal, & Rice, 2010

Automation dependence | Two types of miscalibrations

- Little attention on the human operator's ability
- Another factor: self-confidence in performing a task manually[1,2,3]
- Overconfidence is highly likely[4], especially when tasks are difficult[5]
- Both types of miscalibrations should be modelled in human-automation interaction

Trust-reliability miscalibration

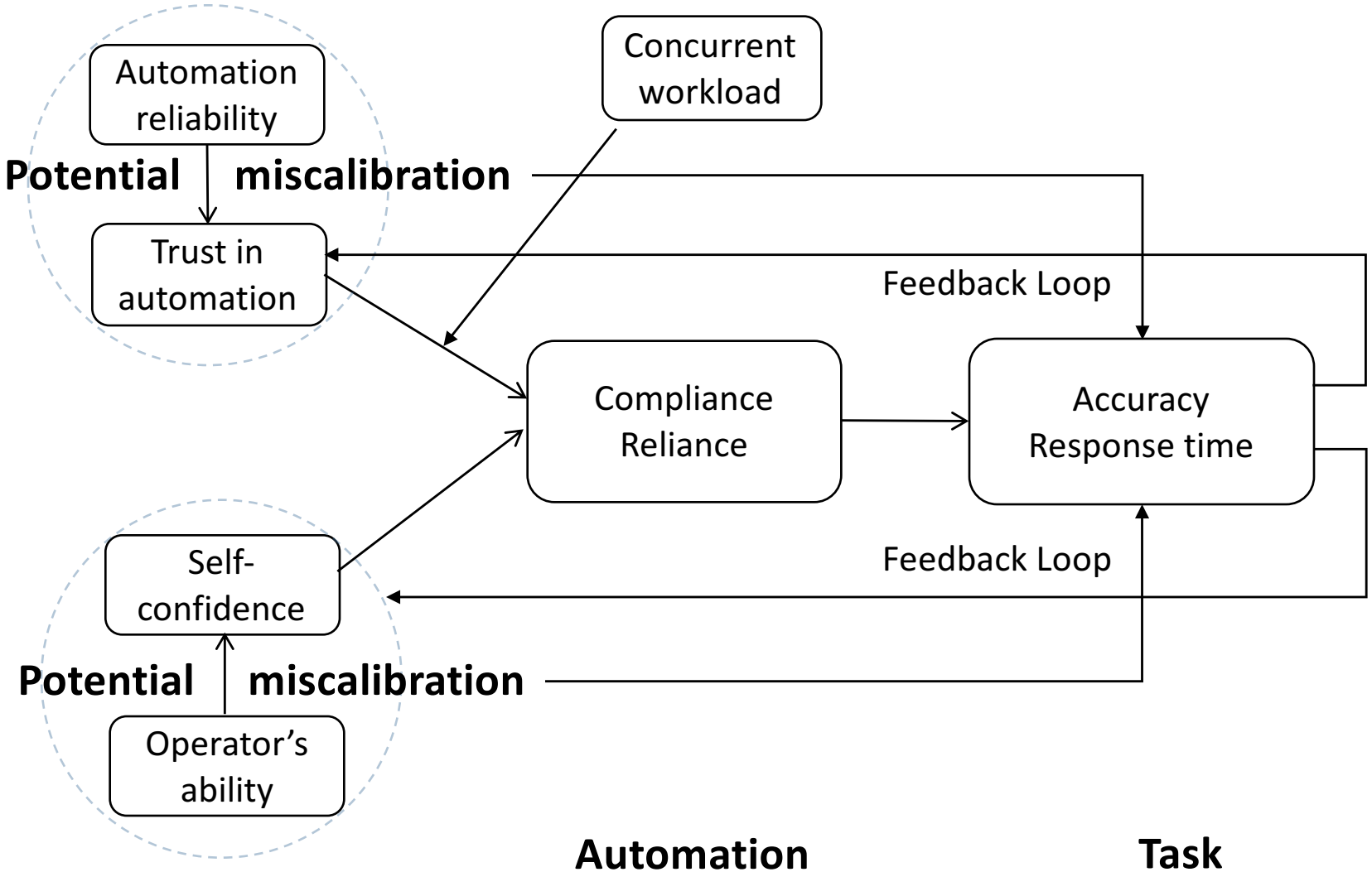




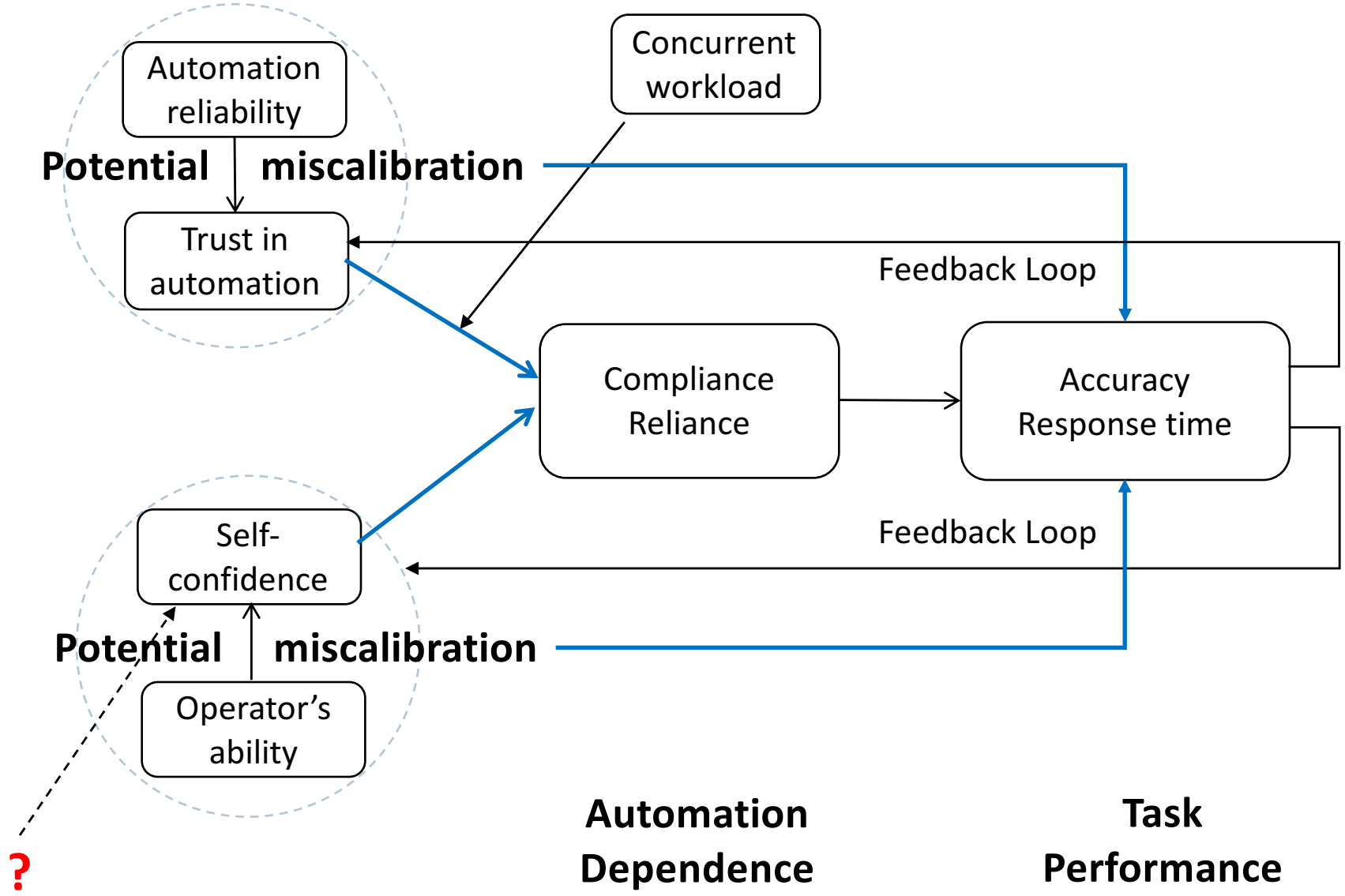
Confidence-ability miscalibration

1. de Vries, Middent, & Bouwhuis, 2003; 2. Lee & Moray, 1994; 3. Moray et al., 2000; 4. Kahneman, 2011; 5. Wickens et al., 2013

Automation dependence | Research model

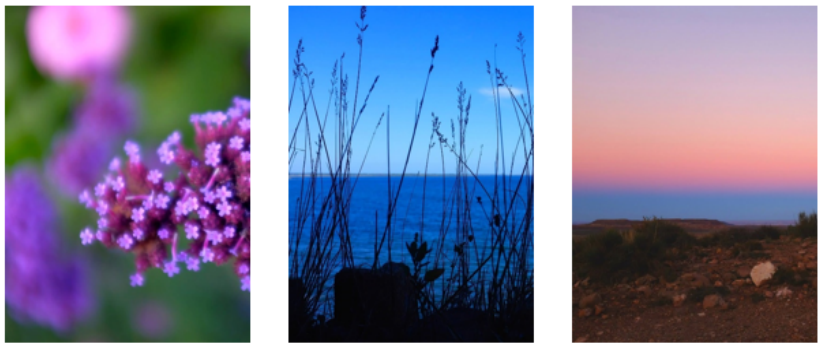


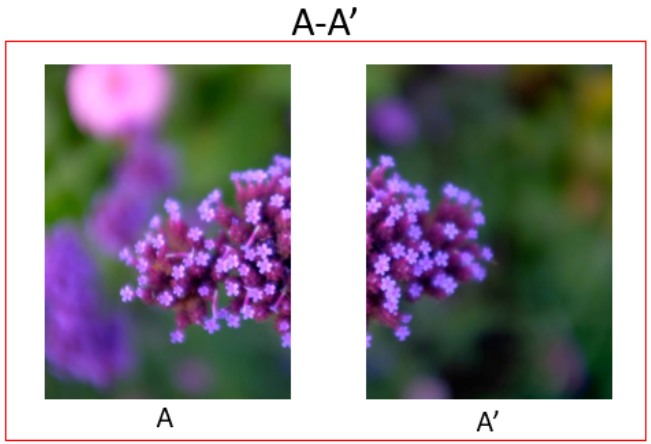
Study 1| Hypotheses



Study 1| Confidence-accuracy inversion[1]

Memorize → Recognize

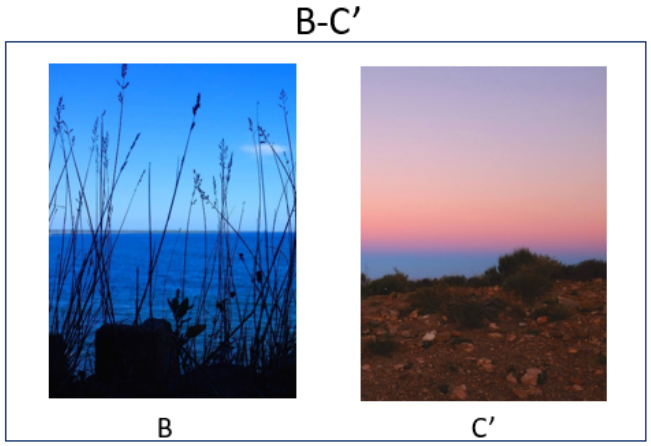




Accuracy ↑

Confidence ↓

Lower overconfidence



Accuracy ↓

Confidence ↑

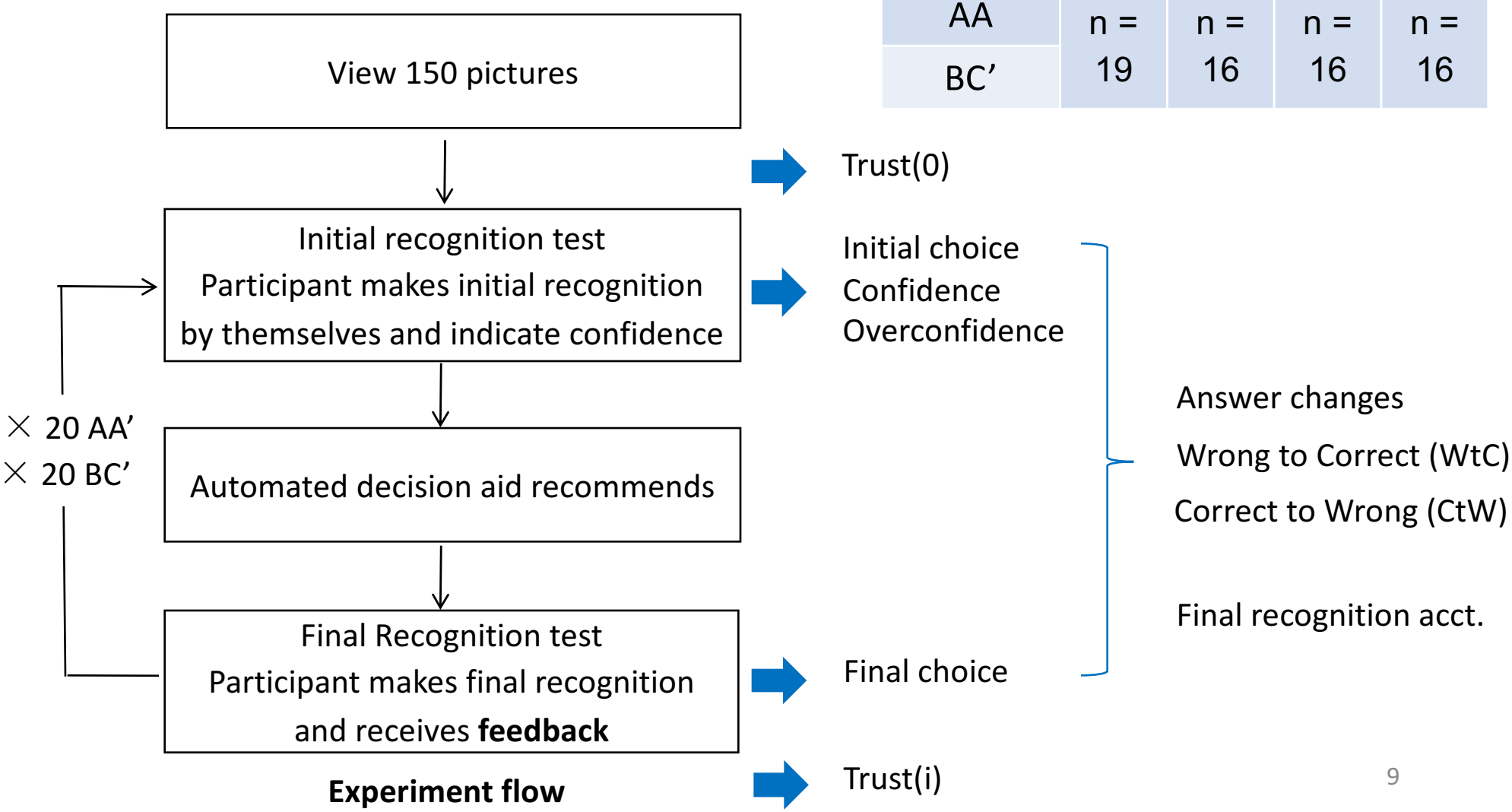
Higher overconfidence

1. Tulving, 1981

Study 1| Experimental Design

67 participants, age = 25.1 (SD = 3.8)

Reliability Picture	60%	70%	80%	90%
AA	n = 19	n = 16	n = 16	n = 16
BC’				



Experiment flow

Study 1| Results summary

No. of answer change = 17.43 **+** 0.025 **Trust***** **–** 0.045 **Confidence** ***

No. of answer change WtC = 22.81 **+** 0.038 **Trust***** **–** 0.046 **Confidence** ***

No. of answer change CtW = 22.81 **–** 0.003 Trust **–** 0.043 **Confidence** ***

- Self-confidence: strong and stable predictor of automation dependence

Study 1| Results summary

No. of answer change = 17.43 **+** 0.025 **Trust***** **–** 0.045 **Confidence** ***

No. of answer change WtC = 22.81 **+** 0.038 **Trust***** **–** 0.046 **Confidence** ***

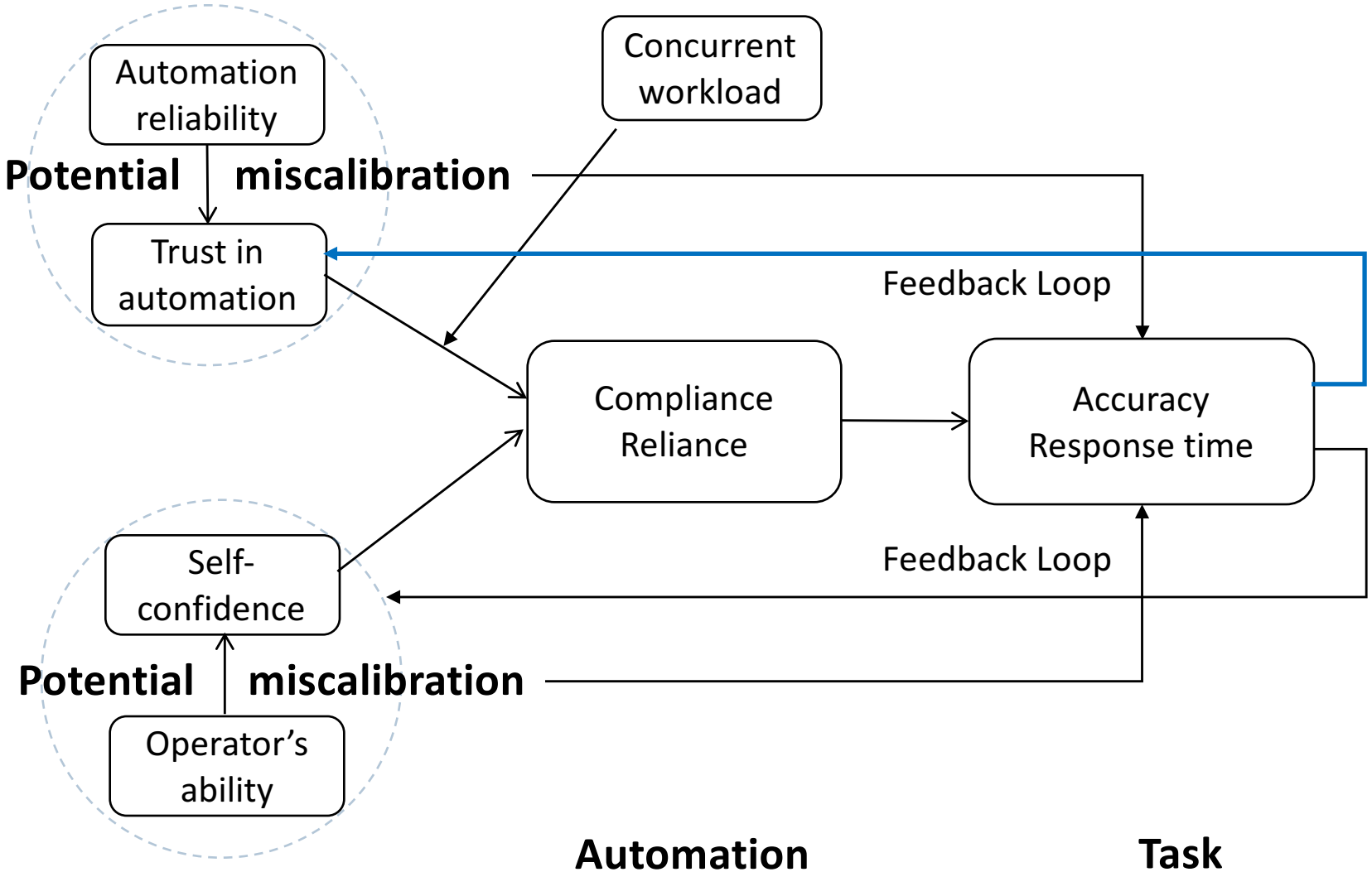
No. of answer change CtW = 22.81 **–** 0.003 Trust **–** 0.043 **Confidence** ***

- Self-confidence: strong and stable predictor of automation dependence

Final recognition accuracy = 19.94 + 0.56 Initial recognition accuracy***
– [if reliability < ability] * 0.38 **Overtrust****
– [if reliability > ability] * 0.26 **Overconfidence***

- Both types of miscalibrations harms performance

Automation dependence | Study 2



Automation dependence | Study 2

Consider the following hypothetical situation

Mark

Clive

Mark and Clive had exactly the same medical condition. They were presented to the hospital with stomachache

Consider the following hypothetical situation

Medical Decision Aid

Mark

Low risk

Endoscopy unnecessary

Clive

Low risk

Endoscopy unnecessary

Considering family history, age, dietary etc, a clinical decision support system suggested both Mark and Clive were at low risk of getting stomach cancer – endoscopy was unnecessary

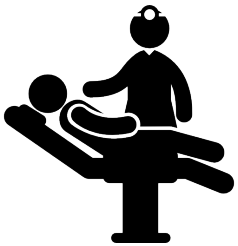
Consider the following hypothetical situation

Medical Decision Aid

Mark

Low risk

Endoscopy unnecessary



pre-cancerous polyps removed

Recovery

Clive

Low risk

Endoscopy unnecessary

diagnosed with stomach cancer years later

Unfortunately

Consider the following hypothetical situation

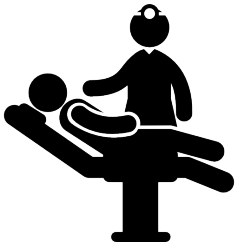
Medical Decision Aid

Trust

Mark

Low risk

Endoscopy unnecessary



Recovery

Similar Decrement ?

Clive

Low risk

Endoscopy unnecessary

Unfortunately

In both cases, the medical decision aid made mistakes – we expect a decrement in trust. But **similar decrement**?

Study 2| Results

Cond.	Initial recognition	Recommend-ation	Final recognition	Trust adjustment
0: 000	Wrong	Wrong	Wrong	-4.0(0.5)
1: 001	Wrong	Wrong	Correct	
2: 010	Wrong	Correct	Wrong	2.0 (0.3)
3: 011	Wrong	Correct	Correct	2.7 (0.5)
4: 100	Correct	Wrong	Wrong	-6.2(1.2)
5: 101	Correct	Wrong	Correct	-4.3(0.6)
6: 110	Correct	Correct	wrong	
7: 111	Correct	Correct	Correct	1.5 (0.2)

Automation success –> trust increases

Automation failure –> trust decreases

*** $p < .001$

Study 2| Results

Cond.	Initial recognition	Recommend-ation	Final recognition	Trust adjustment
0: 000	Wrong	Wrong	Wrong	-4.0(0.5)
1: 001	Wrong	Wrong	Correct	
2: 010	Wrong	Correct	Wrong	2.0 (0.3)
3: 011	Wrong	Correct	Correct	2.7 (0.5)
4: 100	Correct	Wrong	Wrong	-6.2(1.2)
5: 101	Correct	Wrong	Correct	-4.3(0.6)
6: 110	Correct	Correct	wrong	
7: 111	Correct	Correct	Correct	1.5 (0.2)

|Trust decrement| > |Trust increment|

*** $p < .001$

Study 2| Results

Cond.	Initial recognition	Recommend-ation	Final recognition	Trust adjustment
0: 000	Wrong	Wrong	Wrong	-4.0(0.5)
1: 001	Wrong	Wrong	Correct	
2: 010	Wrong	Correct	Wrong	2.0 (0.3)
3: 011	Wrong	Correct	Correct	2.7 (0.5)
4: 100	Correct	Wrong	Wrong	-6.2(1.2)
5: 101	Correct	Wrong	Correct	-4.3(0.6)
6: 110	Correct	Correct	wrong	
7: 111	Correct	Correct	Correct	1.5 (0.2)

Automation successes lead to **greater** increment of trust, if a user is **less capable** of completing the task on his or her own. * $p < .05$

Study 2| Results

Cond.	Initial recognition	Recommend-ation	Final recognition	Trust adjustment
0: 000	Wrong	Wrong	Wrong	-4.0(0.5)
1: 001	Wrong	Wrong	Correct	
2: 010	Wrong	Correct	Wrong	2.0 (0.3)
3: 011	Wrong	Correct	Correct	2.7 (0.5)
4: 100	Correct	Wrong	Wrong	-6.2(1.2)
5: 101	Correct	Wrong	Correct	-4.3(0.6)
6: 110	Correct	Correct	wrong	
7: 111	Correct	Correct	Correct	1.5 (0.2)

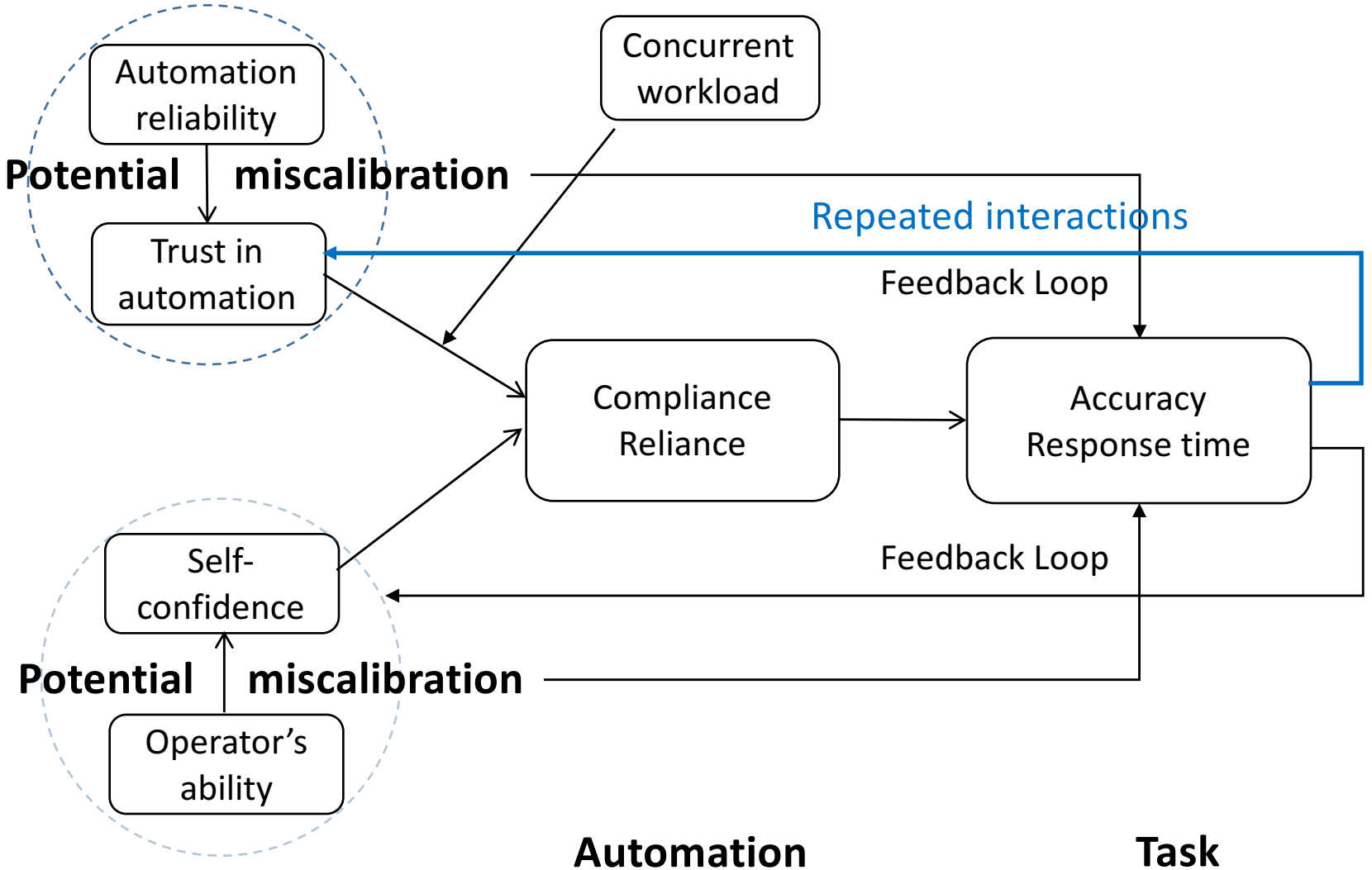
Automation failures lead to **less** decrement of trust, if the outcome is not harmed.

*** $p < .001$

Study 2| Results

- |Trust decrement| > |Trust increment|
- Trust assessment is not entirely rational
  - Not benchmarked strictly against predetermined objective criteria
  - Contrast effect - based on one's ability: a correct recommendation is appreciated more if one cannot perform the task
  - Hindsight bias - based on task outcomes: a wrong recommendation is "forgiven" if it does not harm

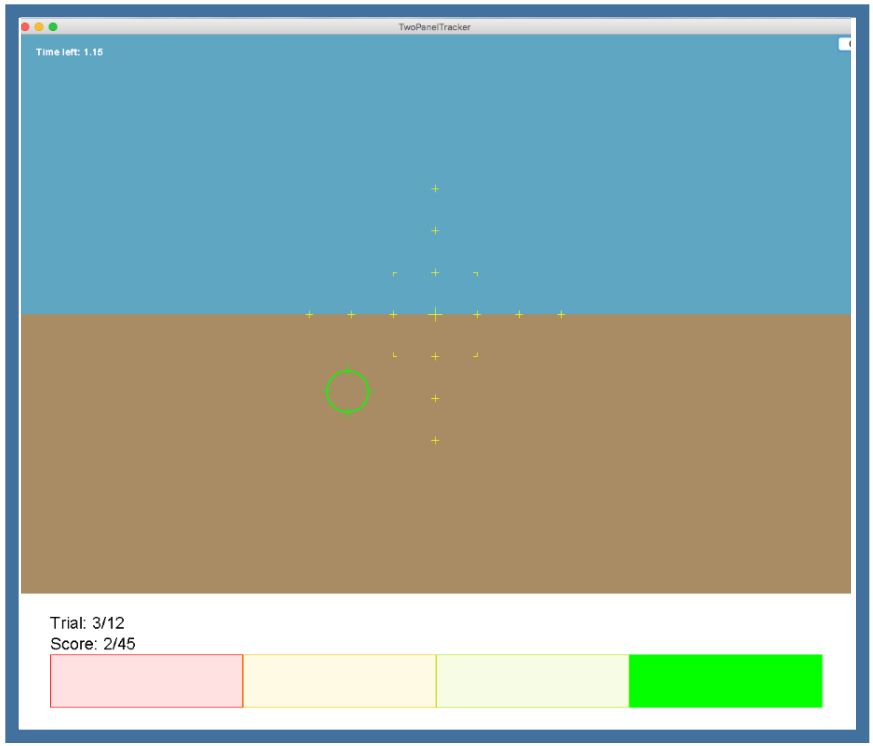
Automation dependence | Study 3

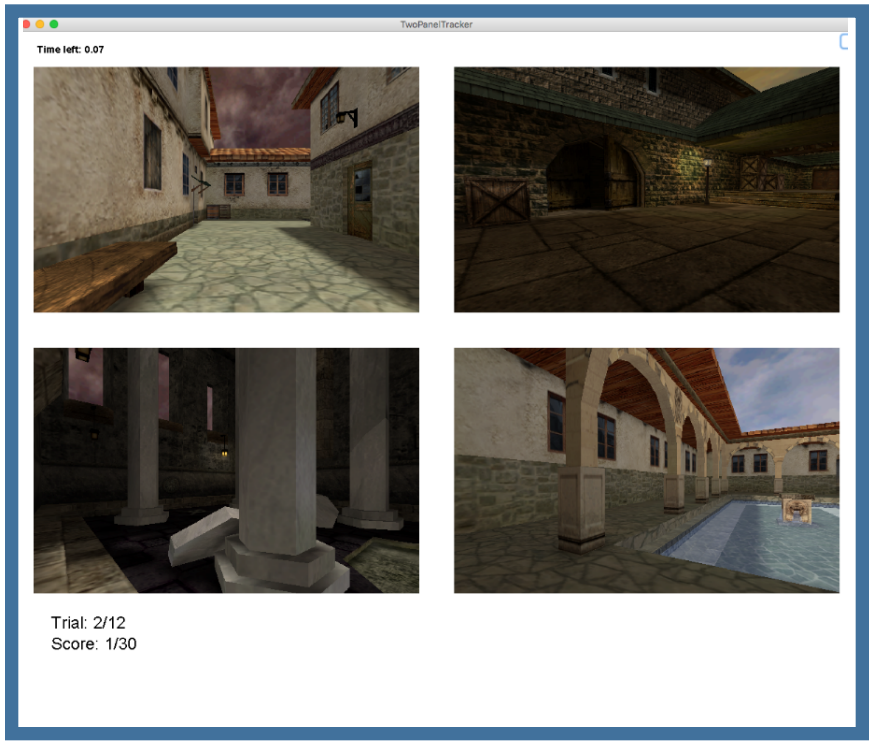


Study 3| Experiment setting

Evaluating Effects of User Experience and System Transparency on Trust in Automation
Thursday 1:30 pm

TOGGLE



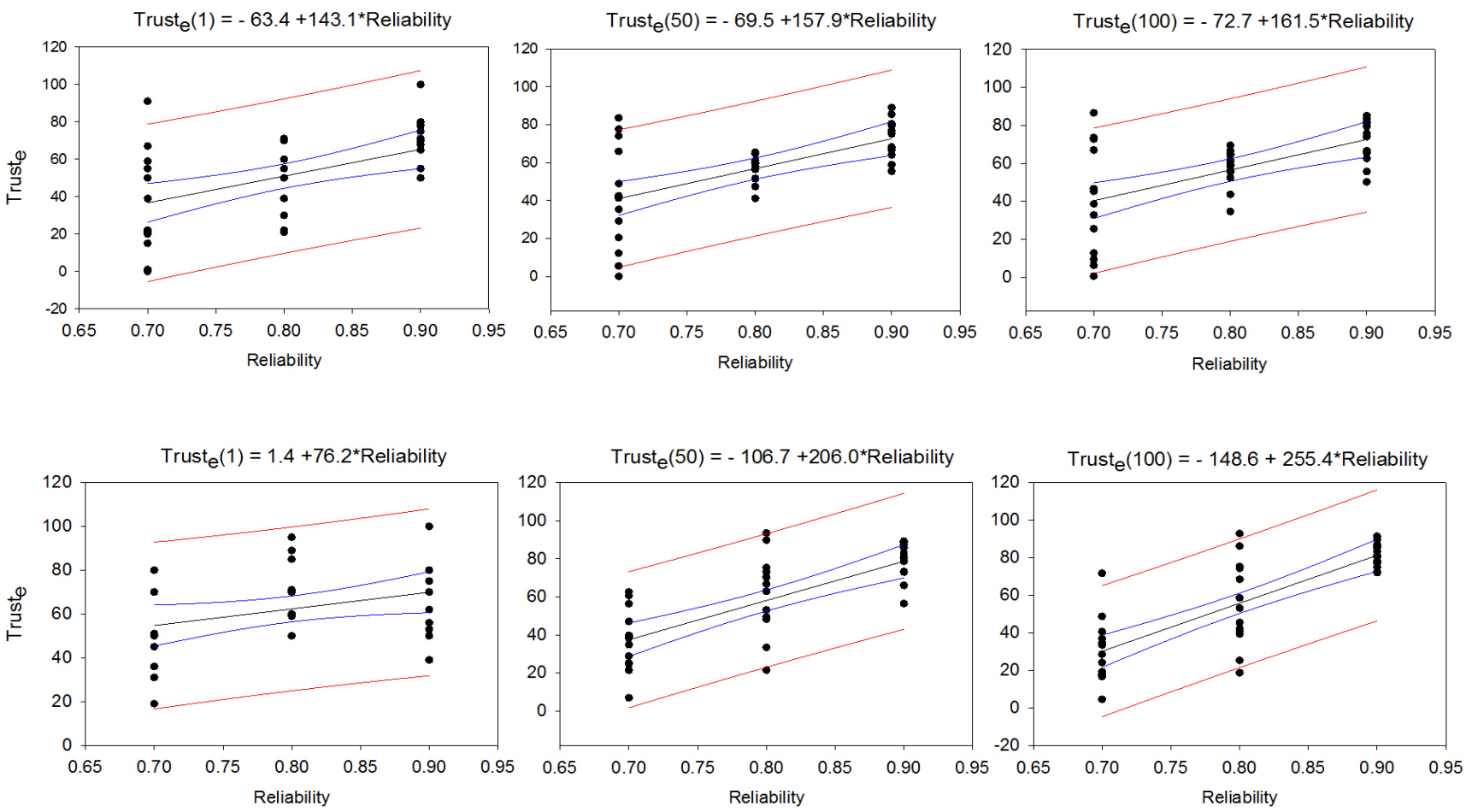


Tracking Task

Detection Task

Study 3| Results

	1st interaction	50th interaction	100th interaction
Binary Alarm	$Trust_e(1) = -63.4 + 143.1*Reliability$	$Trust_e(50) = -69.5 + 157.9*Reliability$	$Trust_e(100) = -72.7 + 161.5*Reliability$
Likeli-Hood Alarm	$Trust_e(1) = 1.4 + 76.2*Reliability$	$Trust_e(50) = -106.7 + 206.0*Reliability$	$Trust_e(100) = -148.6 + 255.4*Reliability$



- Over repeated interactions, users' trust becomes more appropriately calibrated
- With likelihood alarm, the calibration process is faster

Takeaways

- Appropriate calibration is the key
- Focus on the characteristics of the automation and of the human
- To improve human-automation team performance
  - To minimize trust-reliability miscalibration
  - To minimize confidence-ability miscalibration

Acknowledgement