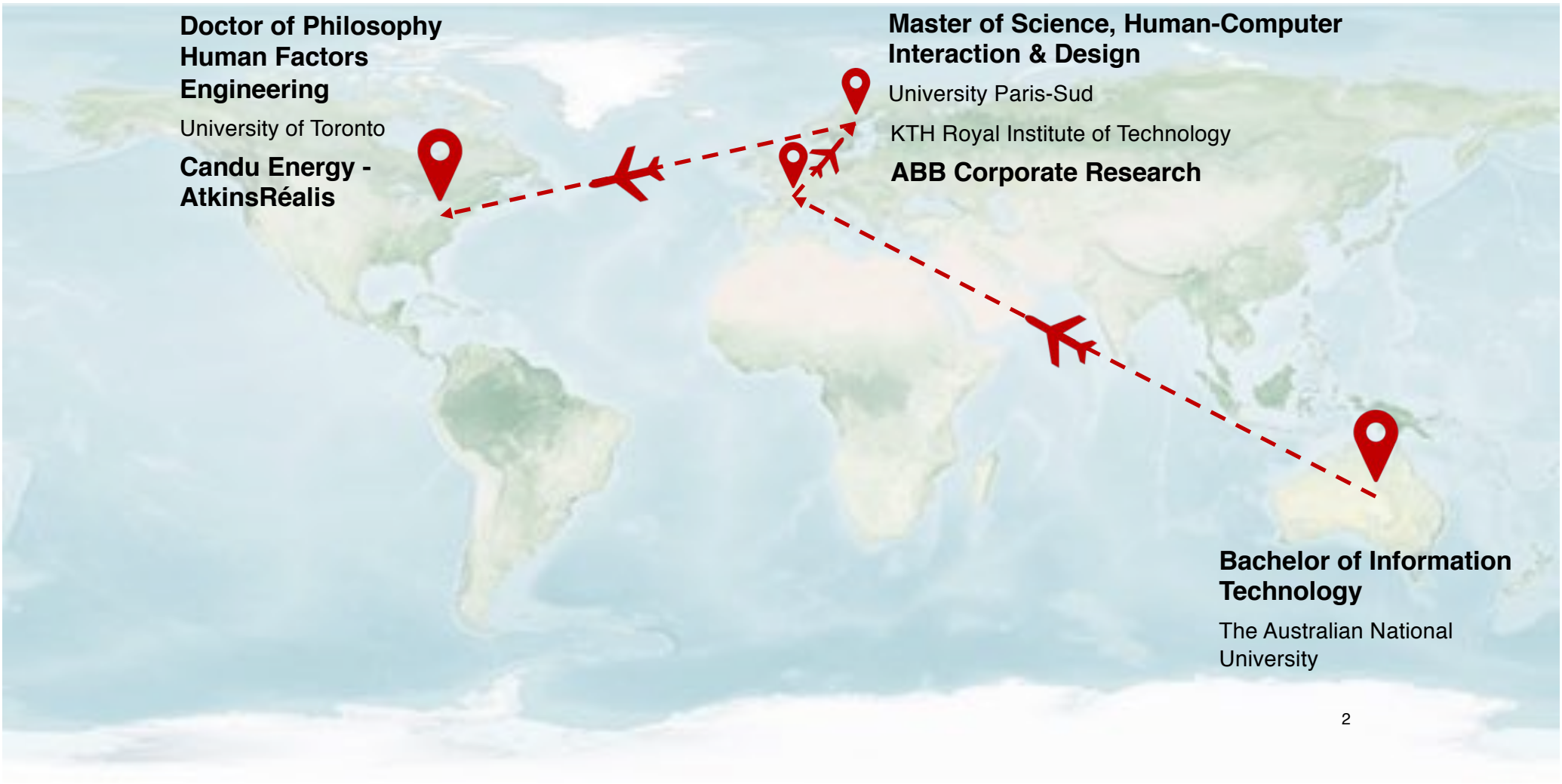
Mechanical & Industrial Engineering
UNIVERSITY OF TORONTO

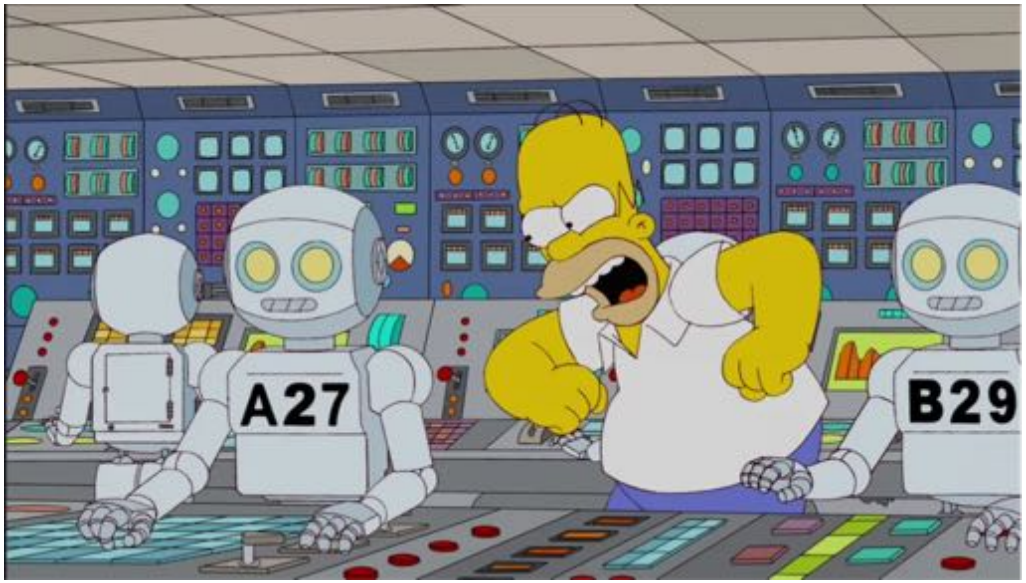
How Machine Learning Transparency Affects Decision Making

by Dr. Fahimeh Rajabiyazdi


Doctor of Philosophy
Human Factors
Engineering
University of Toronto
Candu Energy -
AtkinsRéalis
Master of Science, Human-Computer
Interaction & Design
University Paris-Sud
KTH Royal Institute of Technology
ABB Corporate Research
Bachelor of Information
Technology
The Australian National
University

The costs of over-automation

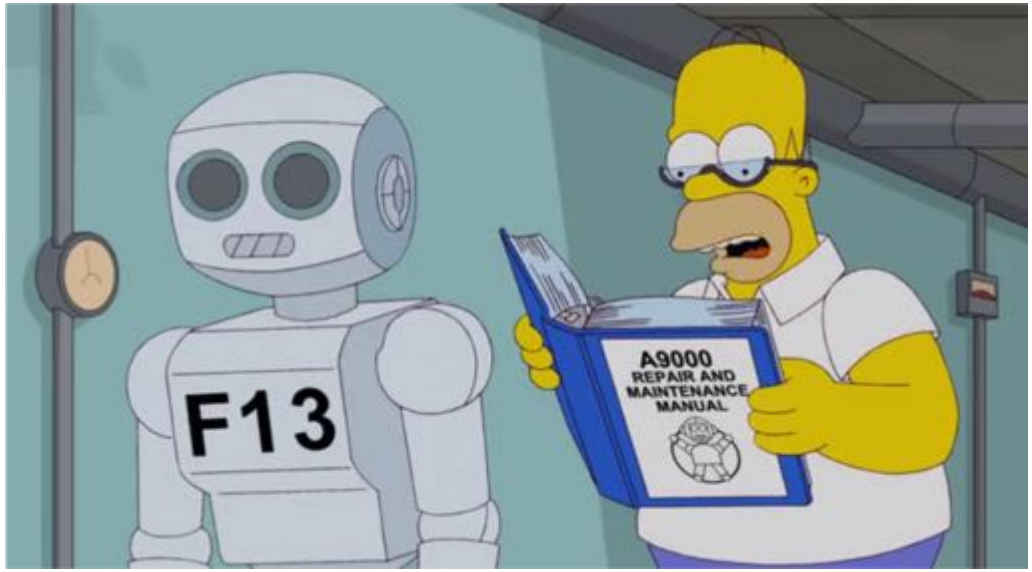
- Loss of Situation Awareness
- Out-of-the-loop unfamiliarity
- Mis-calibrated trust in automation
- Degraded understanding of automation
- Increased operational complexity
- New types of human-automation failures
- Automation bias & complacency
- Increased cognitive demand and sudden workload transitions
- Under-stimulation and loss of vigilance
- De-skilling



The Simpsons, S23, E17, Them, Robot

Automation Transparency

Facilitate human-automation-task interaction by overtly disclosing automation's otherwise hidden complexity through a technology medium.



The Simpsons, S23, E17, Them, Robot

Transparency Objectives

- Support understanding
- Calibrate and resolve trust
- Improve human-automation task performance
- Increase situation awareness

F. **Rajabiyazdi**, Jamieson, G. A.,Skraaning Jr. (2022) "Seeing-through and Seeing-into Automation Transparency: A Scoping Review," submitted to IEEE Transactions on Human-Machine Systems (under review).

F. **Rajabiyazdi** and G. A. Jamieson, "A Review of Transparency (seeing-into) Models," in 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020, pp. 302-308.

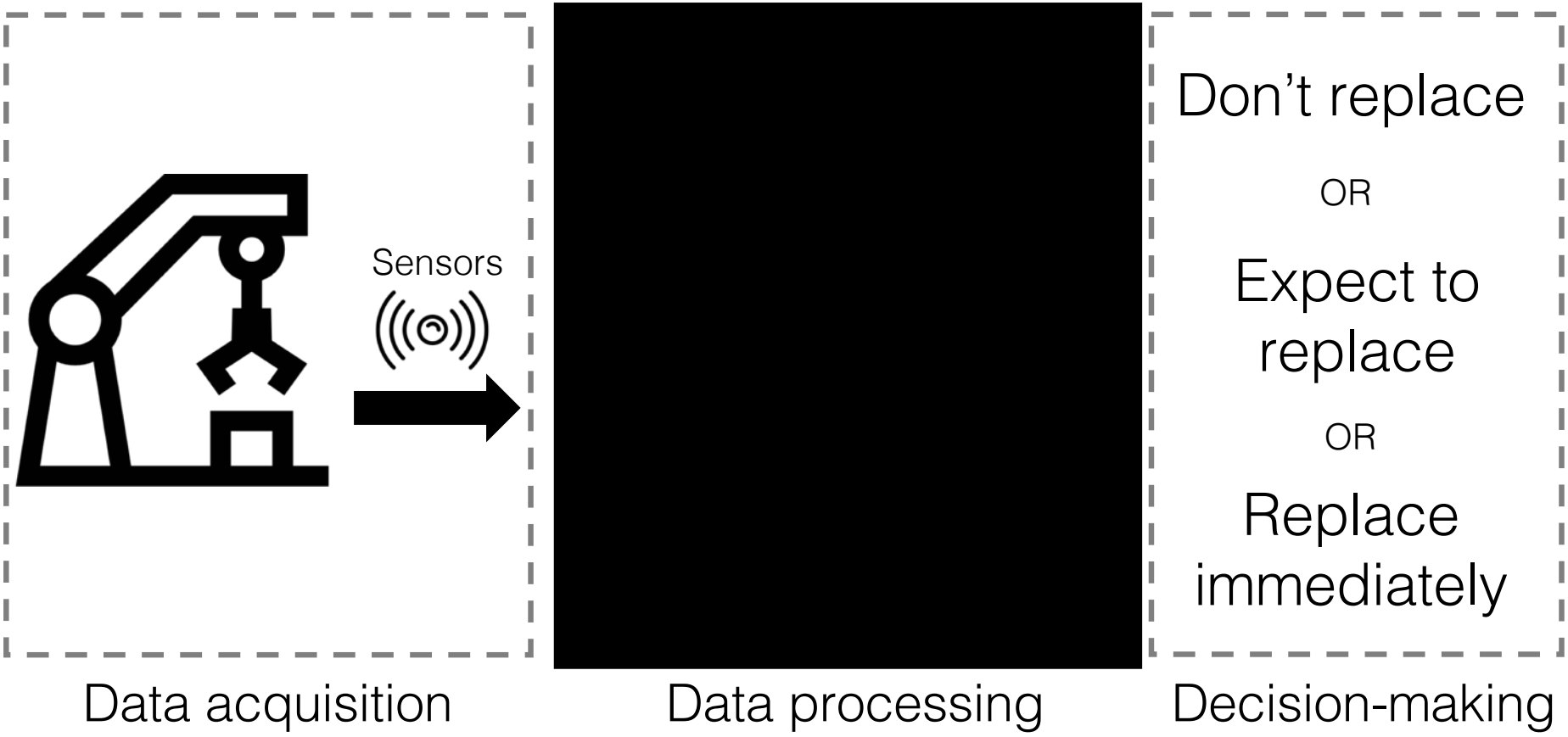
A. F. T. Winfield, S. Booth, L. A. Dennis, T. Egawa, H. Hastie, N. Jacobs, R. I. Muttram, J. I. Olszewska, F. **Rajabiyazdi**, A. Theodorou, M. A. Underwood, R. H. Wortham, and E. Watson, "IEEE P7001: A Proposed Standard on Transparency," Frontiers in Robotics and AI, vol. 8, p. 225, 2021.

Experiment

Motivation

1. Condition-based maintenance is the most prominent application of Artificial Intelligence and Machine Learning

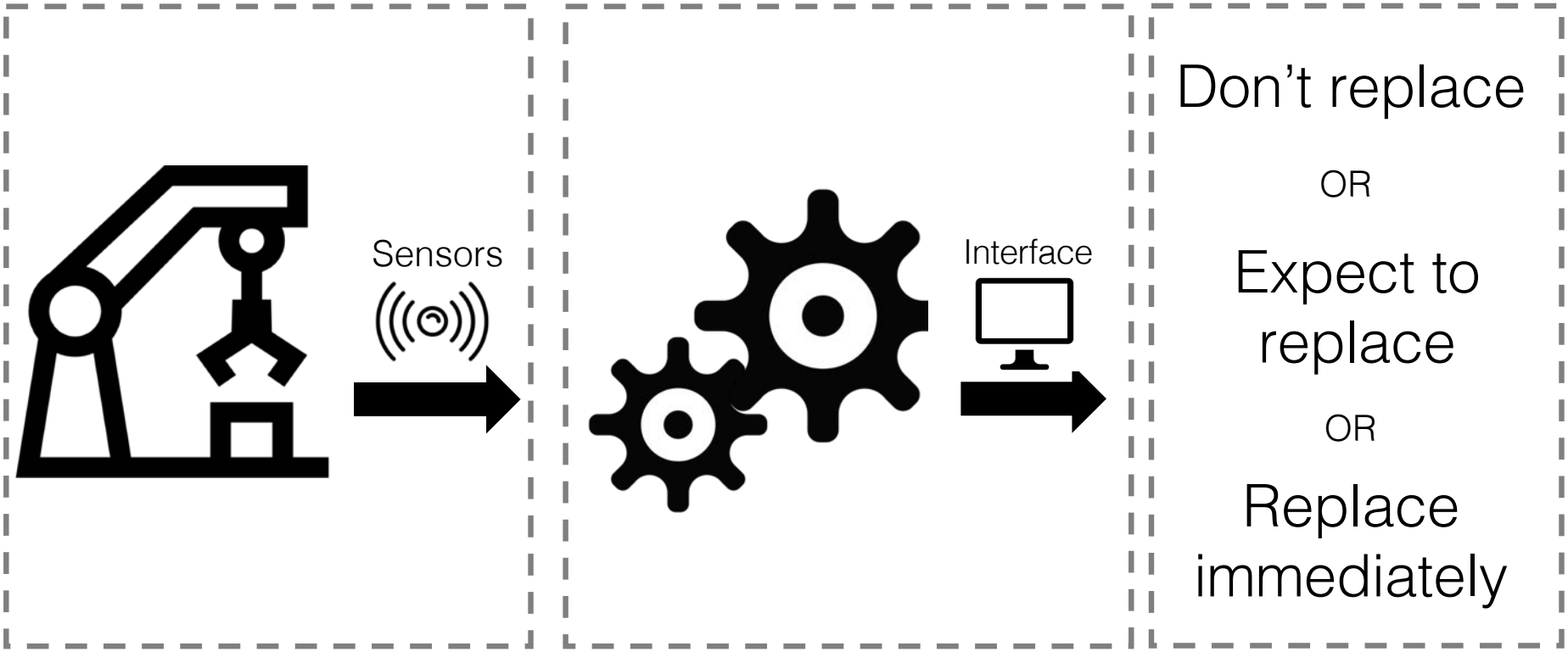
Condition-based Maintenance



Motivation

1. Condition-based maintenance is the most prominent application of AI
   - Human oversight may be required to check that the ML rationale aligns with end-user goals and metrics (Ribeiro, Singh, & Guestrin, 2016).
   - The end-user may need to verify that the training and validation data are representative of real-world conditions.

Condition-based Maintenance



Motivation

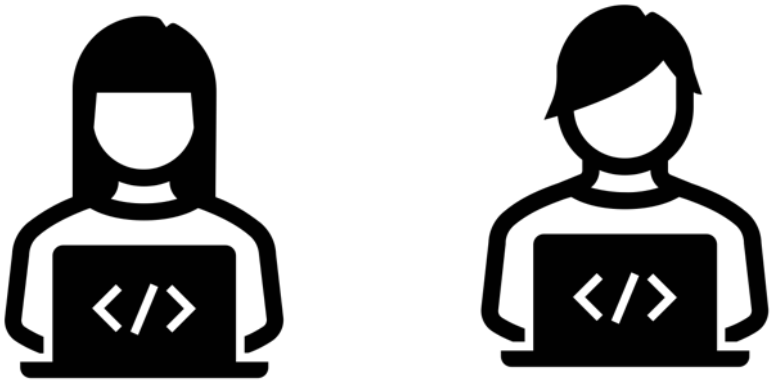
2. Inconsistent and, at times, conflicting results of automation transparency:
   - Positive impact on human task performance and trust calibration (Seong and Bisantz, 2008; Mercado et al. 2016).
   - Negative impact on human task performance but self-reported a better understanding of the ML-based rationale with greater information disclosure (Adhikari et al. 2019).
   - Participants performed worst but calibrated trust with information disclosure as automation capabilities increased (Skraaning and Jamieson, 2019).

Research Question

What are the effects of disclosing the rationale that led to an automated decision on human performance (including reliance decisions, trust, task efficacy, and workload)?

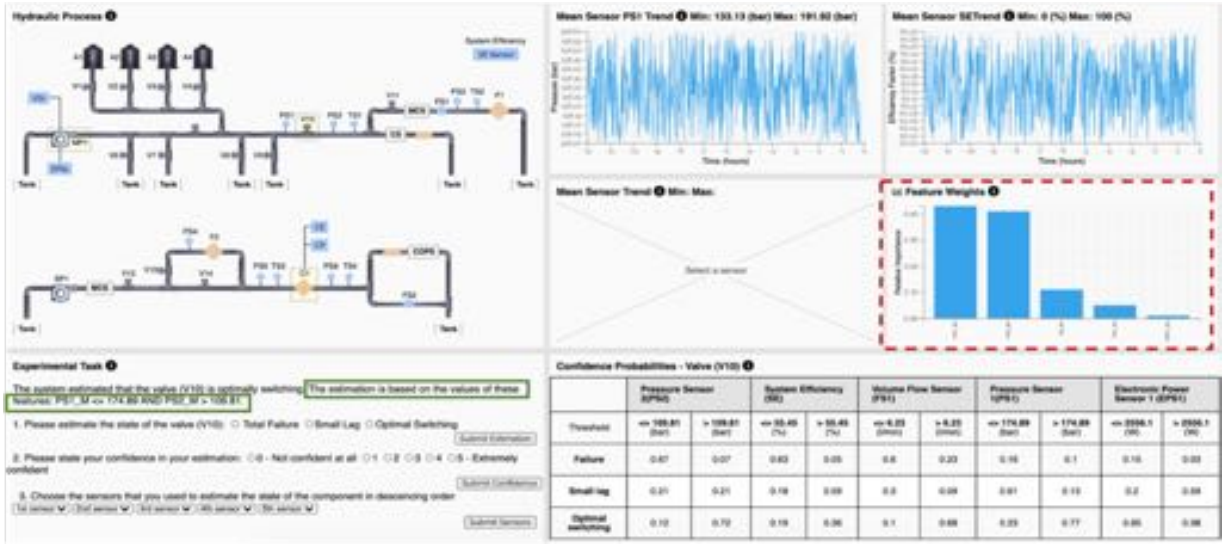
But first...

We need an apparatus



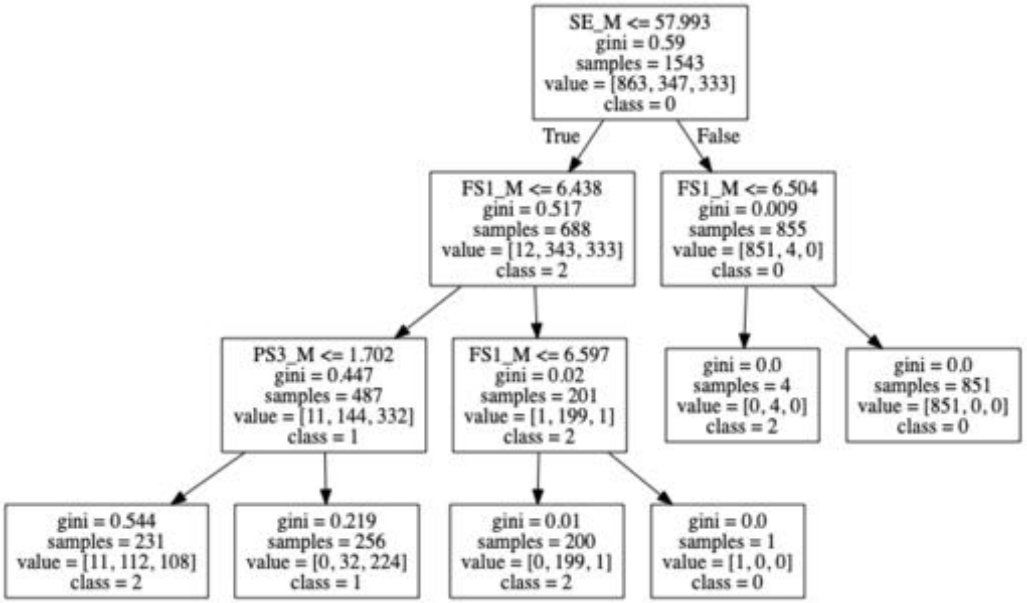
Apparatus

- Open source hydraulic system data
- ML predicting the condition of four hydraulic components: cooler, valve, pump, and accumulator

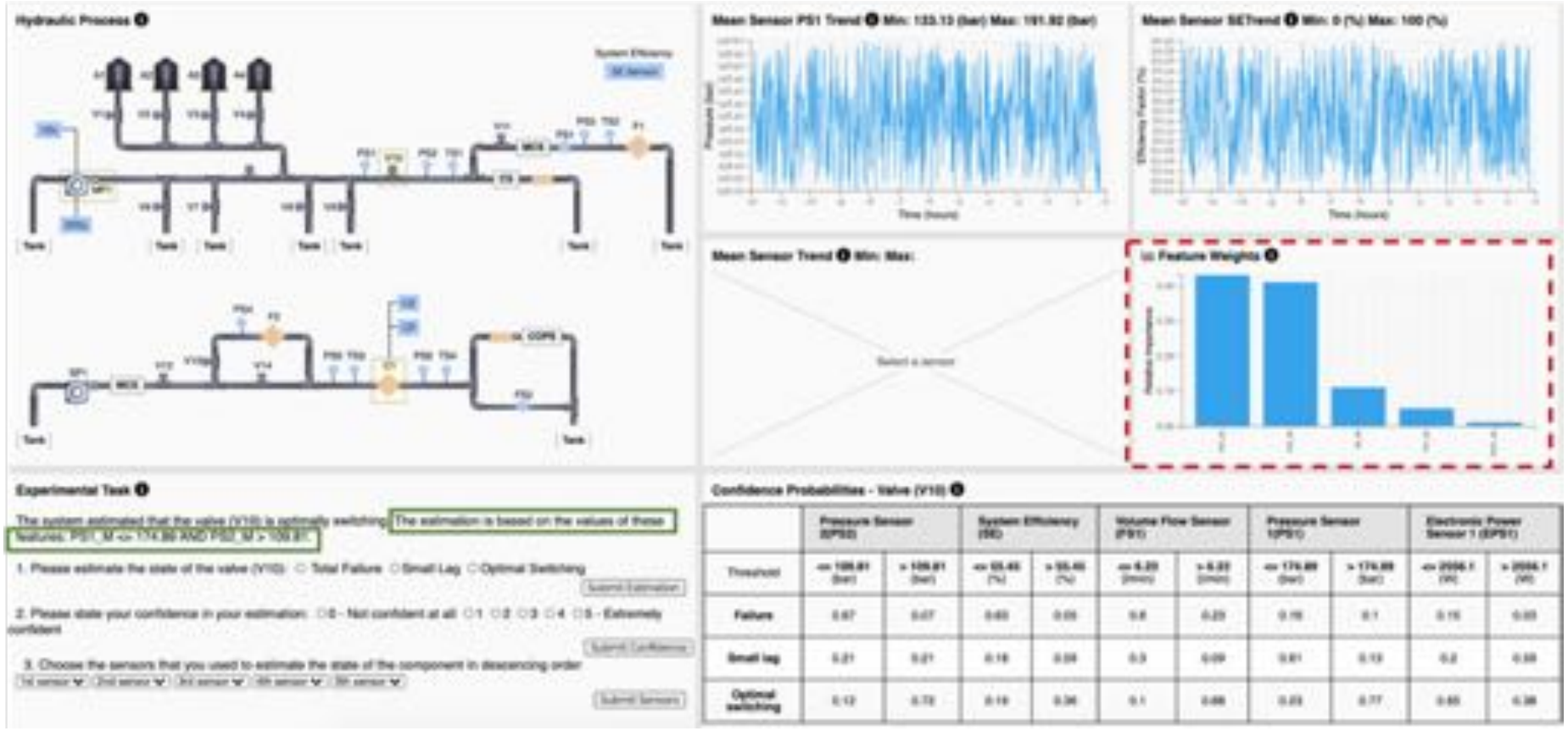


Machine Learning Model

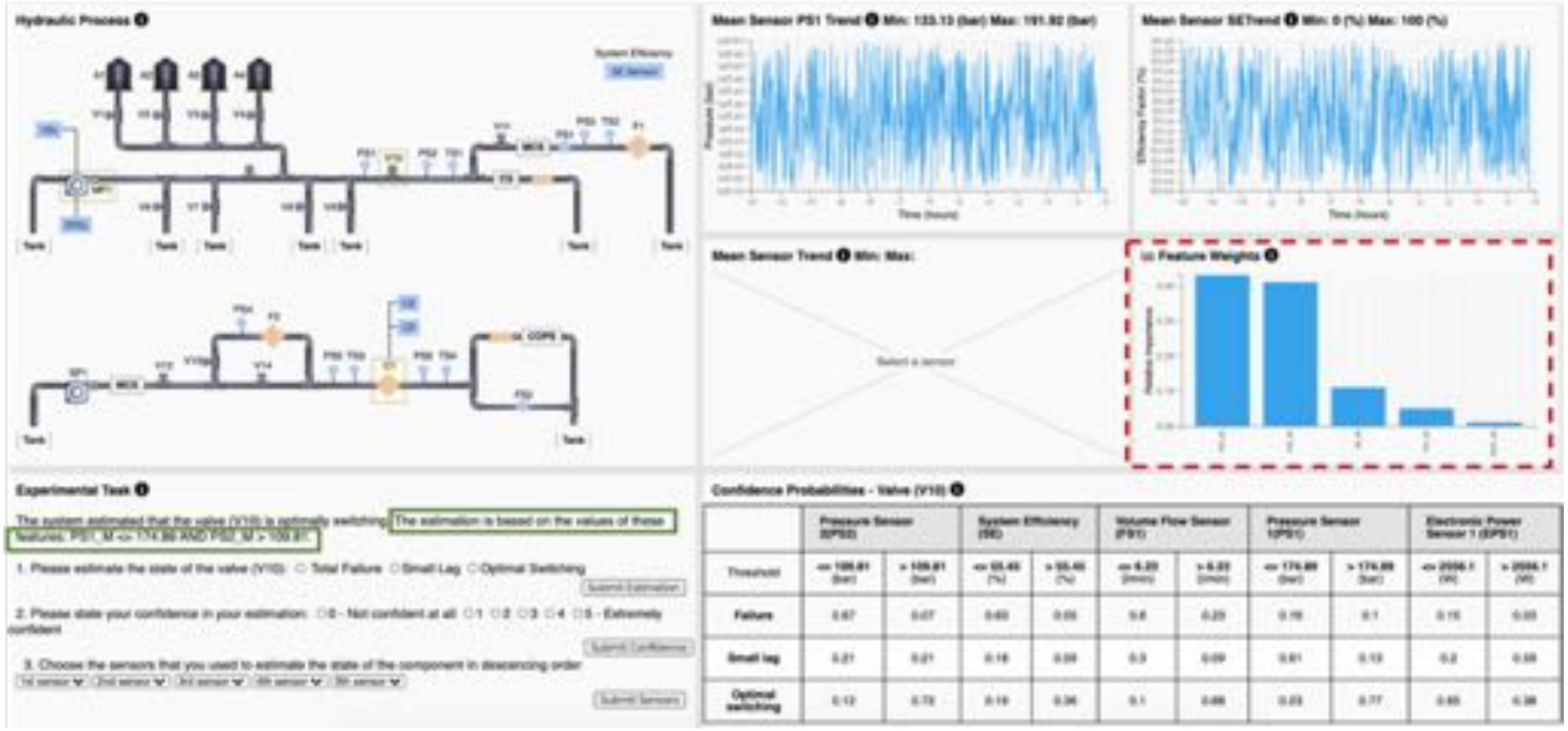
- Configured the number of trees, the depth of trees, the number of features to enhance comprehensibility.
- Integrated ML process with the domain knowledge of hydraulic system process.
- Determined the features and their thresholds.



Transparency Condition – Feature Weight



Transparency Condition – Decision Rules



	Pressure Sensor (EPS2)		System Efficiency (SE)		Volume Flow Sensor (FS1)		Pressure Sensor 1(PS1)		Electronic Power Sensor 1 (EPS1)	
Threshold	<= 109.81 (bar)	> 109.81 (bar)	<= 55.45 (%)	> 55.45 (%)	<= 6.23 (l/min)	> 6.23 (l/min)	<= 174.89 (bar)	> 174.89 (bar)	<= 2056.1 (W)	> 2056.1 (W)
Failure	0.87	0.07	0.83	0.03	0.8	0.23	0.16	0.1	0.15	0.03
Small lag	0.21	0.21	0.16	0.04	0.3	0.09	0.81	0.13	0.2	0.08
Optimal switching	0.12	0.72	0.19	0.36	0.1	0.68	0.03	0.77	0.65	0.38

Experiment design

Transparency conditions

1. Local Feature Weight Graph
2. Decision Rules
3. Combined (Local Feature Weight Graph + Decision Rules)

Within-subject design (randomized and counterbalanced) with 24 (14 female, 10 male) chemical engineering undergraduate and graduate students.

Experimental Task: participants estimated the state of a hydraulic component given three possible states.

Dependent measures: reliance decisions, trust, task efficacy, and workload

Contributions

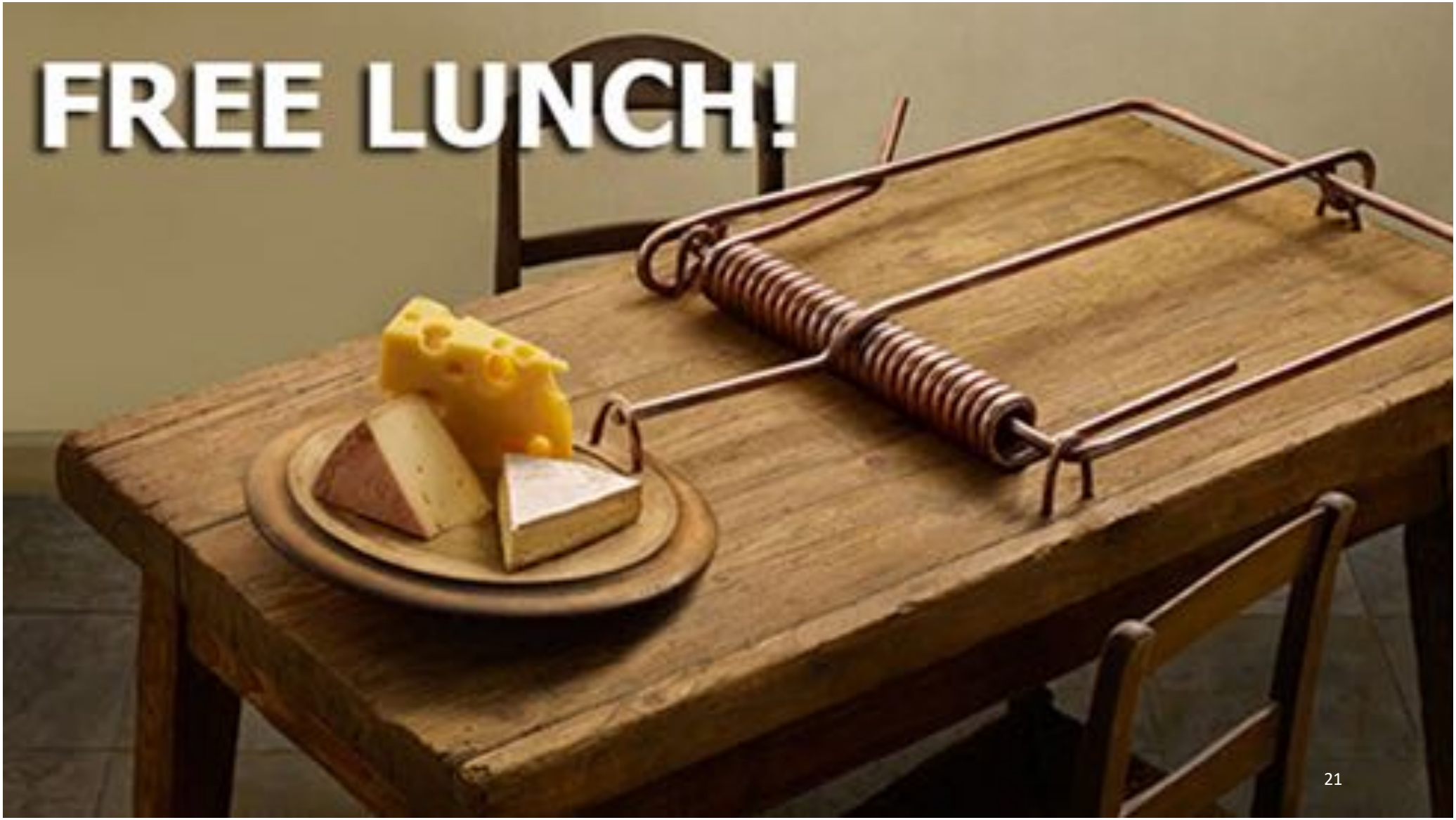
1. No evidence to corroborate the common belief that presenting a rationale for a decision aid's conclusion will positively impact any of the dependent measures.

2. Co-created a micro-world platform that has been used successfully ever since to conduct Explainable AI experiments.

F. **Rajabiyazdi**, G. A. Jamieson, and D. Quispe, "An Empirical Study on Automation Transparency (i.e., seeing-into) of an Automated Decision Aid System for Condition-Based Maintenance," in Proceedings of the 21st Congress of the International Ergonomics Association (IEA 2021), Cham, N. L. Black, W. P. Neumann, and I. Noy, Eds., 2022: Springer International Publishing, pp. 675-682

D. Quispe, F. **Rajabiyazdi** and G. A. Jamieson, "A Machine Learning-Based Micro-World Platform for Condition-Based Maintenance,"2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, 2020, pp. 288-295.

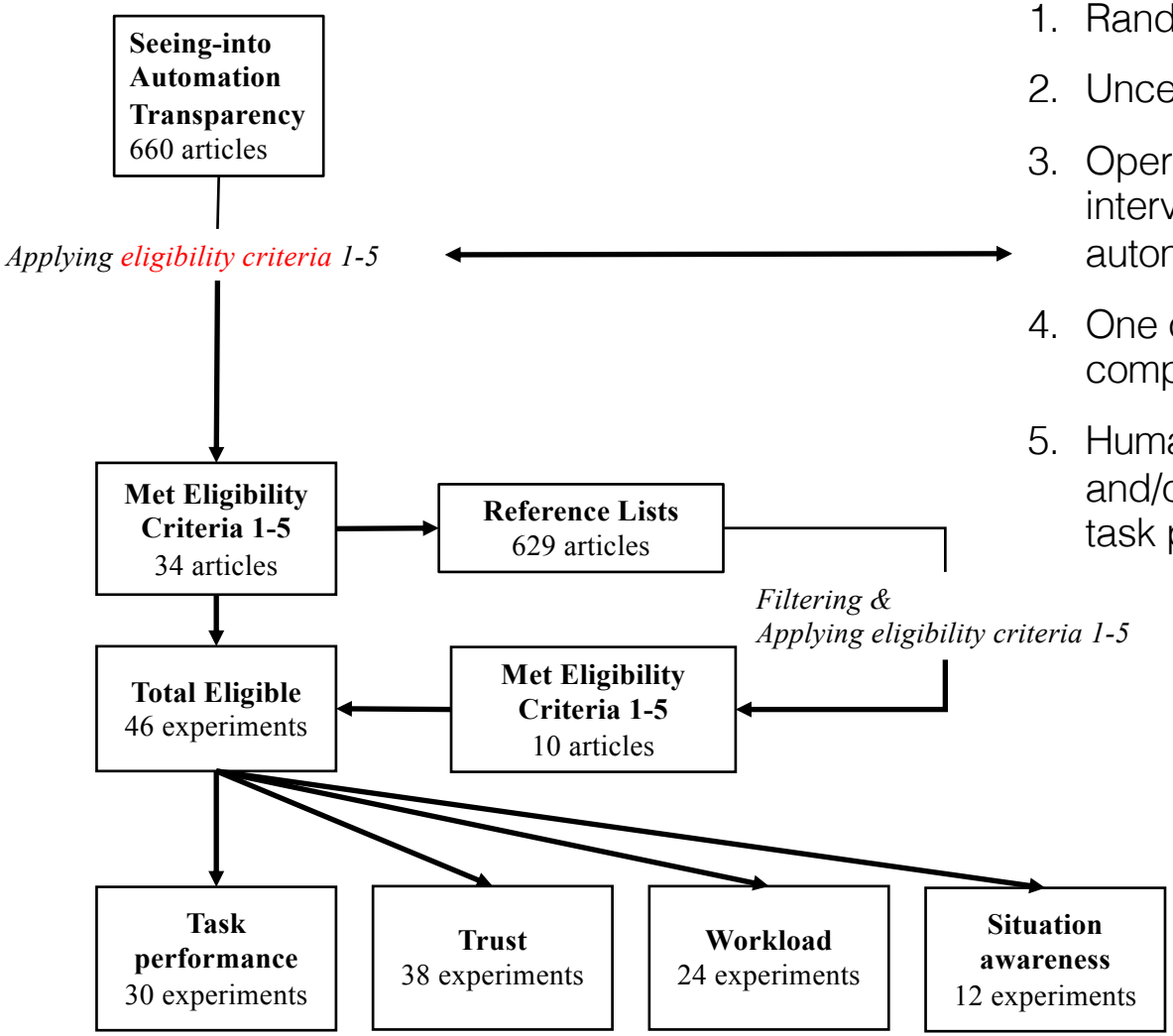
Meta-analysis


FREE LUNCH!

Background

	Bhaskara et al. (2020)	Van de Merwe et al. (2022)	Sargent, Walter, & Wickens (07/2023)	Our Study
No. studies	5 studies	17 studies	81 studies	46 studies
Basis of comparisons	SAT model	Type of tasks	Not specified	Logic model
Type of review	Narrative review	Narrative review	Statistical review	Systematic review & a meta-analysis
Outcomes	Summative claims about the effects of transparency on SA, trust, WL, and task performance.	Summative claims about the effects of transparency on SA, WL, and task performance.	Partial statistical claims about the effects of transparency on performance, WL, trust, SA.	Statistical claims about the effects of transparency on task performance.

Selection Process

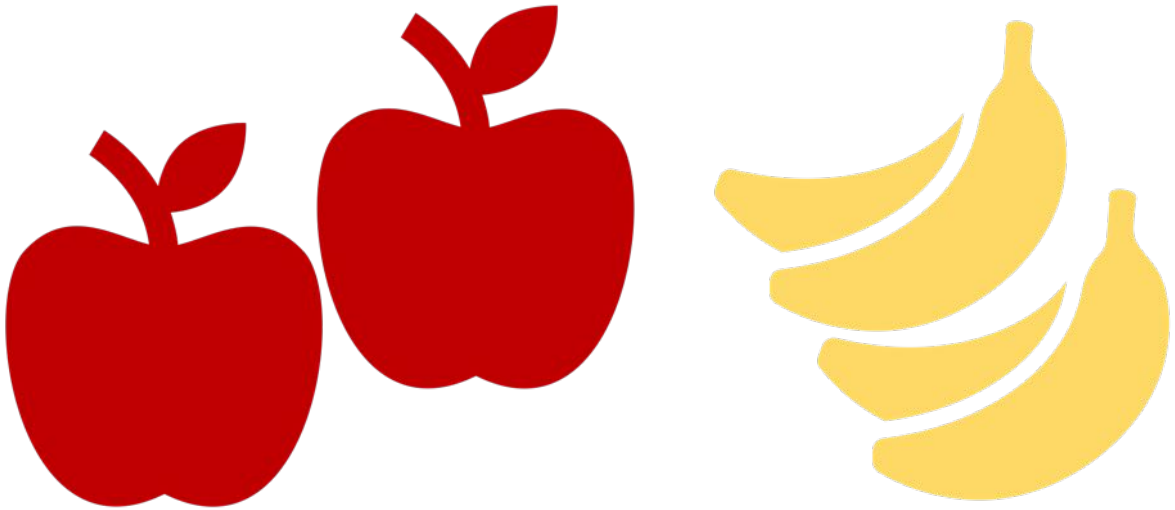


Eligibility criteria

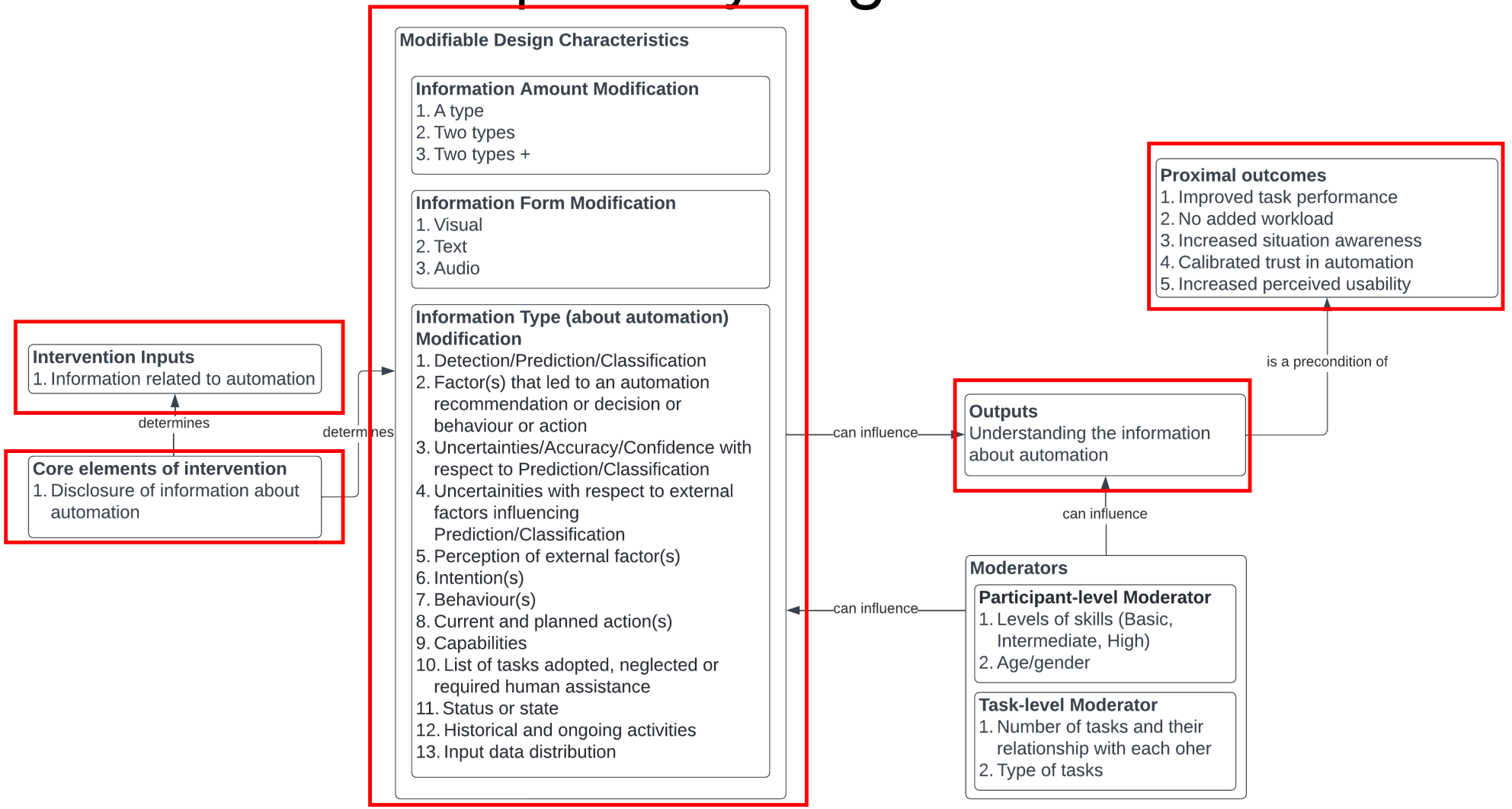
1. Randomized controlled trials
2. Uncertain and vulnerable situations
3. Operationalized automation transparency intervention as the disclosure of information about automation.
4. One or more of our pre-defined transparency comparisons.
5. Human performance measures, including trust and/or workload and/or situation awareness and/or task performance and/or reliance.

Transparency is a mess

Are we comparing apples to bananas?



Automation Transparency Logic Model



On what basis should we compare the eligible studies?

- Complexity Assessment Tool for Systematic Reviews (Lewin et al., 2016; Lewin et al., 2017).
- Assessed the intervention complexity on six core dimensions:

1. Active Components of Automation Transparency (AT)
2. Participants' Actions Targeted by AT (Experimental Tasks)
3. Organisational Levels Targeted by AT (not available in the literature)
4. Flexibility in AT Implementation (not available in the literature)
5. Experimenters' Skills in Delivering AT (not available in the literature)
6. Participants' Skills Targeted by AT

On what basis should we compare the eligible studies?

- Complexity Assessment Tool for Systematic Reviews (Lewin et al., 2016; Lewin et al., 2017).
- Assessed the intervention complexity on six core dimensions:

1. **Active Components of Automation Transparency (AT)**
2. Participants' Actions Targeted by AT (Experimental Tasks)
3. Organisational Levels Targeted by AT (not available in the literature)
4. Flexibility in AT Implementation (not available in the literature)
5. Experimenters' Skills in Delivering AT (not available in the literature)
6. Participants' Skills Targeted by AT

Active Components of Automation Transparency

Comparison 1

Disclosing **one** type of information about automation *vs*. not disclosing it (control)

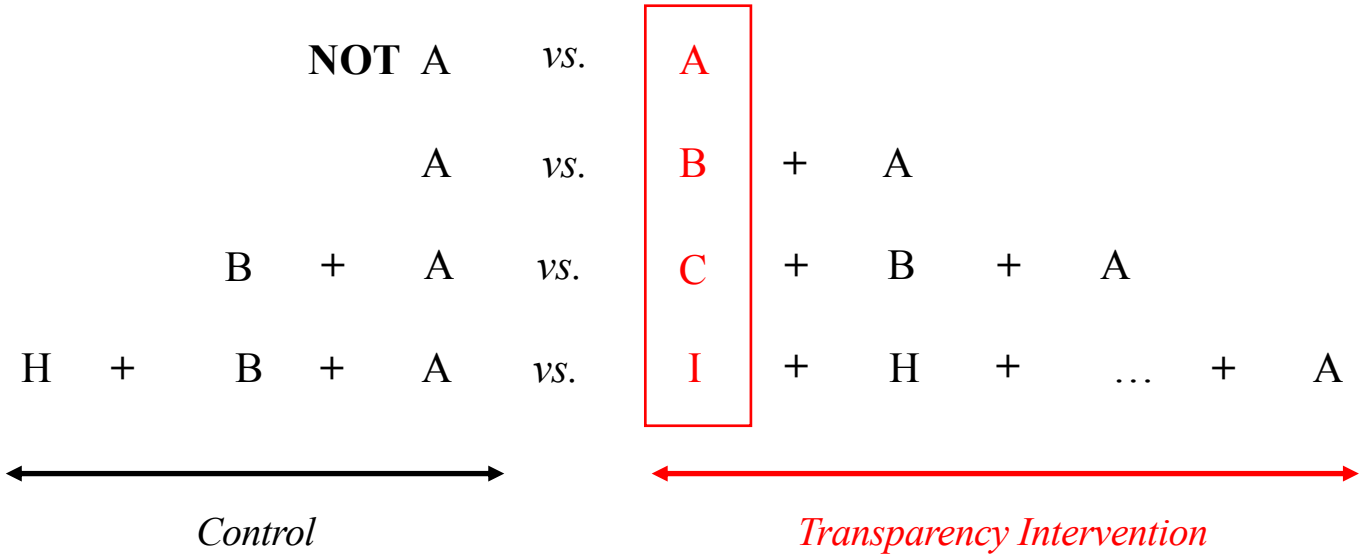
Comparison 2

Disclosing **two** types of information about automation *vs*. not disclosing them (control)

Comparison 3

Disclosing **more than two** types of information about automation *vs*. not disclosing them (control)

Comparison 1: disclosing **one** type of information about automation *vs.* not disclosing it



Note: Each letter represents a type of information about automation (e.g., automation status).

Effect Size Meta-Analysis

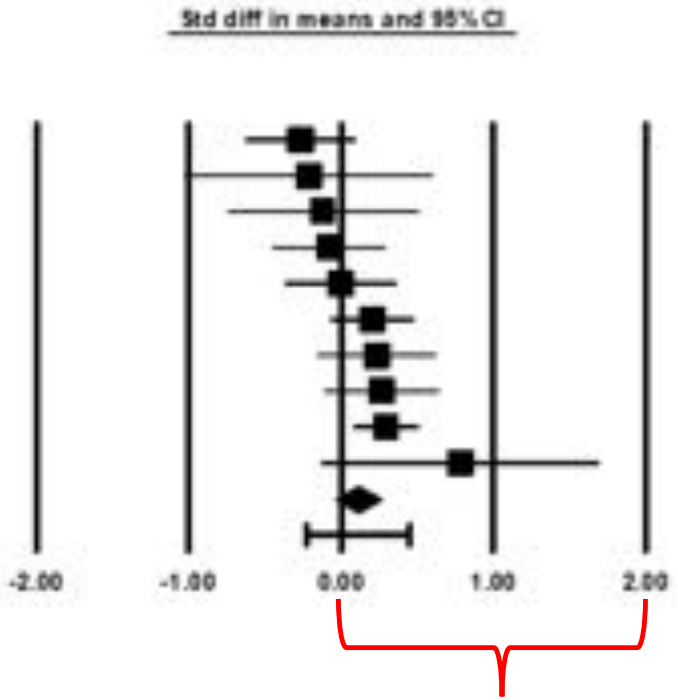
1. Compute effect size (Cohen's *d*) and variance for each study based on experimental design, sample size, & the F-test statistic.

2. Compute a weighted mean of these effect sizes under random-effects model.
    - Assumption of random-effect model: True effect size varies from study to study.

3. Compute the distribution of true effect using mean effect size, Tau-squared, number of studies, Confidence Interval.

Results

Study name	Std diff in means	Standard error	Variance	Upper limit	Lower limit	Z-Value	p-Value	Outcome
Nayyar et al. (2020)	-0.265	0.183	0.034	0.095	-0.624	-1.443	0.149	Response time
Heldin et al. (2020)	-0.212	0.415	0.172	0.602	-1.025	-0.510	0.610	Combined
Wright et al. (2016)	-0.119	0.318	0.101	0.503	-0.742	-0.375	0.707	Combined
Bahaskara et al. (2021)	-0.078	0.189	0.036	0.292	-0.447	-0.412	0.680	Combined
Loft et al. (2021)	-0.002	0.186	0.035	0.363	-0.367	-0.009	0.993	Combined
Stowers et al. (2020)	0.265	0.141	0.020	0.480	-0.071	1.456	0.146	Combined
Hussein et al. (2020)	0.233	0.197	0.039	0.619	-0.153	1.185	0.236	Combined
Mercado et al. (2016)	0.268	0.191	0.036	0.642	-0.107	1.402	0.161	Combined
Guznov et al. (2020)	0.295	0.109	0.012	0.509	0.082	2.712	0.007	Hit rate
Olatunji et al. (2020)EXP3	0.783	0.464	0.215	1.693	-0.126	1.688	0.091	Response time
Pooled	0.112	0.075	0.006	0.259	-0.035	1.498	0.134	
Prediction Interval	0.112			0.452	-0.227			



0.2 is small,
0.5 is moderate,
0.8 is a large effect size.

Where is transparency beneficial and where is transparency harmful?

- Complexity Assessment Tool for Systematic Reviews (Lewin et al., 2016; Lewin et al., 2017).
- Assessed the intervention complexity on six core dimensions:

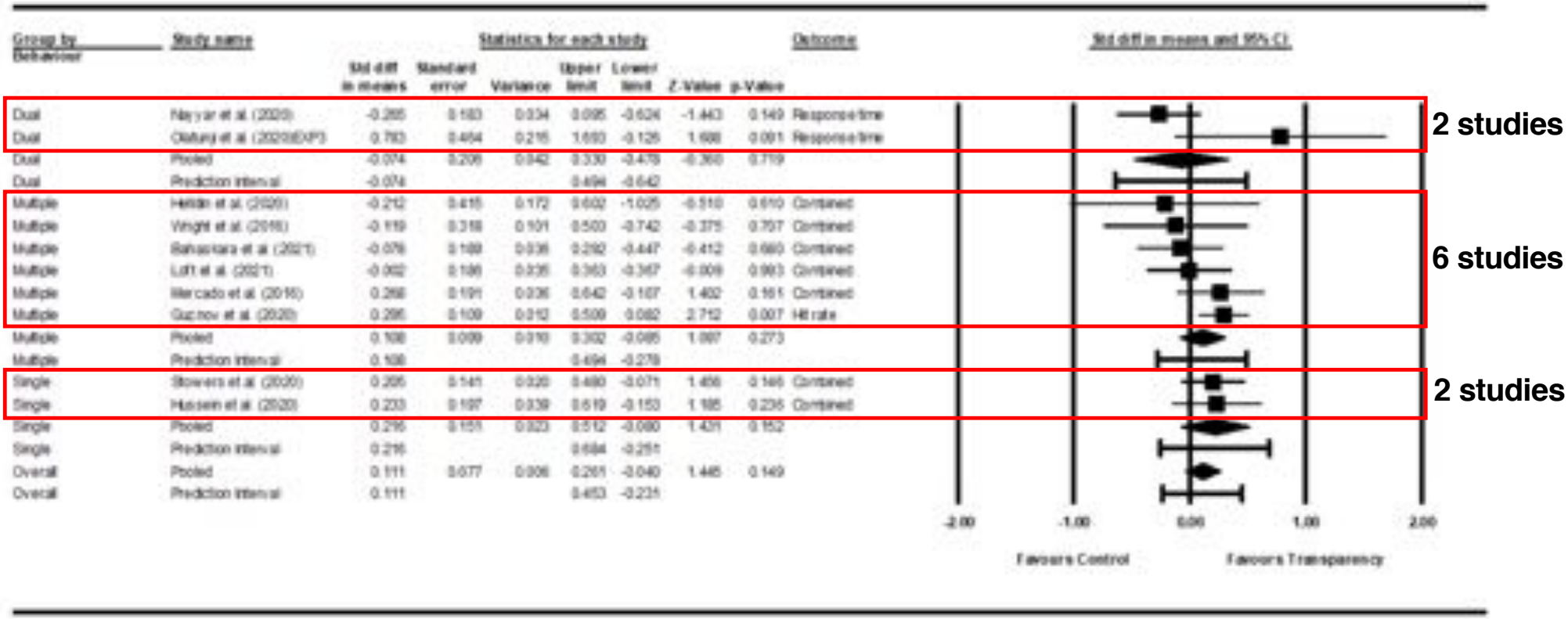
1. Active Components of Automation Transparency (AT)
2. Participants' Actions Targeted by AT (Experimental Tasks)
3. Organisational Levels Targeted by AT (not available in the literature)
4. Flexibility in AT Implementation (not available in the literature)
5. Experimenters' Skills in Delivering AT (not available in the literature)
6. Participants' Skills Targeted by AT

Where is transparency beneficial and where is transparency harmful?

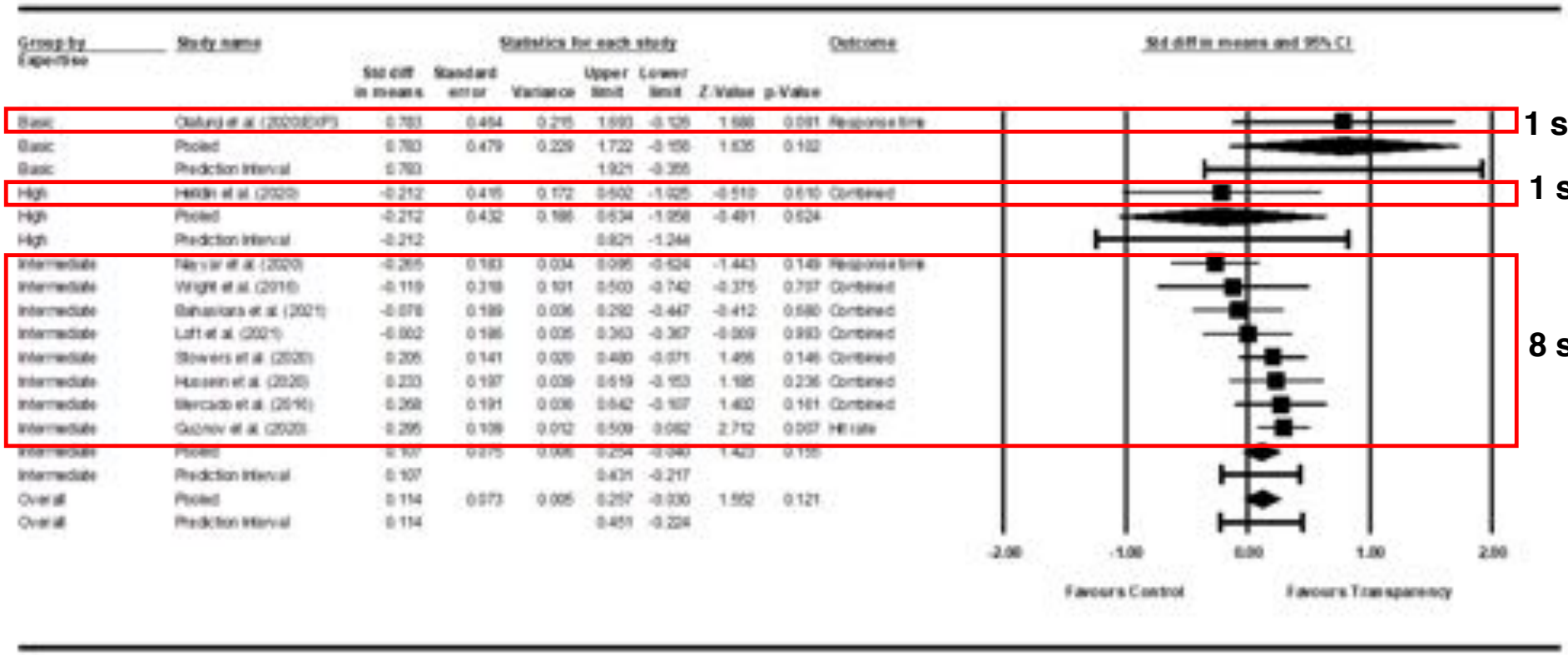
- Complexity Assessment Tool for Systematic Reviews (Lewin et al., 2016; Lewin et al., 2017).
- Assessed the intervention complexity on six core dimensions:

1. Active Components of Automation Transparency (AT)
2. **Participants' Actions Targeted by AT (Experimental Tasks)**
3. Organisational Levels Targeted by AT (not available in the literature)
4. Flexibility in AT Implementation (not available in the literature)
5. Experimenters' Skills in Delivering AT (not available in the literature)
6. **Participants' Skills Targeted by AT**

Participants' Actions Targeted by AT – Single, Dual, Multiple

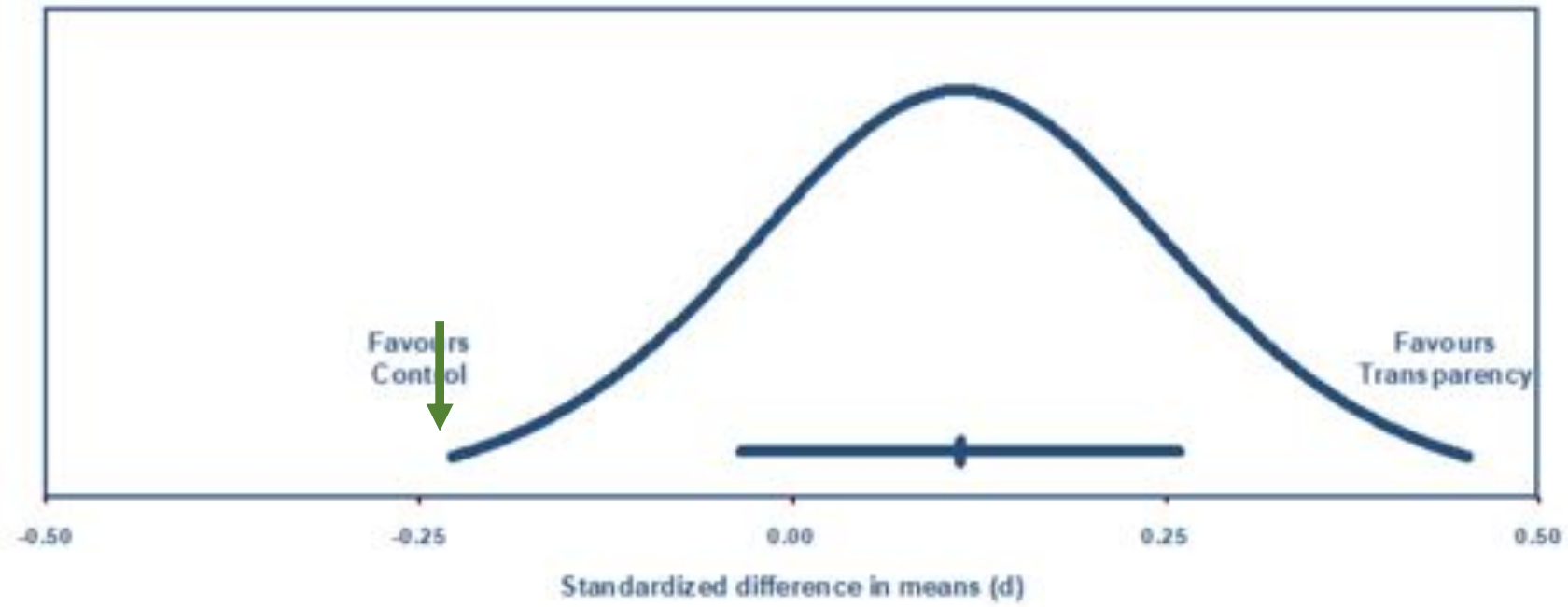


Participants' Skills Targeted by AT – Basic, Intermediate, High



What will the impact be if we implement automation transparency in a new system?

Prediction Interval



Limitations

1. Incomplete reporting, selective reporting, or not reporting data in a format that could be used in a meta-analysis.
2. Risk of publication bias and study quality assessment not yet conducted.
3. Studies that are not comparable to others.
4. Not yet conducted a meta-analysis on the effects of transparency on trust (38 studies), workload (20 studies), and situation awareness (SA) (12 studies).

Should automation be transparent?

Automation transparency is a design principle that is consistently presented in the literature as a means to improve human performance with automated systems.

In this dissertation, rigorous literature, empirical, and statistical examinations demonstrate **little evidence that automation transparency is a generalizable principle.**

Should automation be transparent?

- Need better analysis of existing evidence
  - Expand range of outcome variables
  - Fill in missing data
  - Assess risk of publication bias
  - Assess study quality
- To assemble more evidence
  - Adopt a logic model
  - Apply standards for reporting

Acknowledgment

Mentors from University of Toronto:

- Prof. Greg A. Jamieson
- Prof. Chris Beck
- Prof. Olivier St-Cyr

Special thanks to:

- David Quispe
- Dr. Gyrd Skraaning
- Dr. Sarah Simmons
- Nima Mirjalali