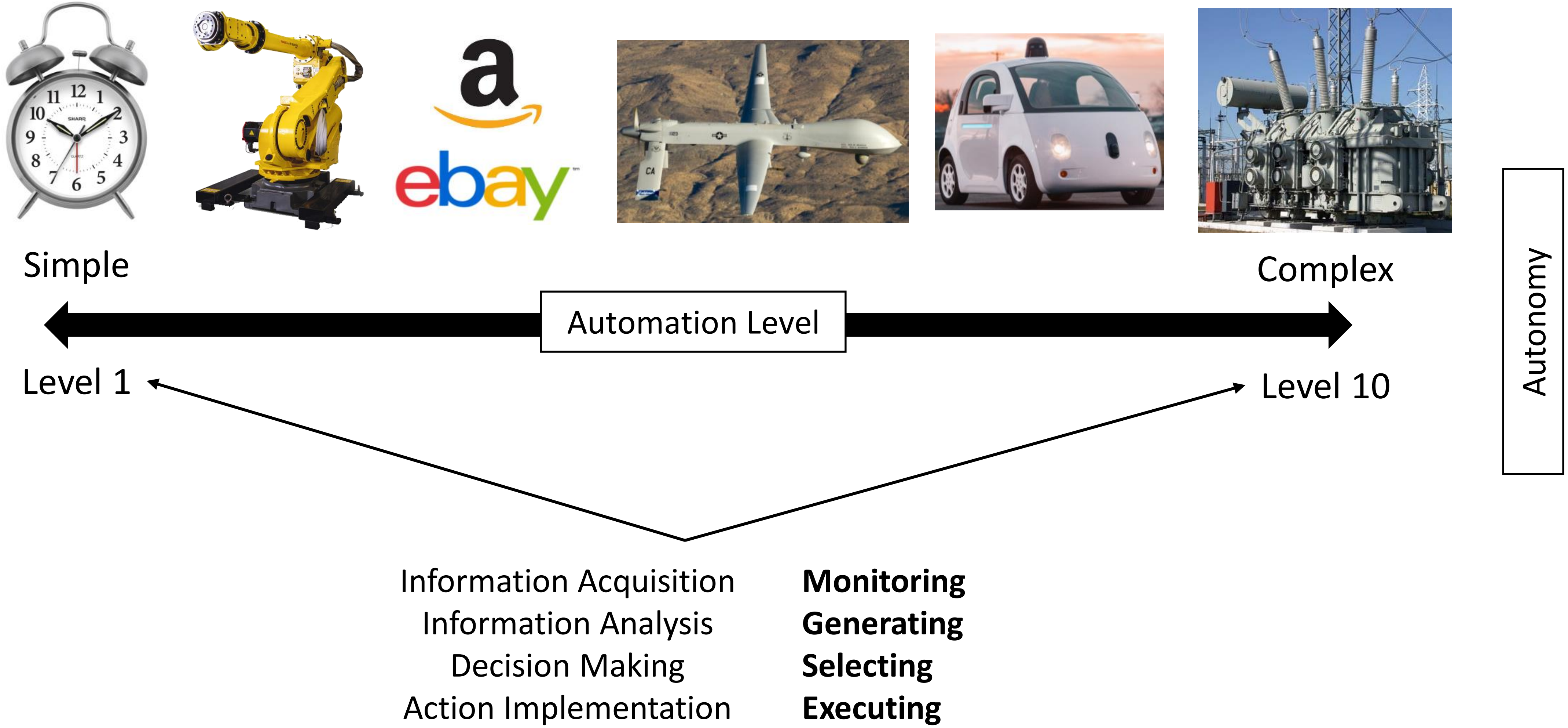
“Good Enough” Agents: Investigating Designed Imperfections in Human-AI Teams Across Parallel Domains

Sebastian S. Rodriguez

Doctoral Final Defense

UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN

Evolving automation



Simple

Complex

Automation Level

Level 1

Level 10

Autonomy

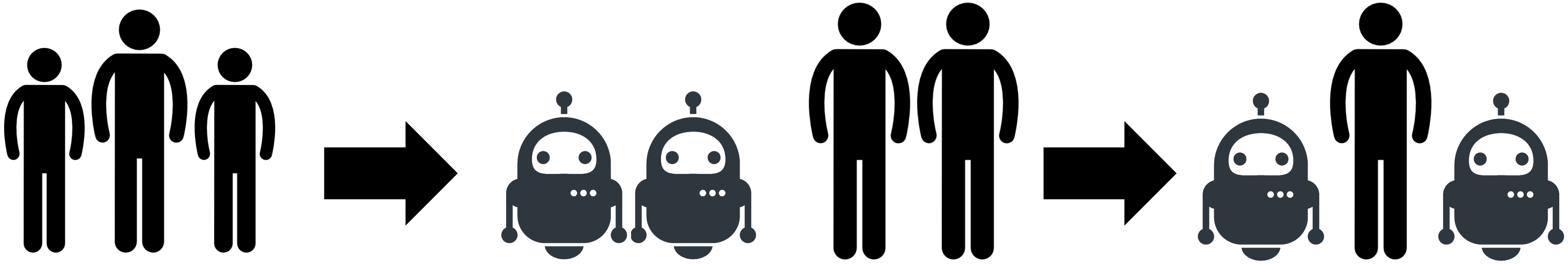
Information Acquisition **Monitoring**

Information Analysis **Generating**

Decision Making **Selecting**

Action Implementation **Executing**

Human-agent interaction



Human Teams

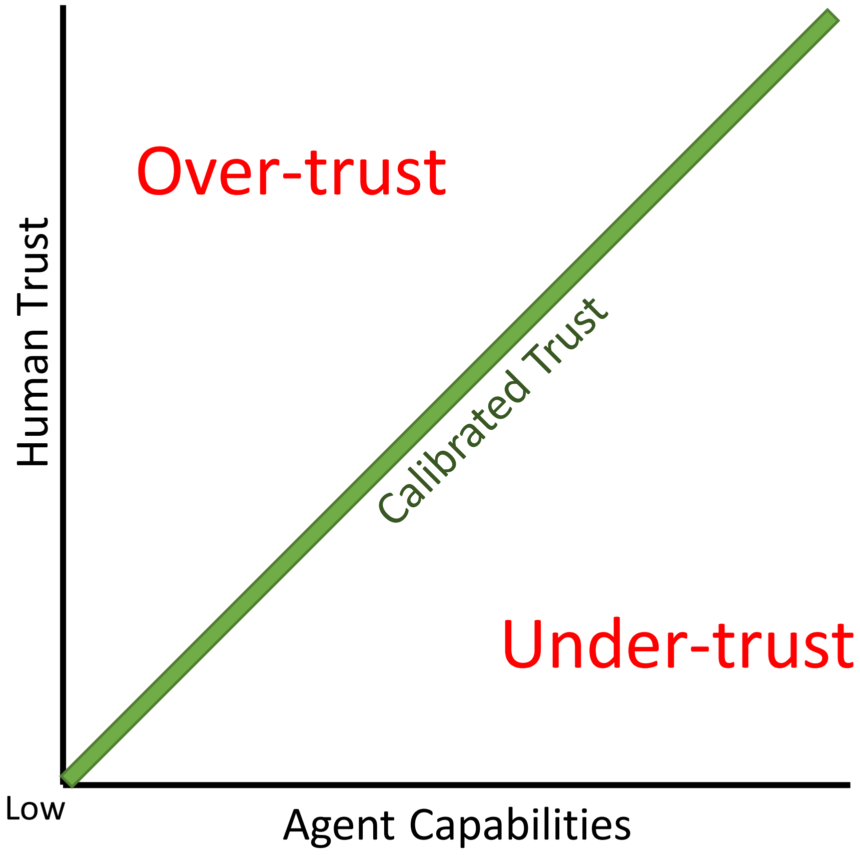
HABA-MABA Teams
Function Allocation

Human-Agent Teams
Interdependent Activity

Inter-predictable
Commonly Grounded
Directable

Inappropriate use: we’re bad at calibrating

- Trust is a facet researched often w.r.t automation
- Truth is, humans are ***bad*** at calibrating trust[a]
  - We expect too much
  - We expect too little
  - We can’t make up our minds
- Hitting the sweet spot is challenging: calibration like human-human trust[b]



a V. L. Pop, A. Shrewsbury, and F. T. Durso. 2015. Individual differences in the calibration of trust in automation. Human Factors 57, 4, Jun 2015.

b J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society 46, 1, Jan 2004.

Under-trust has been addressed

Explanations

Control settings

Information displays

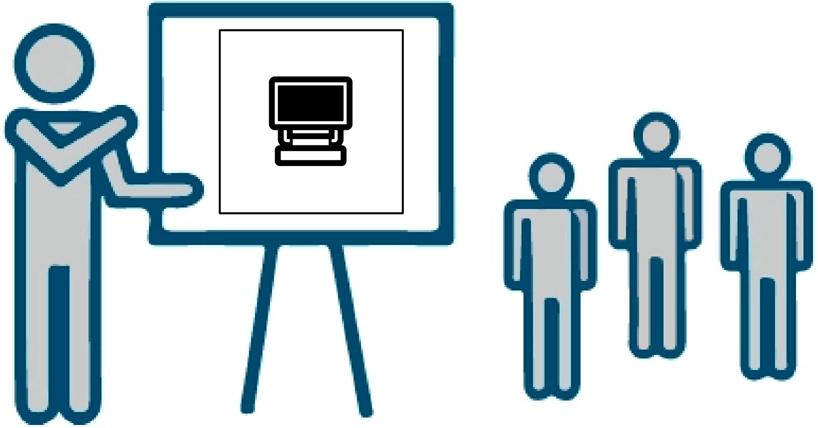
Anthropomorphism

UX design

Trust is shown to positively influence technology acceptance[a]
But could these features[b] generate excessive trust?

a K. Wu, Y. Zhao, Q. Zhu, X. Tan, H. Zheng. A meta-analysis of the impact of trust on technology acceptance mode: Investigation of moderating influence of subject and context type. International Journal of Information Management 31, 6, Dec 2011.
b H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, J. Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. CHI 2020 Proceedings. 2020.

Over-trust is being addressed



Viable approaches: training[a] and repeated exposure[b] to manual control

Requires time and effort to learn and become skilled at calibrated trust

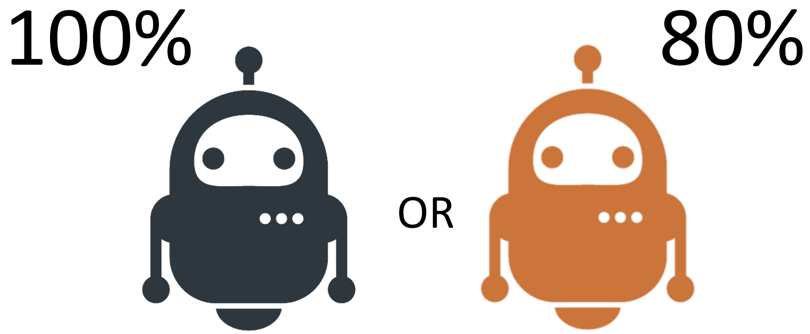
Any other approaches?

a J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society 46, 1, Jan 2004.

b M. Itoh. Necessity of supporting situation understanding to prevent over-trust in automation. International Electronic Journal of Nuclear Safety and Simulation 43, 46, 2010.

“Good Enough”-ness

- We expect AI to be optimal[a]
- What if it doesn’t need to be?
- Reliability is often a feature not manipulated as part of agent design
  - Research states AI reliability should be >70% for any benefit[b]
- At which reliability is the most benefit found?
- Design goal or minimum returns?



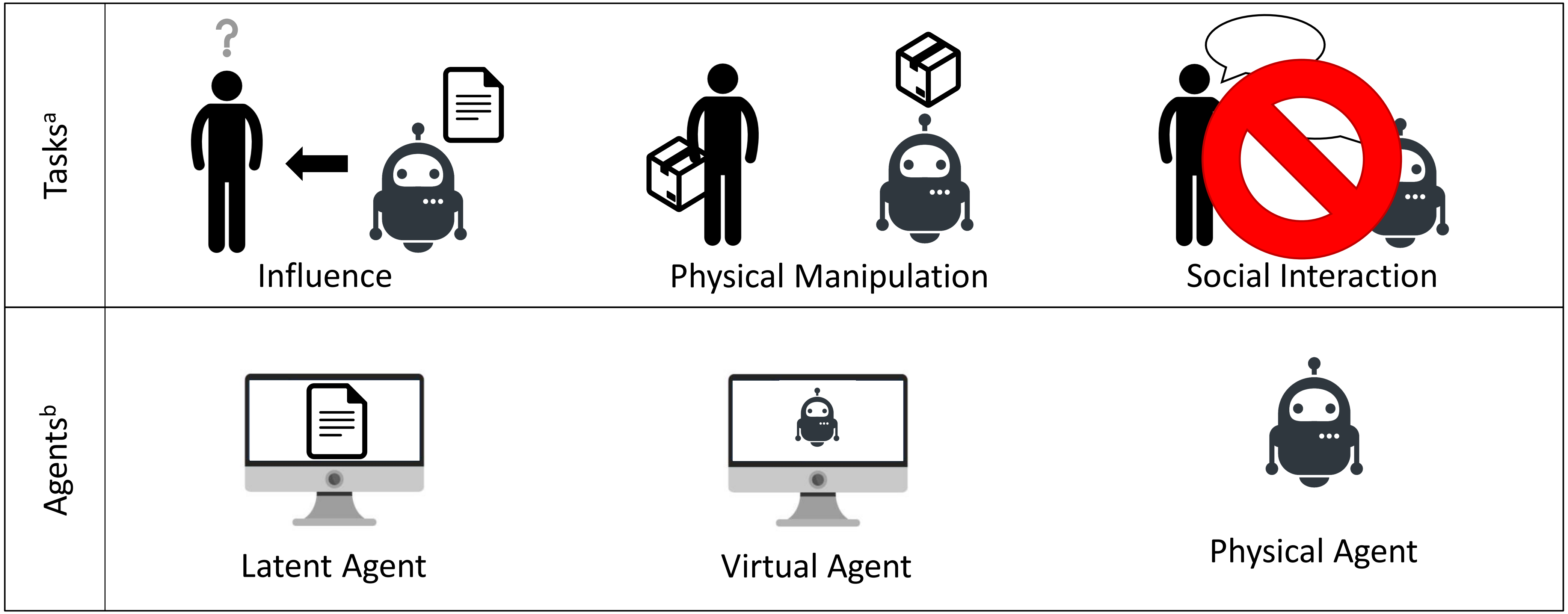
“Good Enough” Agent

An agent engrained with small error, to simulate less than perfect performance

[a] M. Endsley. From Here to Autonomy: Lessons Learned From Human-Automation Research. Human Factors, 59, 1. Feb 2017.

[b] B. Schelble, C. Flathmann, N. McNeese. Towards Meaningfully Integrating Human-Autonomy Teaming in Applied Settings. International Conference on Human-Agent Interaction, Nov 2020.

Domains of operation



Tasks[a]	Influence	Physical Manipulation	Social Interaction
Agents[b]	Latent Agent	Virtual Agent	Physical Agent

[a] C. Esterwood, K. Essenmacher, H. Yang, F. Zeng, L. Robert. A Meta-Analysis of Human Personality and Robot Acceptance in Human-Robot Interaction. Conference on Human Factors in Computing Systems. 2021.

[b] J. Li. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. International Journal of Human-Computer Studies. 2014.

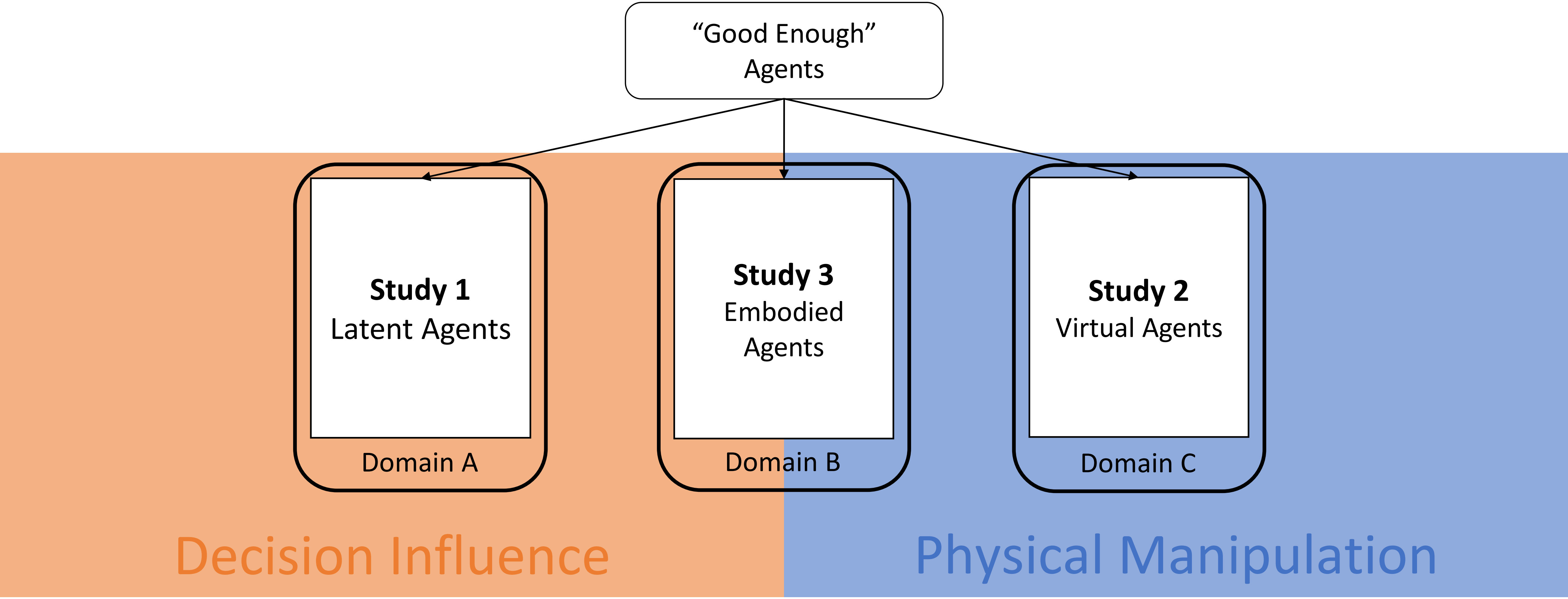
Research Questions

How does lowering agent reliability (i.e., “Good Enough” agents) affect task performance and human trust across different domains?

What is the relationship between human individual differences, agent reliability, and technology use?

How do different task requirements interact with the human-AI interaction dynamic?

Dissertation Approach



“Good Enough” Agents

Study 1 Latent Agents

Domain A

Study 3 Embodied Agents

Domain B

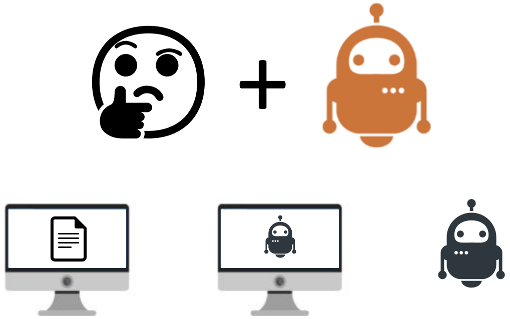
Study 2 Virtual Agents

Domain C

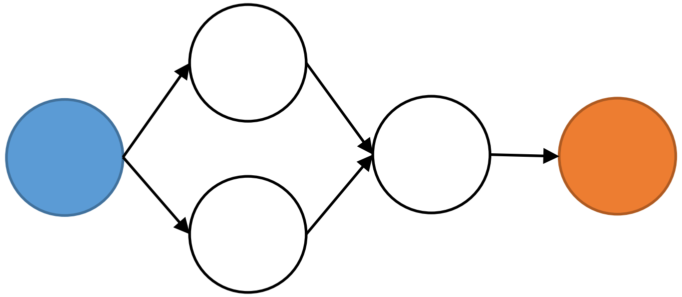
Decision Influence

Physical Manipulation

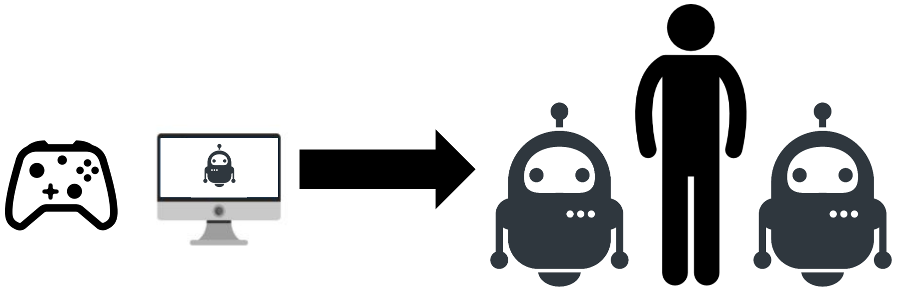
Dissertation Contributions



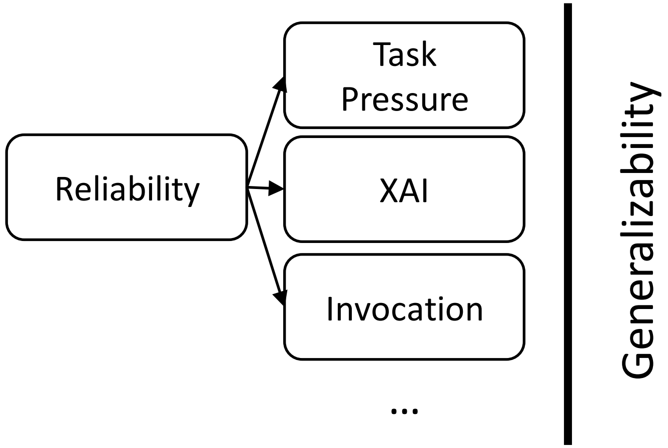
Evidence showing “Good Enough” agents lead to positive outcomes in different contexts



Holistic mediation models for human-AI interaction



Validation of 3 game-based tasks for human-AI interaction research



An ontology of task features that impede generalizability in reliability research

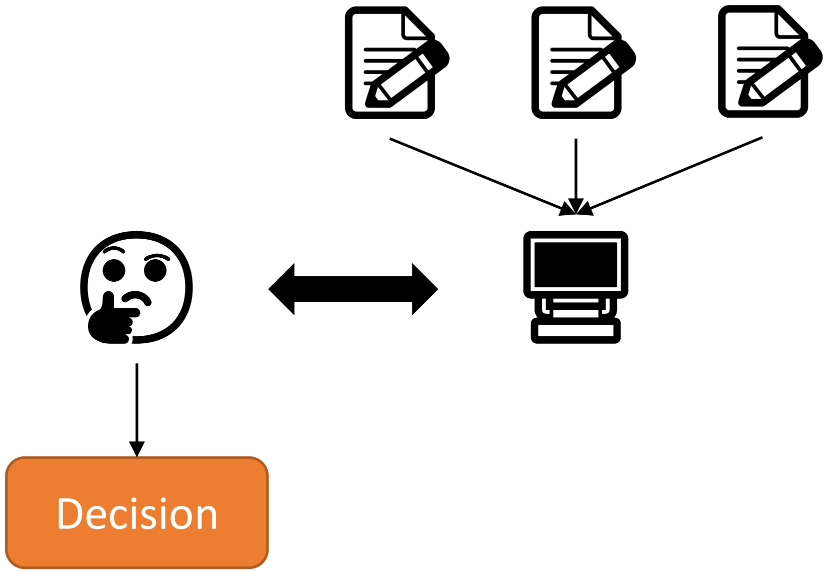
Study 1: Latent Agents

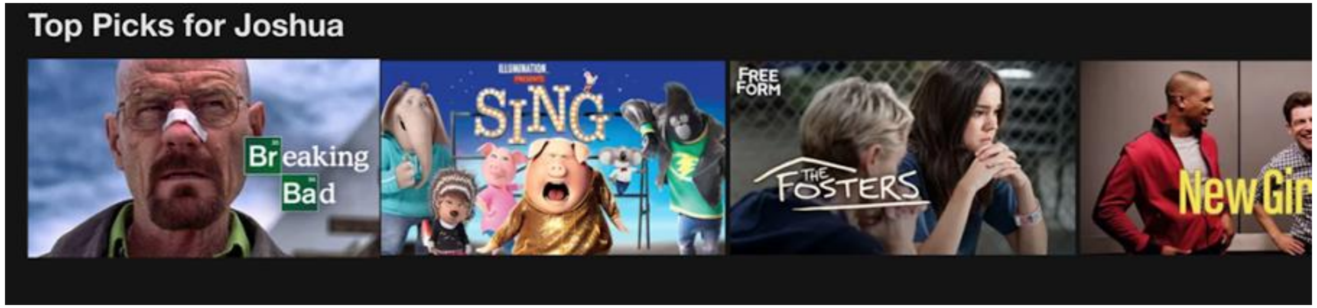
Study 2: Virtual Agents

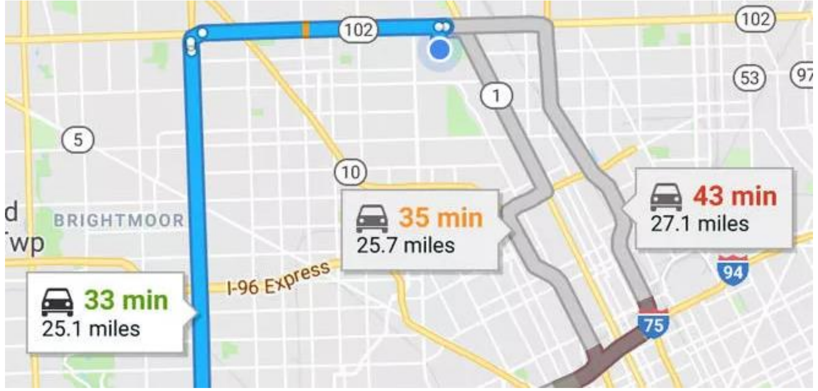
Study 3: Embodied Agents

Decision Support Systems (DSS)

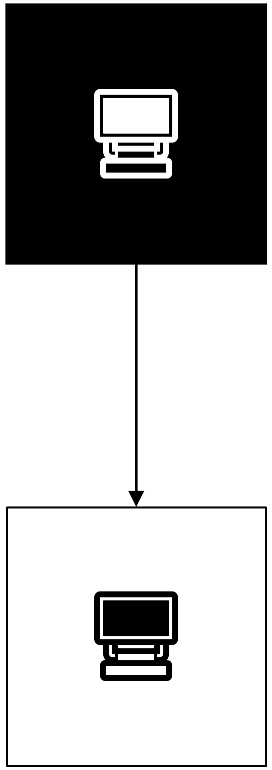
- Used daily!
- Any AI system that assists the sensemaking process
- Exists in multiple domains: navigation, recommender systems, medicine diagnostics




Top Picks for Joshua



From Black to White



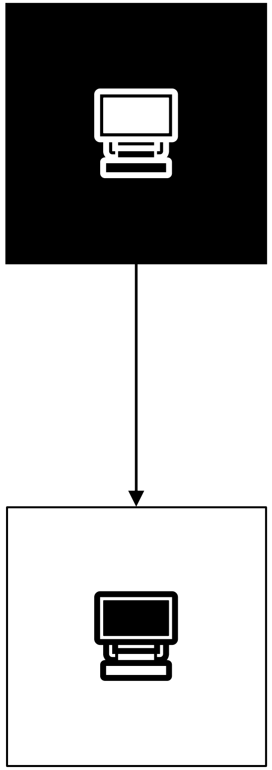
- DSS are sophisticated and complex
  - Challenging even for experts to understand[a]
- Address with *transparency*[b] and *customizability*[c]
  - Increases trustworthiness and adoption
- New issue: Increasing and improperly calibrated trust

a V. Arnold, P. A. Collier, S. A. Leech, and S. G. Sutton. Impact of intelligent decision aids on expert and novice decision-makers judgments. Accounting & Finance, 44, 1. 2004.

b V. Arnold, N. Clark, P. A. Collier, S. A. Leech, and S. G. Sutton. The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. Mis Quarterly. 2006

c B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan, and A. Kobsa. Inspectability and control in social recommenders. Conference on Recommender systems. 2012.

From Black to White



- DSS research done: complacency, performance degradation, situation awareness
  - DSS feature X → metric Z
- Arnold et al. suggest knowledge plays a large factor in judgment and performance[a]
  - DSS feature X → knowledge → metric Z
- *Could white-box features harm knowledge?*

[a] V. Arnold, N. Clark, P. A. Collier, S. A. Leech, and S. G. Sutton. The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. Mis Quarterly. 2006

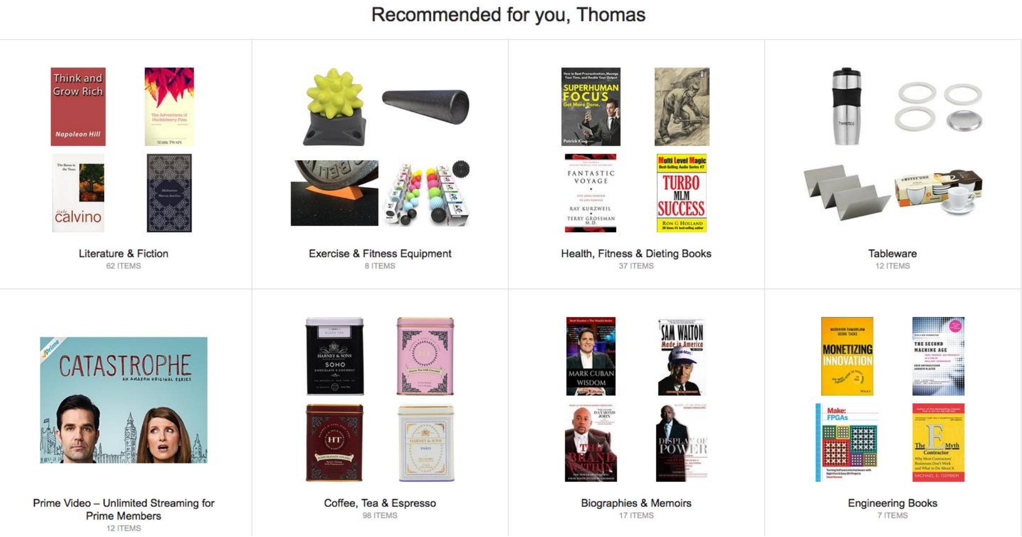
Study 1: Latent Agents

Research Question

How do latent agents affect domain knowledge acquisition and retention, and how is it affected by lowered reliability?

Task Context

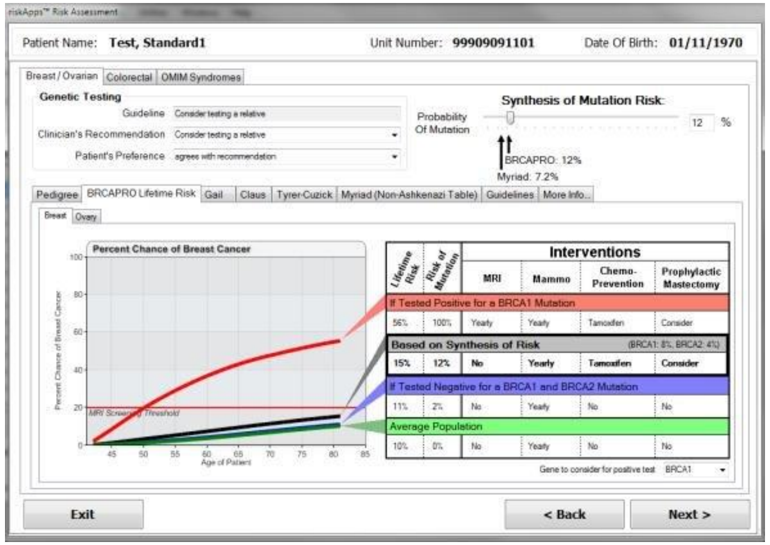
DSS tasks can be either subjective or objective



Recommended for you, Thomas

Literature & Fiction 62 ITEMS	Exercise & Fitness Equipment 8 ITEMS	Health, Fitness & Dieting Books 37 ITEMS	Tableware 12 ITEMS
Prime Video – Unlimited Streaming for Prime Members 12 ITEMS	Coffee, Tea & Espresso 98 ITEMS	Biographies & Memoirs 17 ITEMS	Engineering Books 7 ITEMS

Success = Satisfaction



riskApps™ Risk Assessment

Patient Name: **Test, Standard1** Unit Number: **99909091101** Date Of Birth: **01/11/1970**

Genetic Testing

Guideline: Consider testing a relative

Clinician's Recommendation: Consider testing a relative

Patient's Preference: agrees with recommendation

Synthesis of Mutation Risk:

Probability Of Mutation: 12 %

BRCAPRO: 12%

Myriad: 7.2%

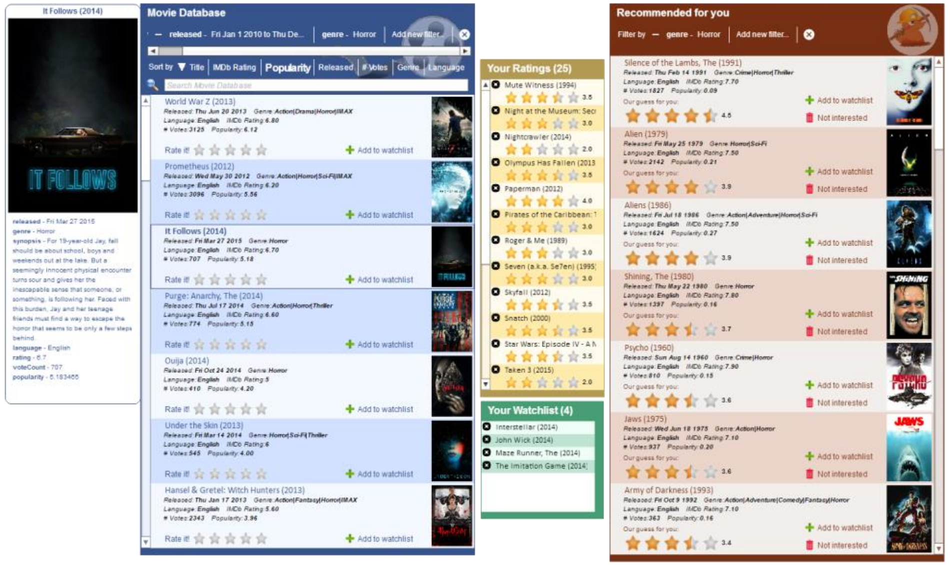
Lifetime Risk	Risk of Mutation	MRI	Mammo	Chemo-Prevention	Prophylactic Mastectomy
If Tested Positive for a BRCA1 Mutation					
56%	100%	Yearly	Yearly	Tamoxifen	Consider
Based on Synthesis of Risk (BRCA1: 8%, BRCA2: 4%)					
15%	**12%**	**No**	**Yearly**	**Tamoxifen**	**Consider**
If Tested Negative for a BRCA1 and BRCA2 Mutation					
11%	2%	No	Yearly	No	No
Average Population					
10%	0%	No	Yearly	No	No

Gene to consider for positive test: BRCA1

Success = Performance

Task Selection: Recommender Systems

DSS tasks can be either subjective or objective



It Follows (2014)

released - Fri Mar 27 2015
genre - Horror
synopsis - For 19-year-old Jay, fall should be about school, boys and weekends out at the lake. But a seemingly innocent physical encounter turns sour and gives her the inescapable sense that someone, or something, is following her. Faced with this burden, Jay and her teenage friends must find a way to escape the horror that seems to be only a few steps behind.
language - English
rating - 6.7
voteCount - 707
popularity - 5.183466

Movie Database

— released - Fri Jan 1 2010 to Thu De... | genre - Horror | Add new filter...

Sort by ▼ Title | IMDb Rating | **Popularity** | Released | # Votes | Genre | Language

Search Movie Database

World War Z (2013)
Released: Thu Jun 20 2013 Genre: Action|Drama|Horror|IMAX
Language: English IMDb Rating: 6.80
Votes: 3125 Popularity: 6.12
Rate it! ☆☆☆☆☆ + Add to watchlist

Prometheus (2012)
Released: Wed May 30 2012 Genre: Action|Horror|Sci-Fi|IMAX
Language: English IMDb Rating: 6.20
Votes: 3096 Popularity: 5.56
Rate it! ☆☆☆☆☆ + Add to watchlist

It Follows (2014)
Released: Fri Mar 27 2015 Genre: Horror
Language: English IMDb Rating: 6.70
Votes: 707 Popularity: 5.18
Rate it! ☆☆☆☆☆ + Add to watchlist

Purge: Anarchy, The (2014)
Released: Thu Jul 17 2014 Genre: Action|Horror|Thriller
Language: English IMDb Rating: 6.60
Votes: 774 Popularity: 5.15
Rate it! ☆☆☆☆☆ + Add to watchlist

Ouija (2014)
Released: Fri Oct 24 2014 Genre: Horror
Language: English IMDb Rating: 5
Votes: 410 Popularity: 4.20
Rate it! ☆☆☆☆☆ + Add to watchlist

Under the Skin (2013)
Released: Fri Mar 14 2014 Genre: Horror|Sci-Fi|Thriller
Language: English IMDb Rating: 6
Votes: 545 Popularity: 4.00
Rate it! ☆☆☆☆☆ + Add to watchlist

Hansel & Gretel: Witch Hunters (2013)
Released: Thu Jan 17 2013 Genre: Action|Fantasy|Horror|IMAX
Language: English IMDb Rating: 5.60
Votes: 2343 Popularity: 3.96
Rate it! ☆☆☆☆☆ + Add to watchlist

Your Ratings (25)

Mute Witness (1994)	3.5
Night at the Museum: Secr	3.0
Nightcrawler (2014)	2.0
Olympus Has Fallen (2013	3.5
Paperman (2012)	4.0
Pirates of the Caribbean: T	3.0
Roger & Me (1989)	3.0
Seven (a.k.a. Se7en) (1995)	3.0
Skyfall (2012)	3.5
Snatch (2000)	3.5
Star Wars: Episode IV - A N	3.5
Taken 3 (2015)	2.0

Your Watchlist (4)

- Interstellar (2014)
- John Wick (2014)
- Maze Runner, The (2014)
- The Imitation Game (2014)

Recommended for you

Filter by — genre - Horror | Add new filter...

Silence of the Lambs, The (1991)
Released: Thu Feb 14 1991 Genre: Crime|Horror|Thriller
Language: English IMDb Rating: 7.70
Votes: 1827 Popularity: 0.09
Our guess for you: 4.5 + Add to watchlist Not interested

Alien (1979)
Released: Fri May 25 1979 Genre: Horror|Sci-Fi
Language: English IMDb Rating: 7.50
Votes: 2142 Popularity: 0.21
Our guess for you: 3.9 + Add to watchlist Not interested

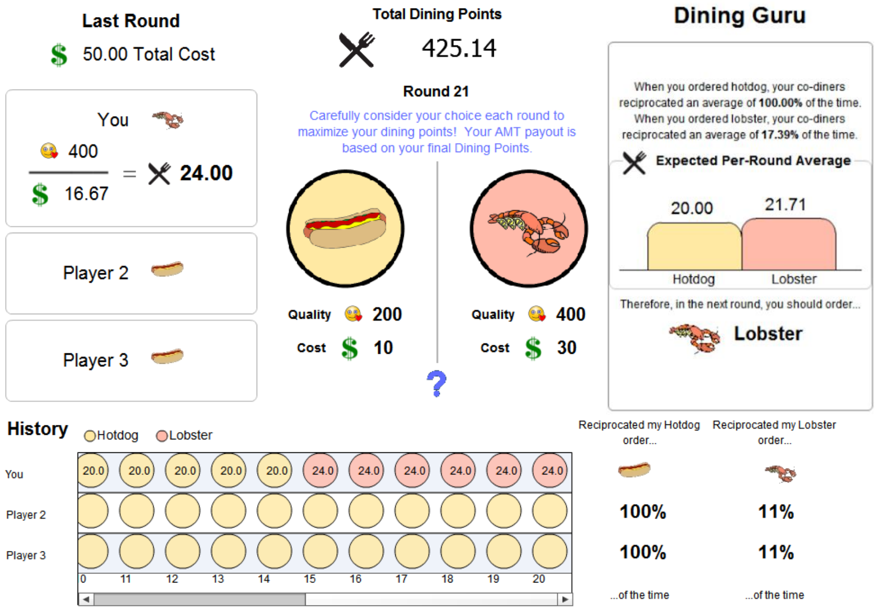
Aliens (1986)
Released: Fri Jul 18 1986 Genre: Action|Adventure|Horror|Sci-Fi
Language: English IMDb Rating: 7.50
Votes: 1624 Popularity: 0.27
Our guess for you: 3.9 + Add to watchlist Not interested

Shining, The (1980)
Released: Thu May 22 1980 Genre: Horror
Language: English IMDb Rating: 7.80
Votes: 1397 Popularity: 0.16
Our guess for you: 3.7 + Add to watchlist Not interested

Psycho (1960)
Released: Sun Aug 14 1960 Genre: Crime|Horror
Language: English IMDb Rating: 7.90
Votes: 810 Popularity: 0.15
Our guess for you: 3.6 + Add to watchlist Not interested

Jaws (1975)
Released: Wed Jun 18 1975 Genre: Action|Horror
Language: English IMDb Rating: 7.10
Votes: 937 Popularity: 0.20
Our guess for you: 3.6 + Add to watchlist Not interested

Army of Darkness (1993)
Released: Fri Oct 9 1992 Genre: Action|Adventure|Comedy|Fantasy|Horror
Language: English IMDb Rating: 7.10
Votes: 363 Popularity: 0.16
Our guess for you: 3.4 + Add to watchlist Not interested



Last Round

$ 50.00 Total Cost

Total Dining Points

425.14

Round 21

Carefully consider your choice each round to maximize your dining points! Your AMT payout is based on your final Dining Points.

You
400
$ 16.67
= 24.00

Player 2

Player 3

Quality 200
Cost $ 10

Quality 400
Cost $ 30

?

Dining Guru

When you ordered hotdog, your co-diners reciprocated an average of **100.00%** of the time. When you ordered lobster, your co-diners reciprocated an average of **17.39%** of the time.

Expected Per-Round Average

Therefore, in the next round, you should order...

Lobster

History ○Hotdog ○Lobster

	0	11	12	13	14	15	16	17	18	19	20
You	20.0	20.0	20.0	20.0	20.0	24.0	24.0	24.0	24.0	24.0	24.0
Player 2											
Player 3											

	Reciprocated my Hotdog order...	Reciprocated my Lobster order...
Player 2	100%	11%
Player 3	100%	11%
	...of the time	...of the time

Success = Satisfaction

Movie Recommendation Task

Make a watchlist from recommended movies

Predicted to forget knowledge about movies

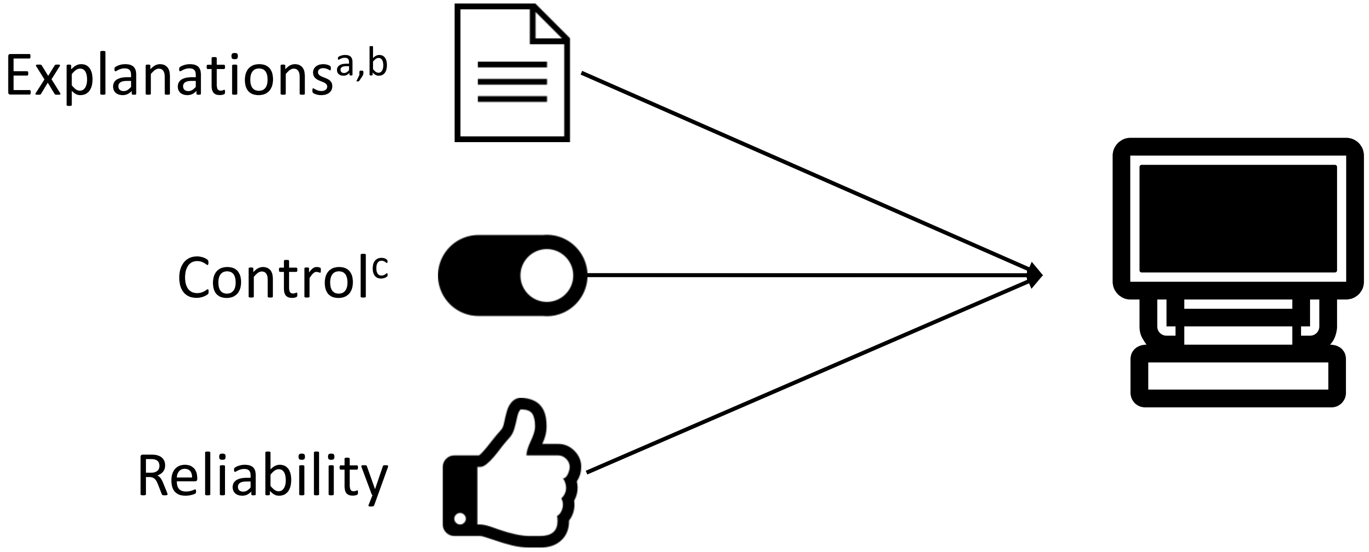
Success = Performance

Diner's Dilemma Game

Eat the most while paying the least

Predicted to learn rules of the game

Independent Variables



Explanations[a,b]

Control[c]

Reliability

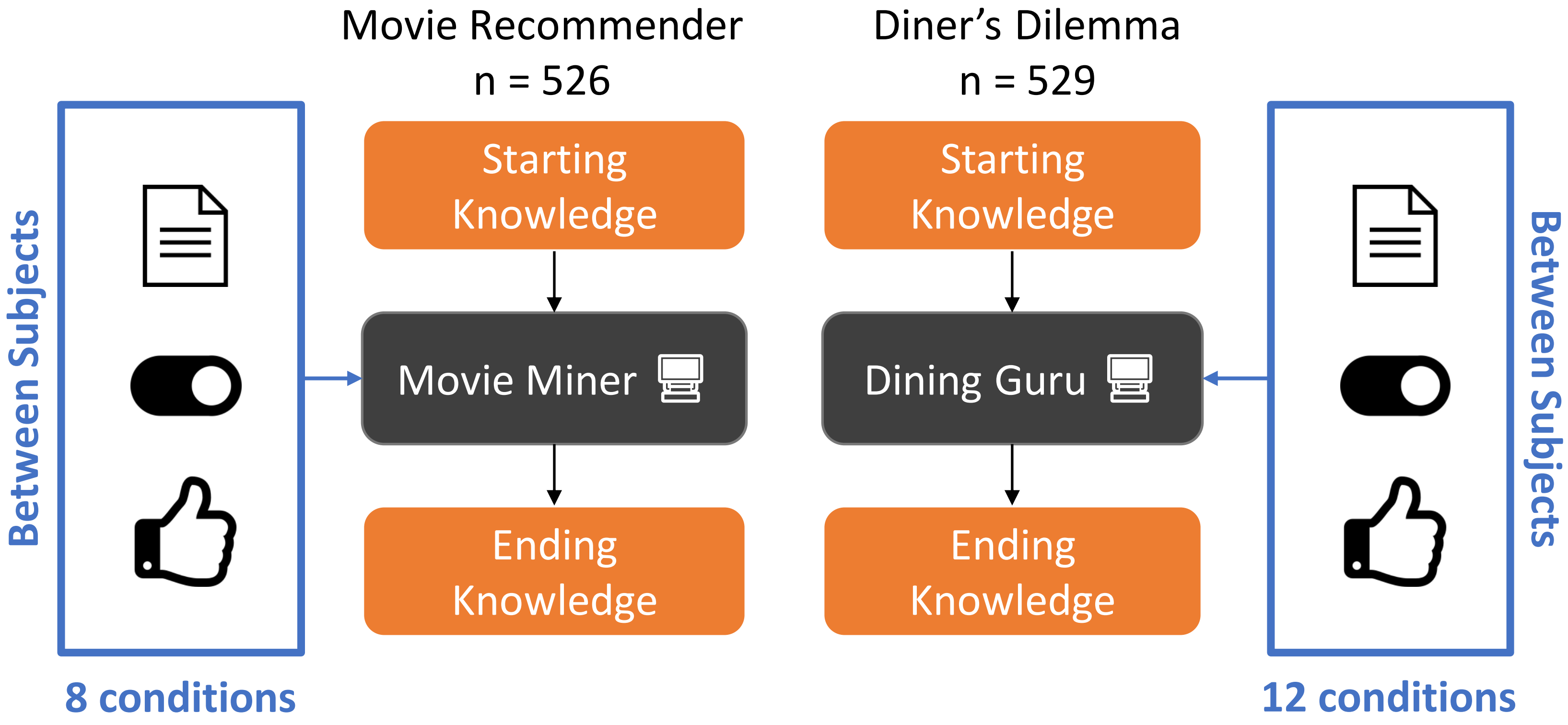
These have granularities, but we'll simplify with dichotomies

[a] N. Tintarev and J. Masthoff. A survey of explanations in recommender systems. In Data Engineering Workshop, 2007 IEEE 23rd International Conference on.

[b] V. Arnold, N. Clark, P. A. Collier, S. A. Leech, and S. G. Sutton. The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. Mis Quarterly. 2006

[c] B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan, and A. Kobsa. Inspectability and control in social recommenders. Conference on Recommender systems. 2012.

Study Design

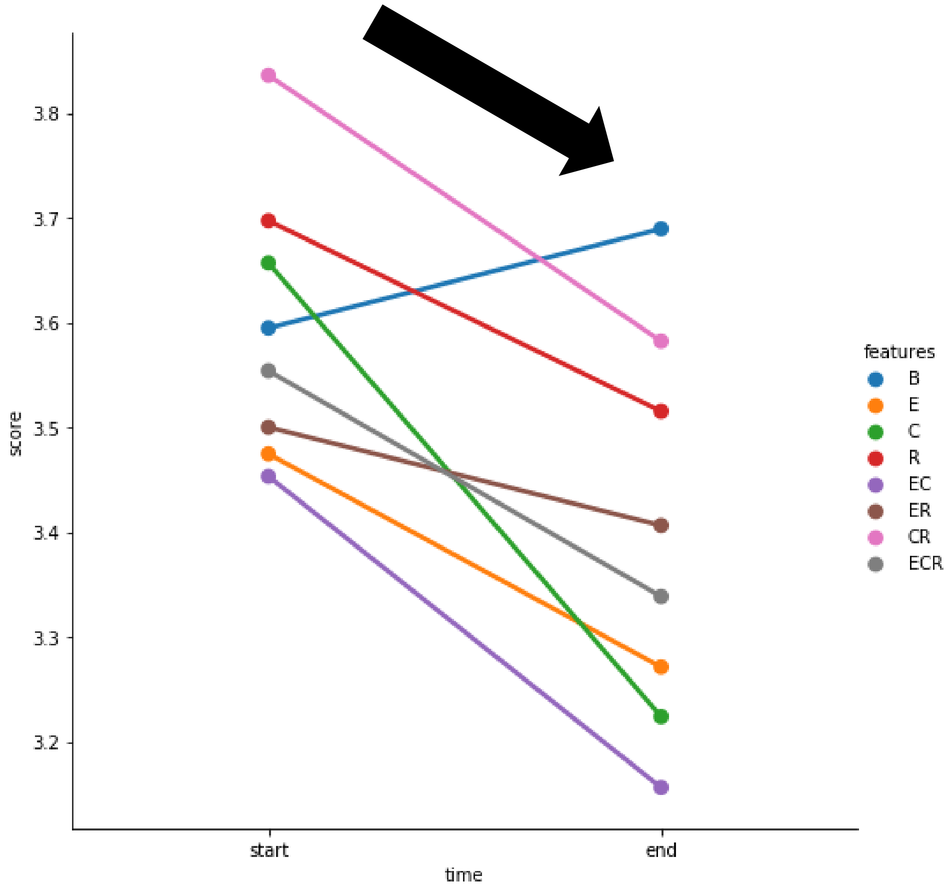


Movie Recommender
n = 526

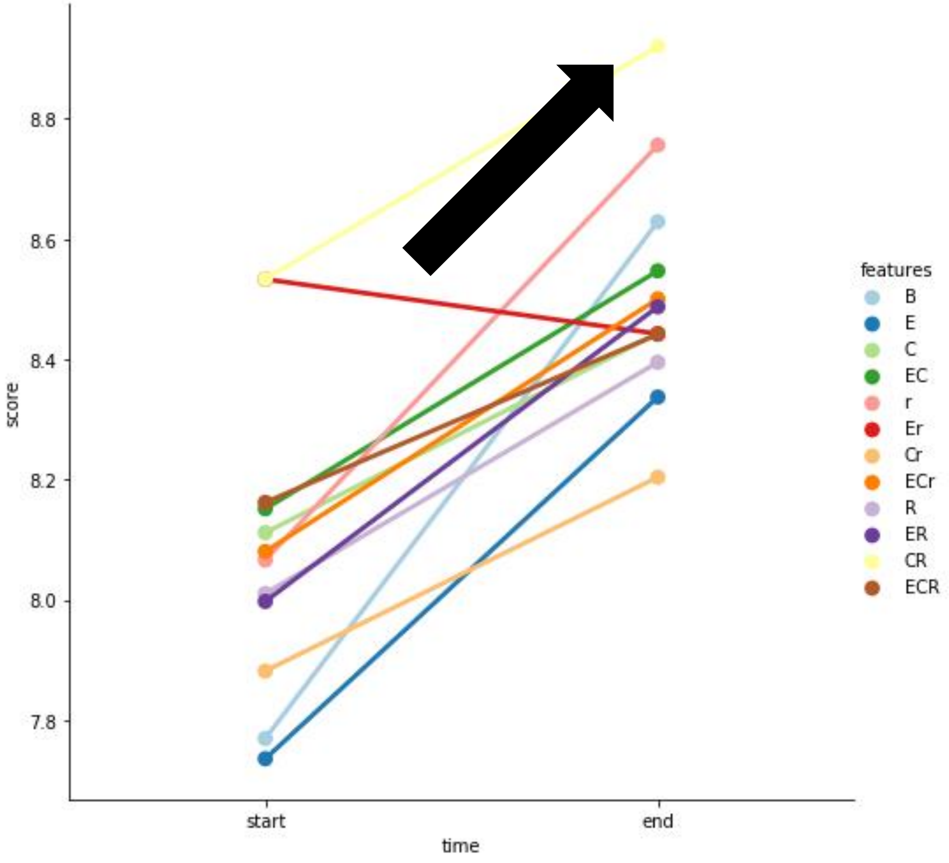
Diner's Dilemma
n = 529

Recruited through Amazon Mechanical Turk

Knowledge changes over time (expected)

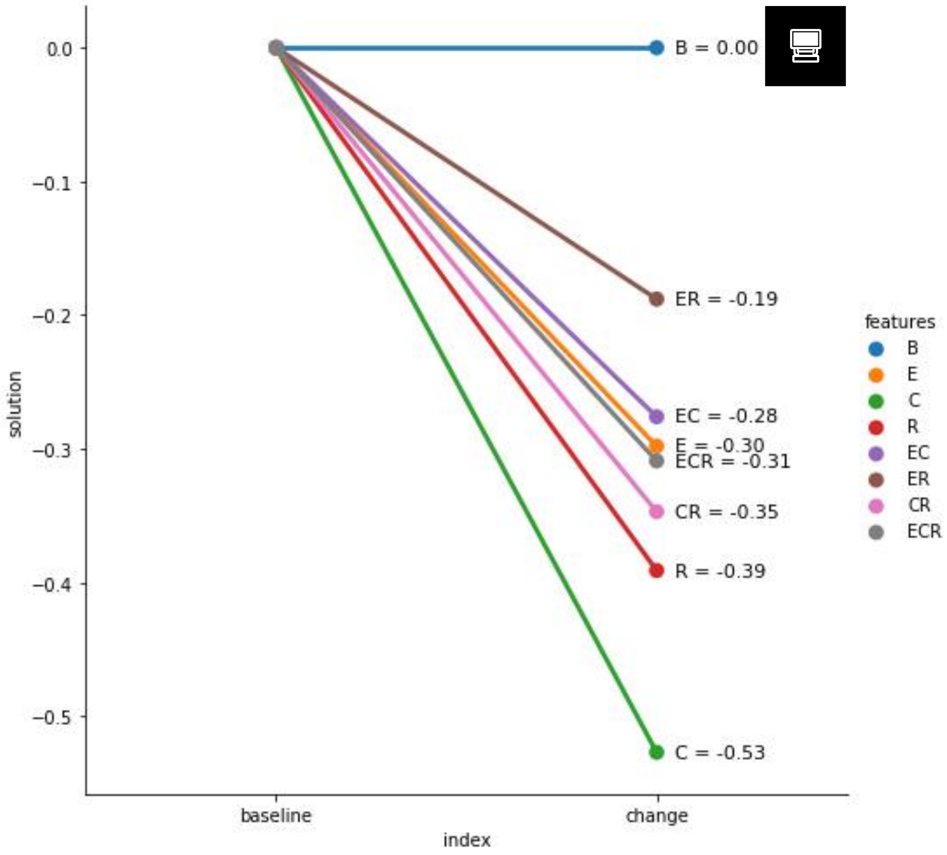


Movie (loss)

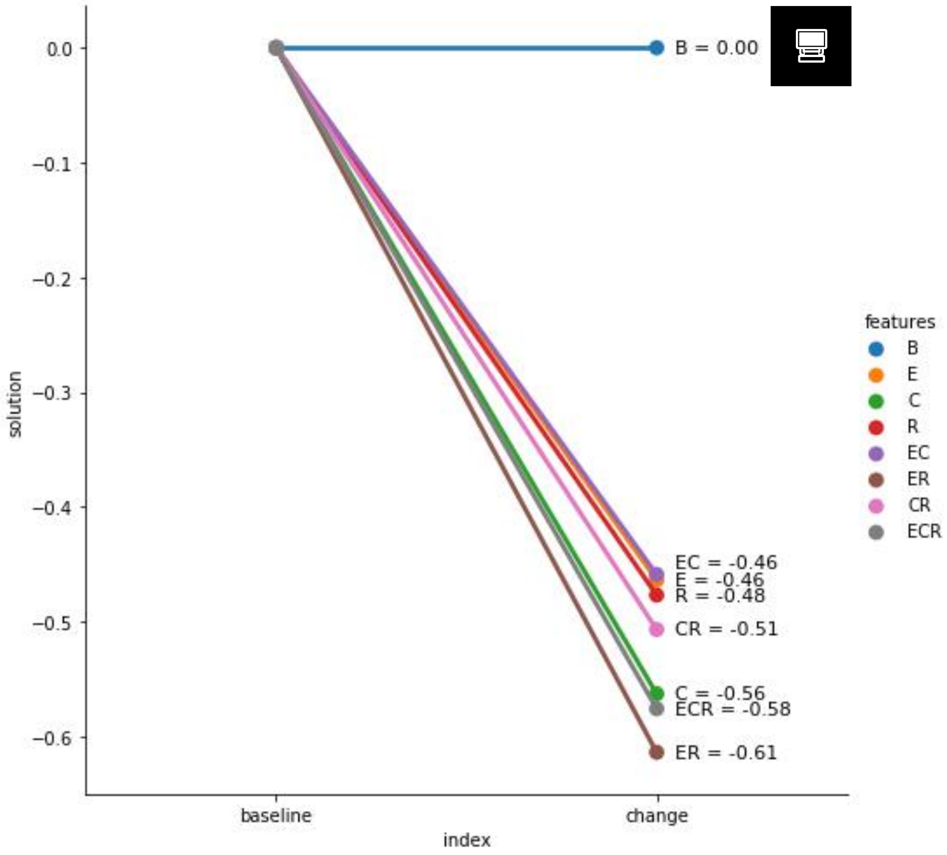


Diner (gain)

Potential gains are best with black boxes



Movie (loss)



Diner (gain)

Worst culprit: control settings

Movie (n = 526)

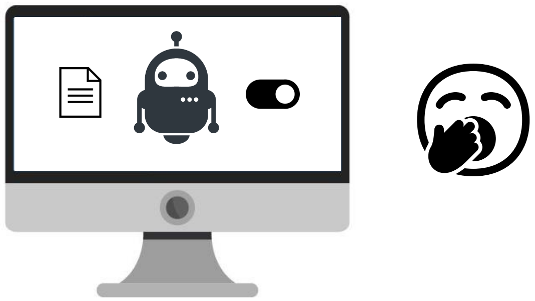
	B	p-value
Explanation	-0.298	0.088
Control	-0.527	0.002**
Reliability	-0.391	0.022*
EC	0.549	0.026*
ER	0.501	0.044*
CR	0.571	0.019*
ECR	-0.714	0.041*

Diner (n = 529)

	B	p-value
Explanation	-0.464	0.015*
Control	-0.563	0.004**
Reliability	-0.477	0.025*
EC	0.568	0.037*
ER	0.327	0.27
CR	0.533	0.078
ECR	-0.5	0.239

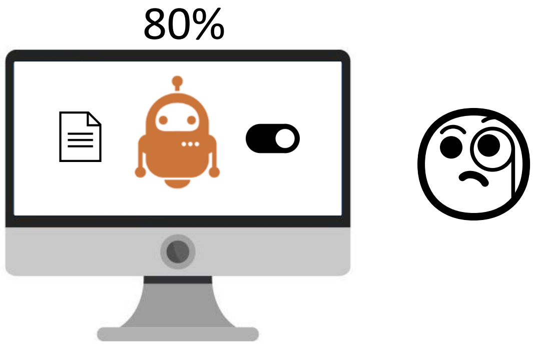
White box harms mitigated by low reliability
Task features led to different interactions; more research needed

Summary



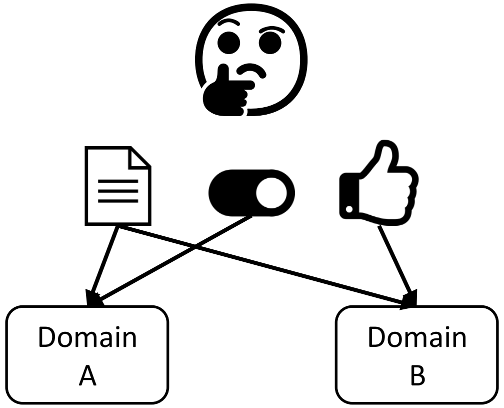
Feature-filled agents led to worsened learning – over-trust

Control was the worst culprit (corrupts the information retrieval process)



Lowering reliability hampers over-trust; users forced to exercise judgment

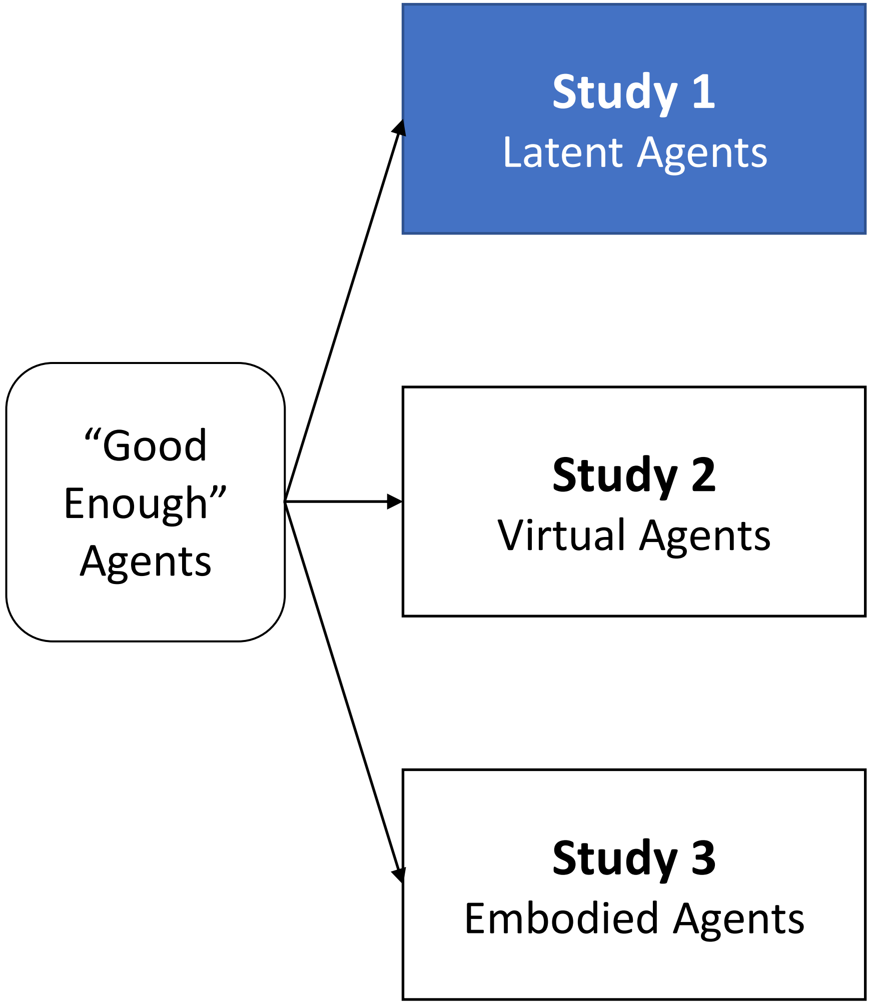
Caveat: lowered reliability itself can also cause wrong insights to be formed



Different tasks led to different feature effects (some overlap)

Task context remains relevant for interpreting outcomes

Dissertation Comparative Analysis



- “Good Enough” agents can help users exercise their judgment more often in influence tasks with decision support systems.
- The “gold” standard (explanations, control) could be “silver” instead.

Study 1: Latent Agents

Study 2: Virtual Agents

Study 3: Embodied Agents

Physical agents in the world

Fire

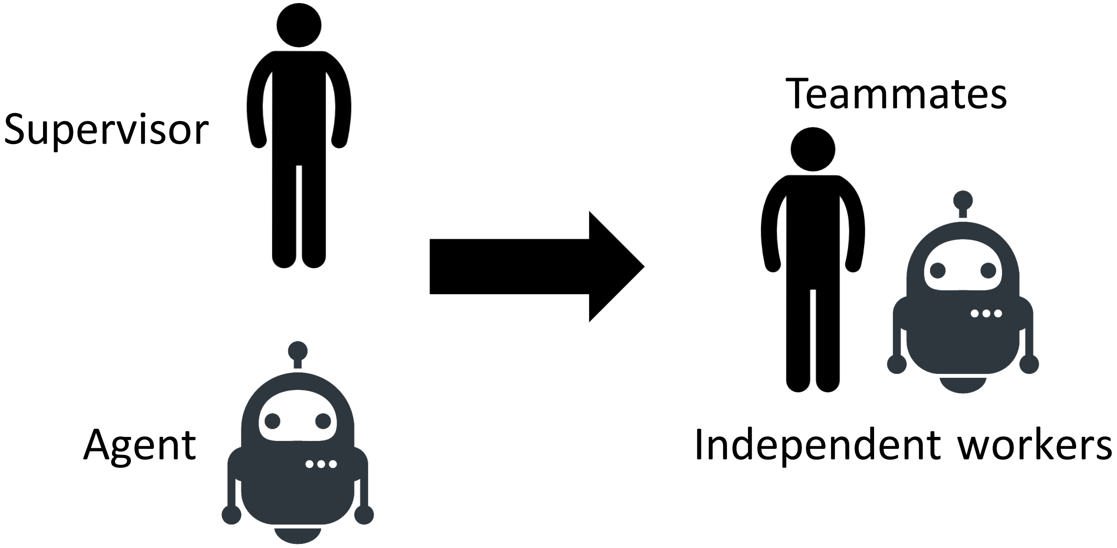
Navigation

Reconnaissance

Domain e.g.: IoBT
(little human intervention)

Emergent collaboration with physical non-controllable entities
Hard to study in a safe manner

Changes from latent agents



Supervisor

Agent

Teammates

Independent workers

- New considerations[a]
  - Non-supervisory role
  - Non-discrete decision making
  - Non-latent agents (virtuality simulating physicality)
- Scenarios difficult to study; simulation used for safety
- How does reliability, trust, and performance interact in simulated physical domains? Often not studied

[a] M. Endsley. From Here to Autonomy: Lessons Learned From Human-Automation Research. Human Factors, 59, 1. Feb 2017

Study 2: Virtual Agents

Research Questions

How is human performance affected by varying reliability in collaborating agents in a continuous pursuit task?

How do human individual differences mediate perception of agents, situation awareness, and performance in continuous domains?

Task Context

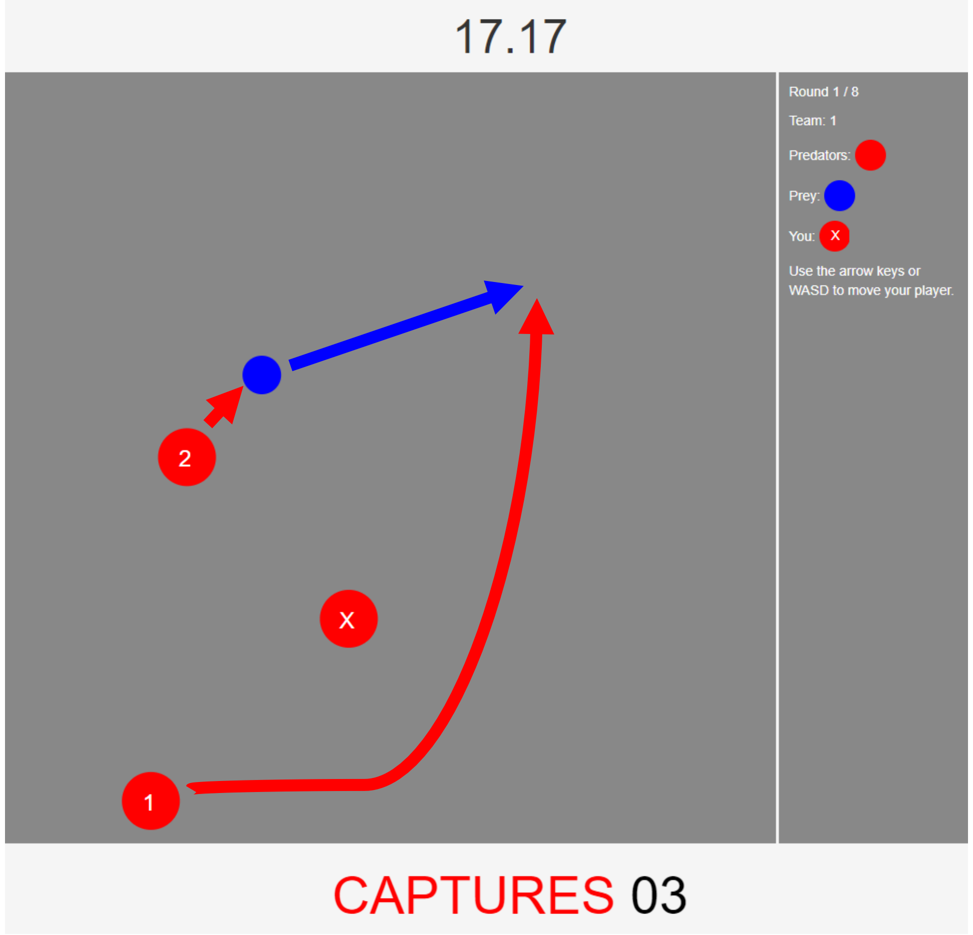
- Select task that models agent capabilities and usage
  - ARL collaboration: aerial drones for multi-agent systems[a,b]
- Target recognition and pursuit
  - Not exclusive to IoBT
  - Fundamental capabilities for physical tasks[b]
  - Covers large part of multi-agent interaction and human-robot interaction

a D. Asher and S. Barton. 2018. Reinforcement learning framework for collaborative agents interacting with soldiers in dynamicmilitary contexts.

b D. Asher, E. Zaroukian, and S. Barton. 2018. Adapting the Predator-Prey Game Theoretic Environment to Army Tactical Edge Scenarios with Computational Multiagent Systems

Task Selection: Predator-Prey Game

- Continuous pursuit task
  - Goal: capture the prey
- Asymmetrical and heterogeneous capabilities to engender collaboration
  - AI prey is faster
  - AI predators use different pursuit strategies (chase or intercept)
- Participants must be aware of their teammates for proper strategy-building
- Can situation awareness (SA) be affected by good performing agents?



17.17

Round 1 / 8

Team: 1

Predators: ●

Prey: ●

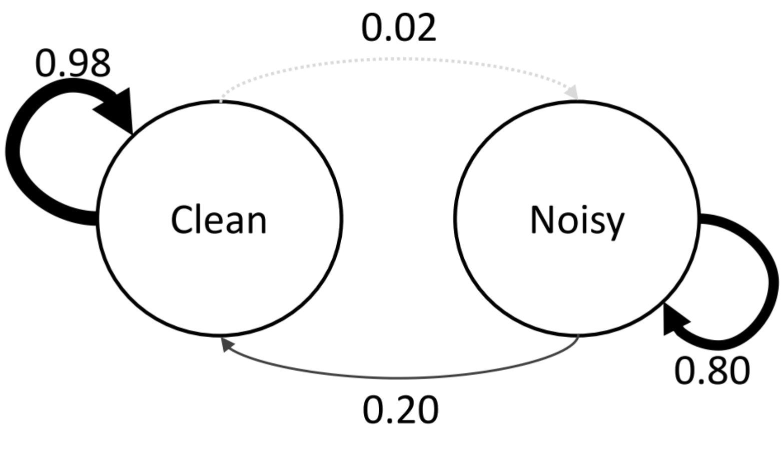
You: X

Use the arrow keys or WASD to move your player.

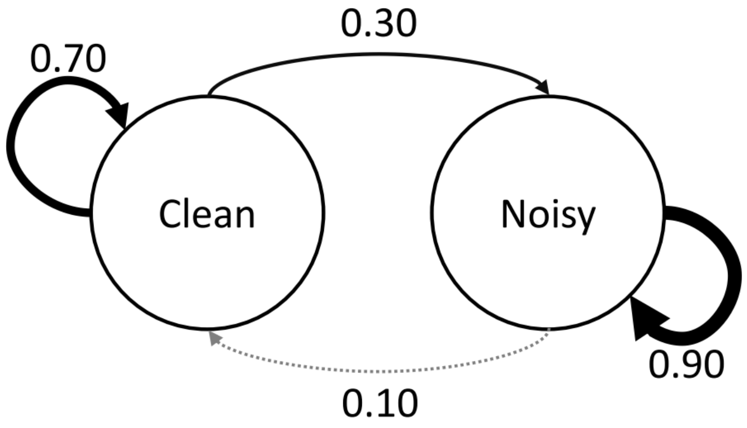
CAPTURES 03

Independent Variable: Reliability

- How do we design reliability?
  - Goal: simulate environmental/system randomness
  - 2 states
    - Clean: no movement interference
    - Noisy: movement interference
  - 2 MDPs
    - High and Low
    - Both able to chase the prey; Low (i.e., "Good Enough") sometimes acts erratically



High

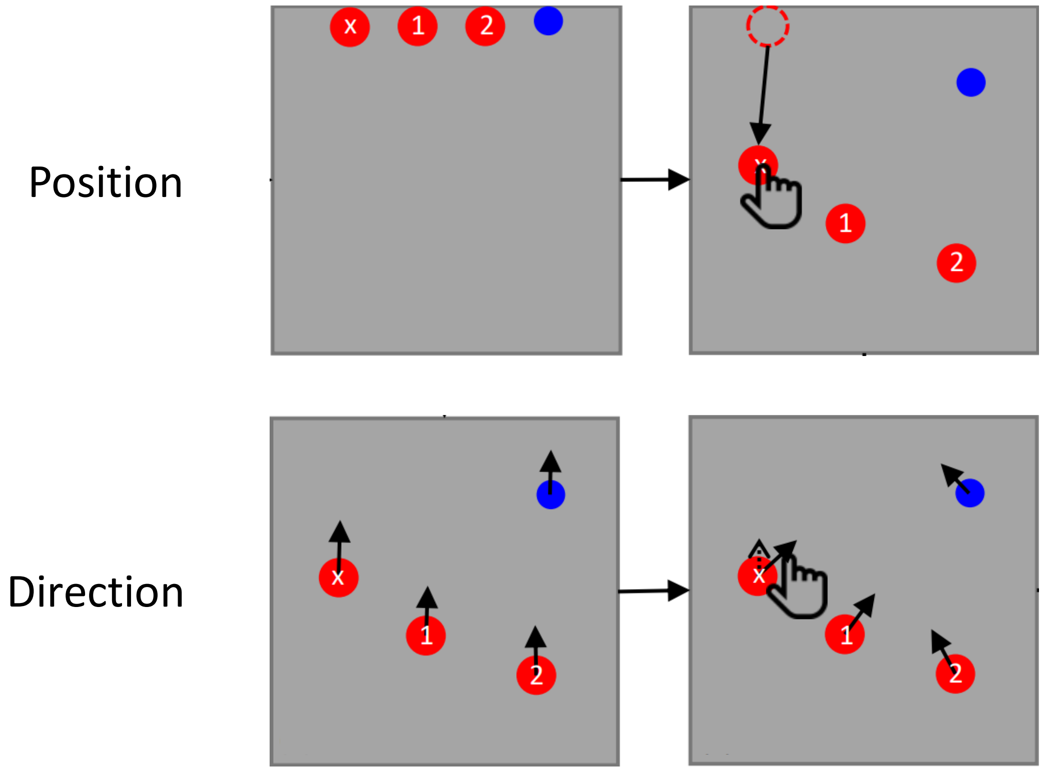


Low

Dependent Variables

Situation Awareness (SA)

State Reconstruction[f]



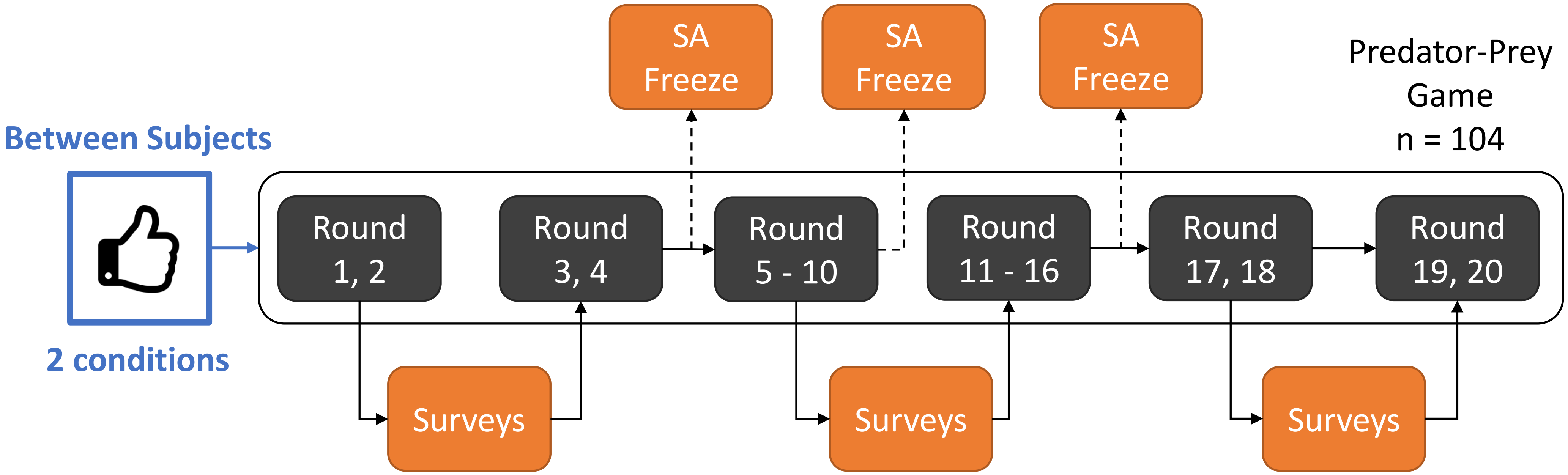
Surveys

Individual Differences (ID)/Subjective Perceptions (SP)

Factor	Construct	Survey
ID	**Motivation**	Intrinsic Motivation Inventory[a]
ID	**Trust Propensity**	Propensity to Trust Inventory[b]
ID	**Complacency Potential**	Automation-induced Complacency Potential Scale[c]
SP	**Trust in Agent**	Trust in Automated Systems[d]
SP	**Workload**	NASA Task Load Index[e]

a Ryan and Deci, 1982 **b** Jessup et al., 2019 **c** Merritt et al., 2019 **d** Jian et al., 2000 **e** Hart and Staveland, 1988 **f** Endsley, 1988.

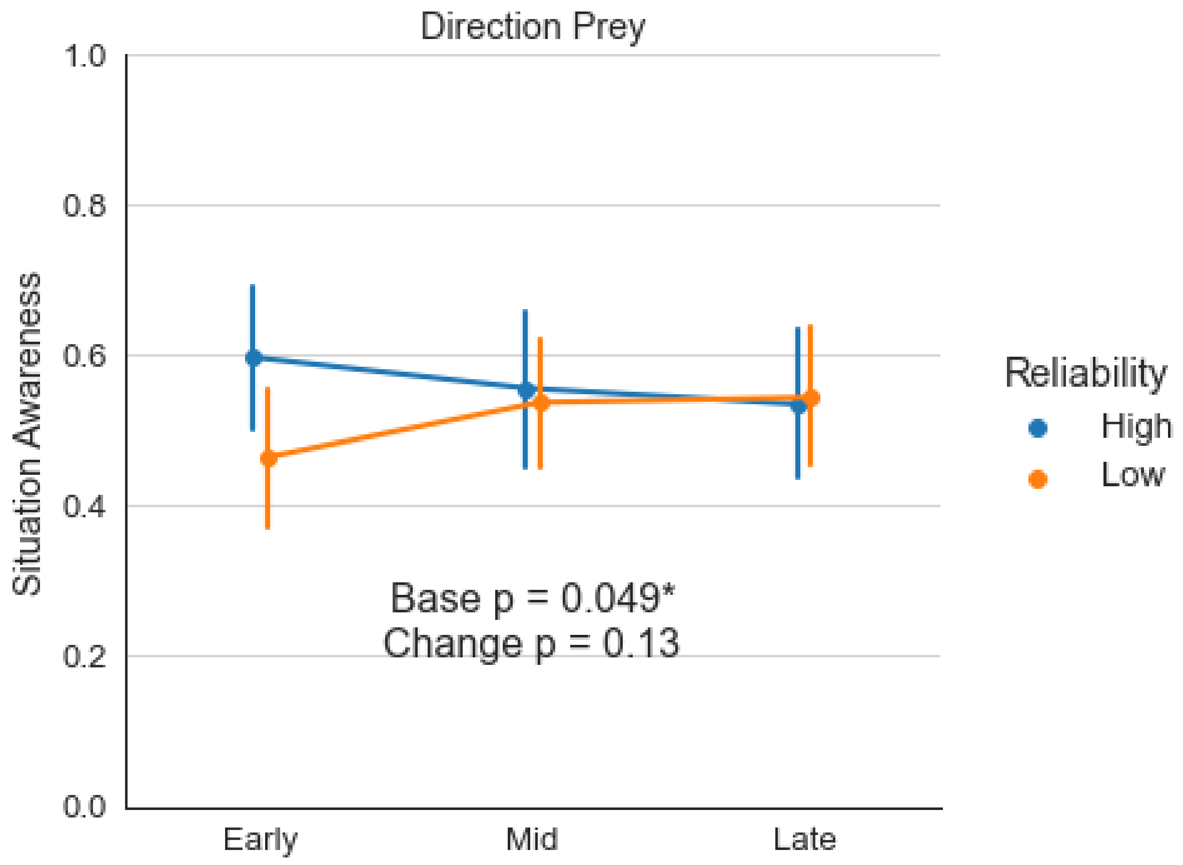
Study Design



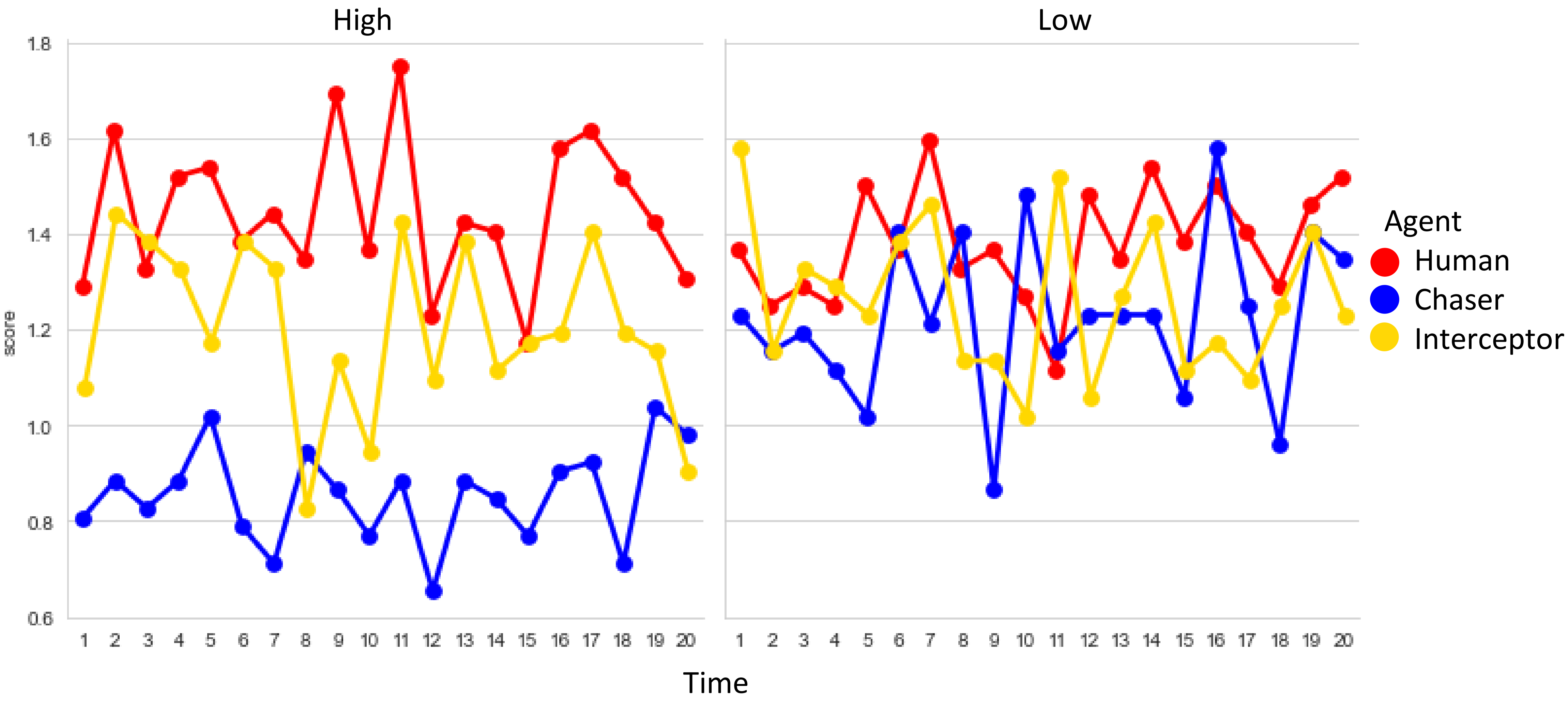
Recruited through Amazon Mechanical Turk

Non-significant changes in SA (mostly)

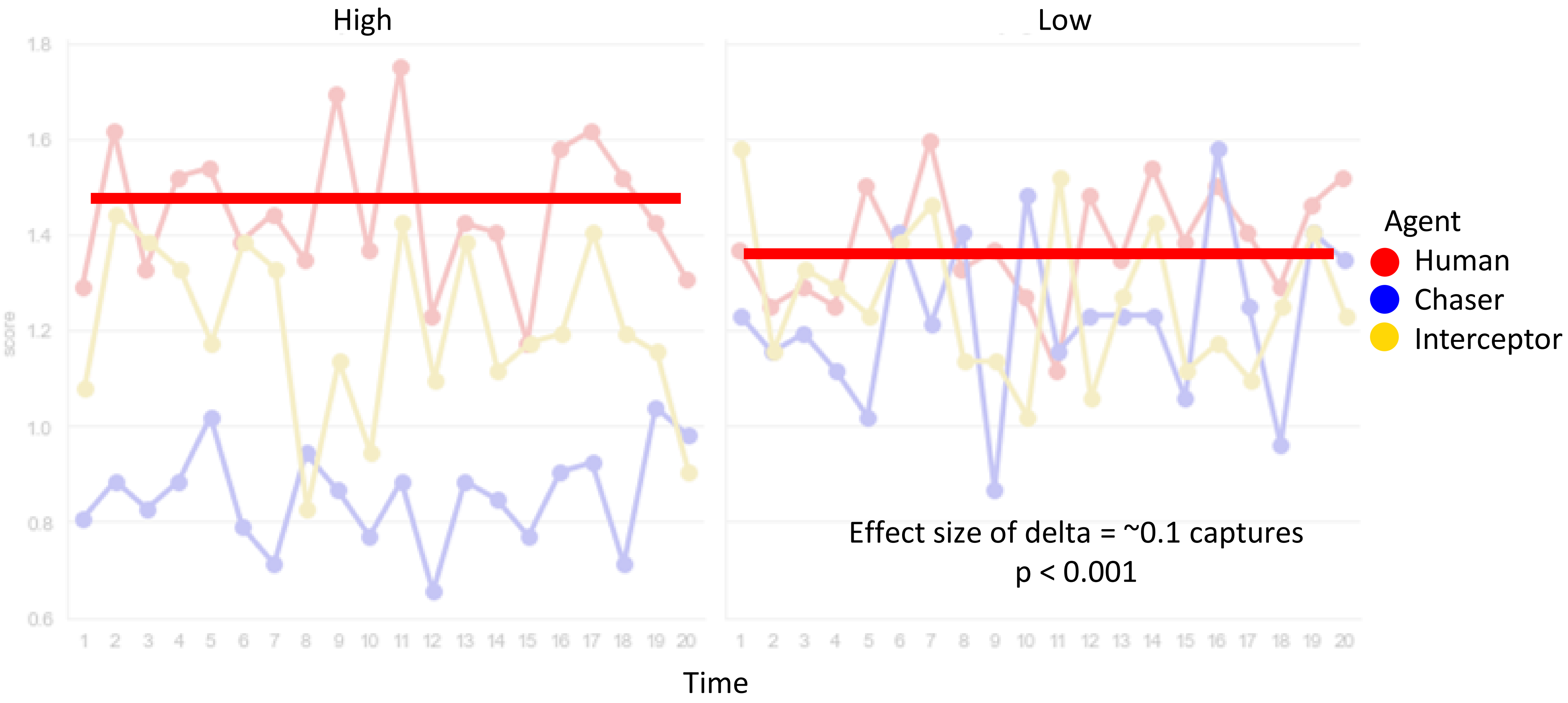
SA Probe	Base p-val	Change p-val
Position – Human	0.81	0.41
Position – Chaser	0.84	0.89
Position – Interceptor	0.33	0.23
Position – Prey	0.54	0.20
Direction – Human	0.30	0.73
Direction – Chaser	0.47	0.29
Direction – Interceptor	0.69	0.35
Direction – Prey	**0.049***	0.13



Minimal changes in human performance

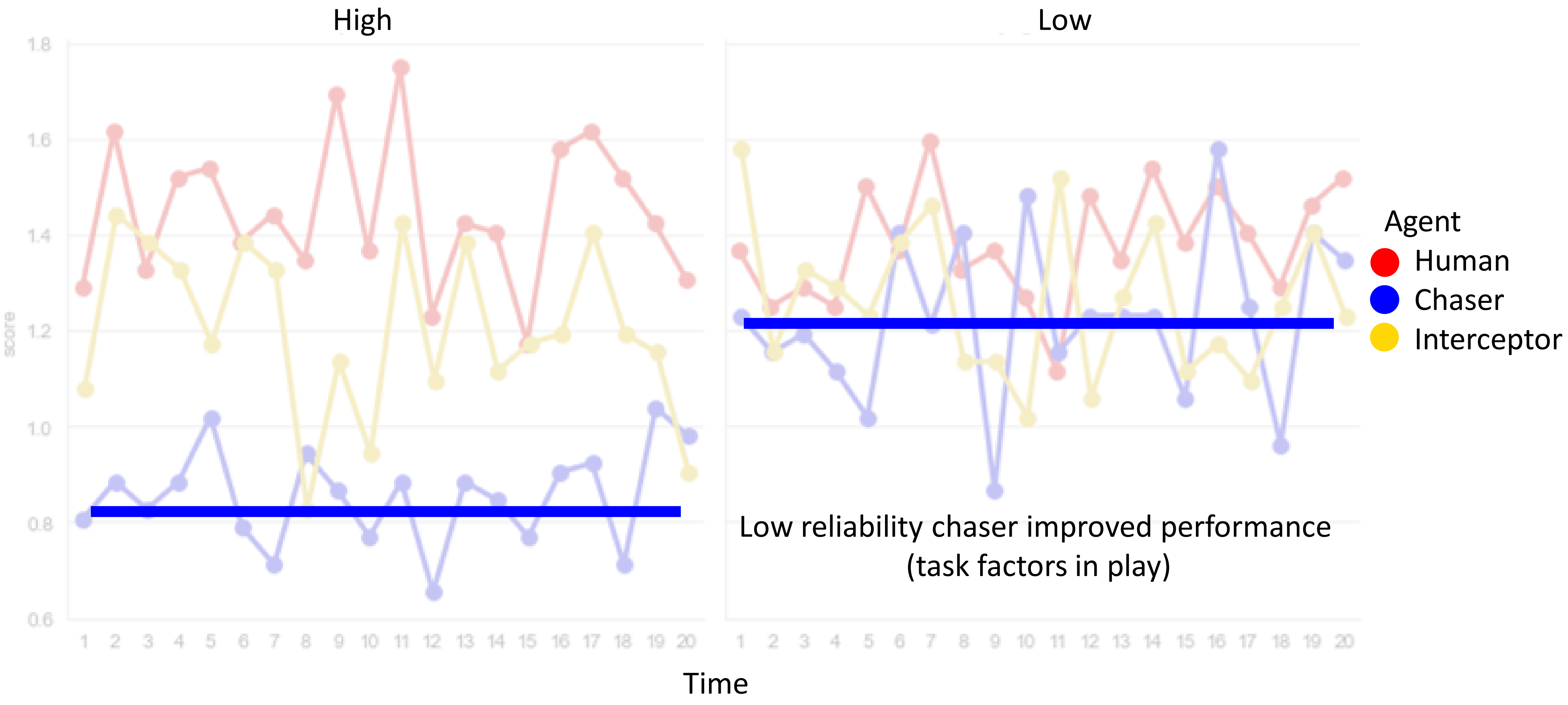


Minimal changes in human performance



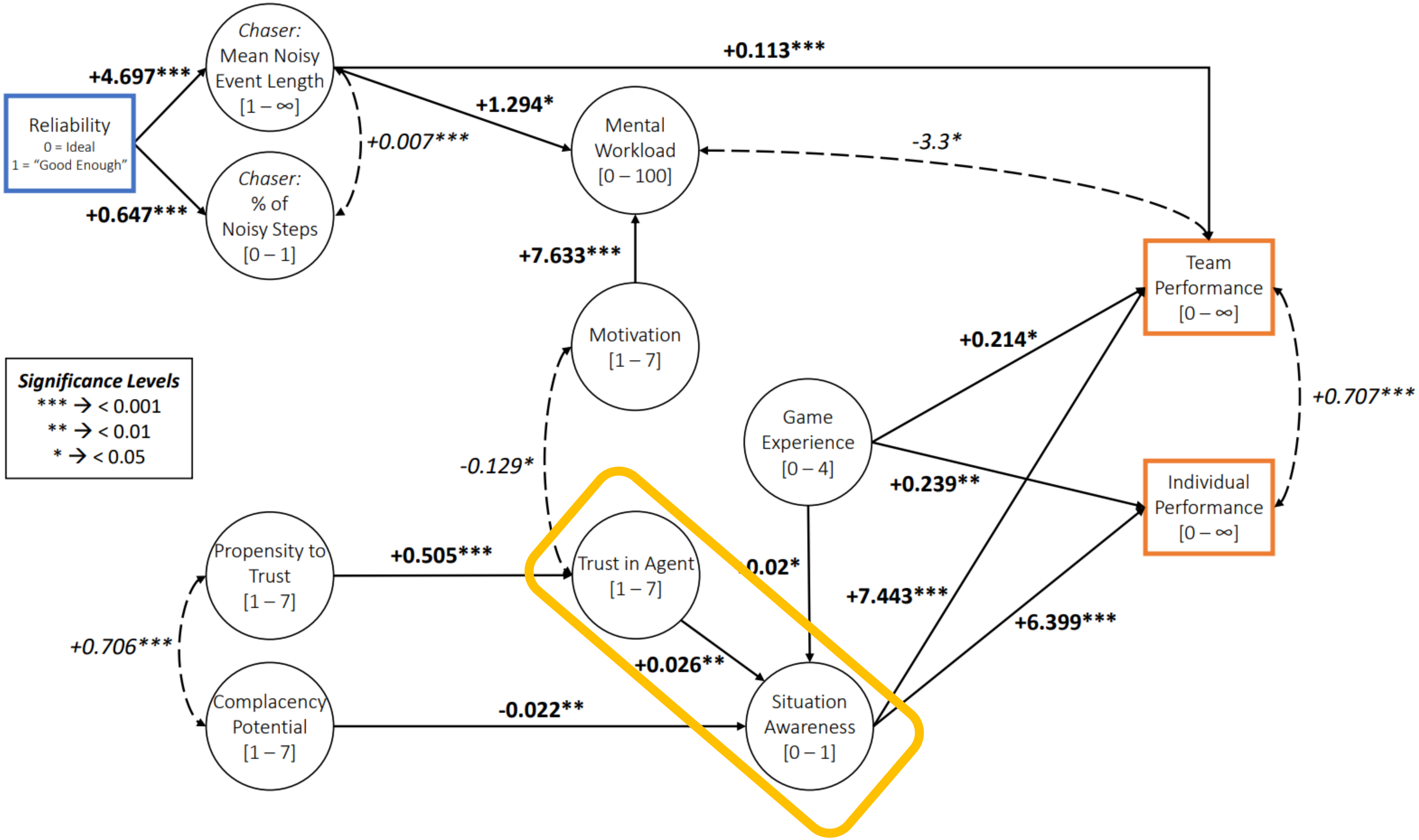
Effect size of delta = ~0.1 captures
p < 0.001

Minimal changes in human performance

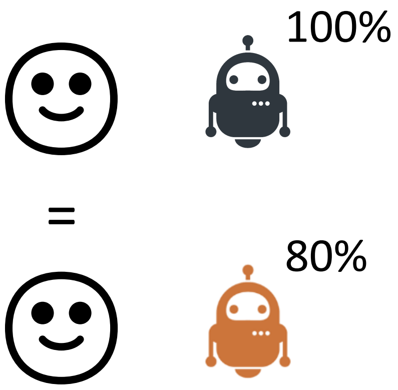


Low reliability chaser improved performance (task factors in play)

Final SEM: Differing trust effects

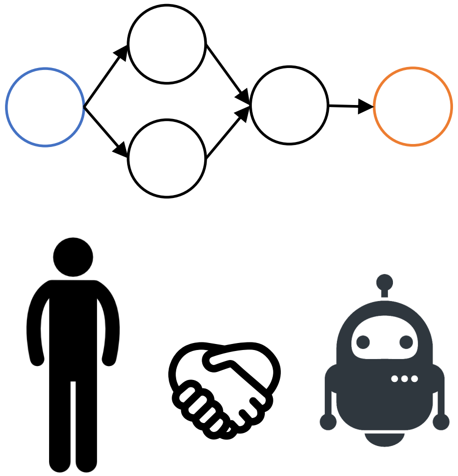


Summary



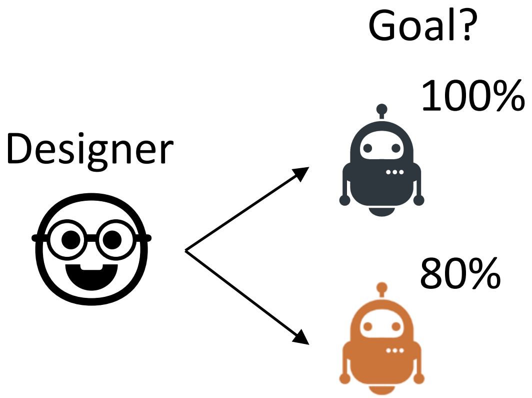
Reduced reliability led to improved team performance without affecting SA

Performance may differ per domain: the PPG may have benefitted from erraticness



The resulting SEM matches models from HF research: physical tasks and continuous environments are subject to similar complacency dynamics

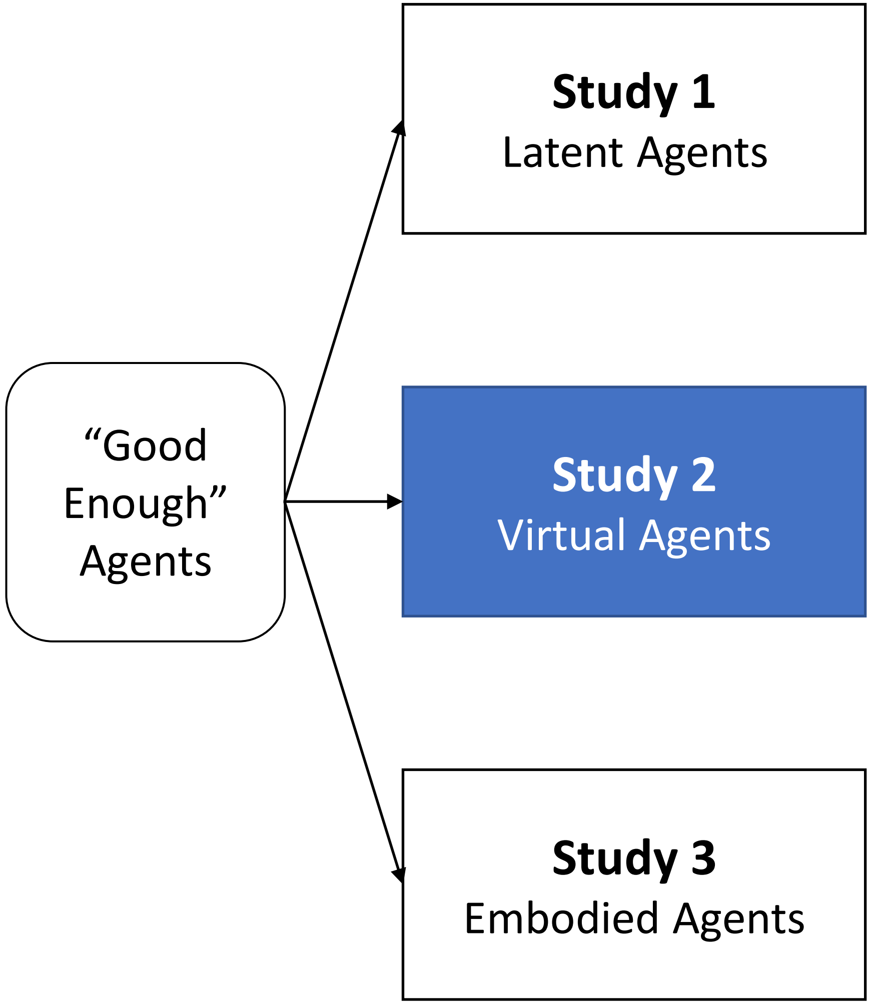
One exception: trust. Task type may have different influences



“Good enough” agents can be imperceptible, yet bring benefits

Could high reliability have minimal returns?

Dissertation Comparative Analysis



- "Good Enough" agents can help users exercise their judgment more often in influence tasks with decision support systems.
- The "gold" standard (explanations, control) could be "silver" instead.

- "Good Enough" agents can contribute to improved situation awareness for behavioral predictions.
- Trust positively affected situation awareness; a new insight from non-supervisory control tasks.

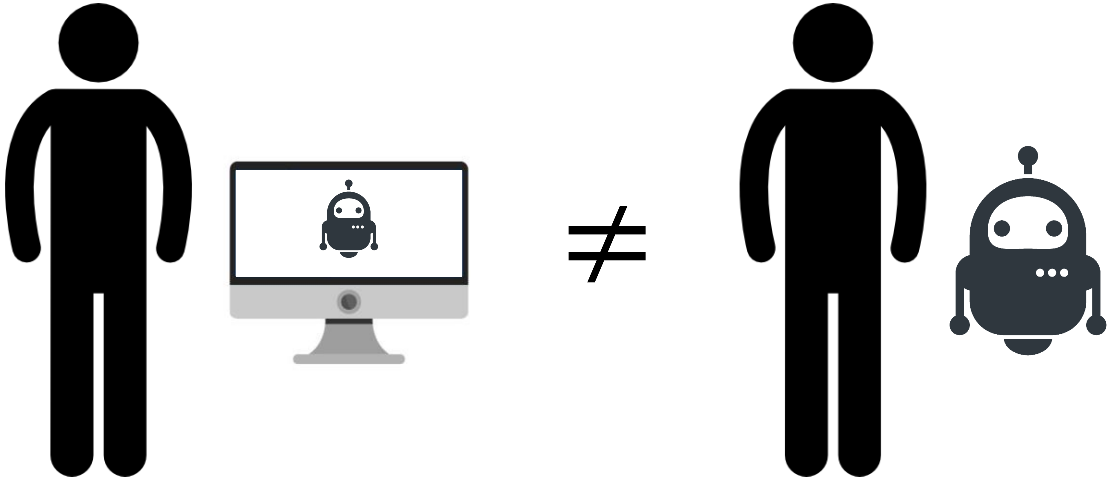
Study 1: Latent Agents

Study 2: Virtual Agents

Study 3: Embodied Agents

Differences in virtuality vs. physicality

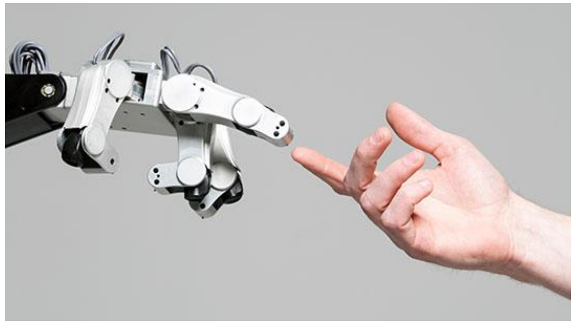
- Interaction with agents can be virtual or physical
- Physical interactions subject to social embodiment[a]
  - Human responses can be different[b]
- HAI research often researches tasks with virtual agents (see study 1 and 2)



a K. M. Lee. Presence, Explicated. Communication Theory, 14, 1. 2004. **b** L. W. Barsalou, P. M. Niedenthal, A. K. Barbey, and J. A. Ruppert. Social Embodiment. 2003.

HRI and feasibility

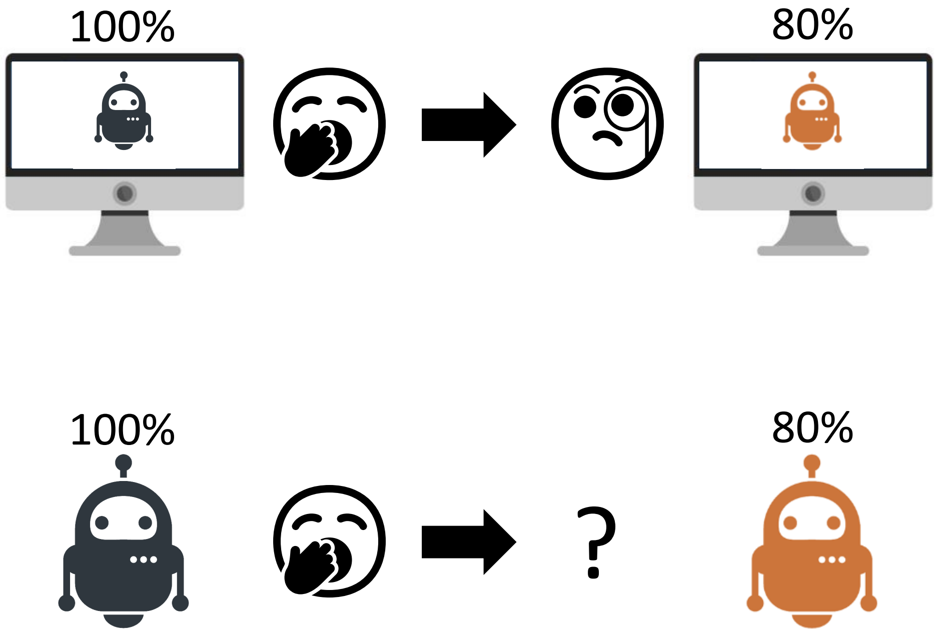
- Physical agents and trust researched through human-robot interaction
    - Very limited: expensive and needs guarantees of safety
- Recent alternative: Virtual Reality[a]
    - Simulate any robot and related interactions
- Research on trust in VR upcoming



[a] O. Liu, D. Rakita, B. Mutlu, and M. Gleicher. Understanding Human-Robot Interaction in Virtual Reality. International Symposium on Robot and Human Interactive Communication. 2017.

Embodiment and reliability interactions

- Can our previous reliability findings replicate with embodied agents?
- We can’t really build or order a robot for this – use VR instead



Study 3: Embodied Agents

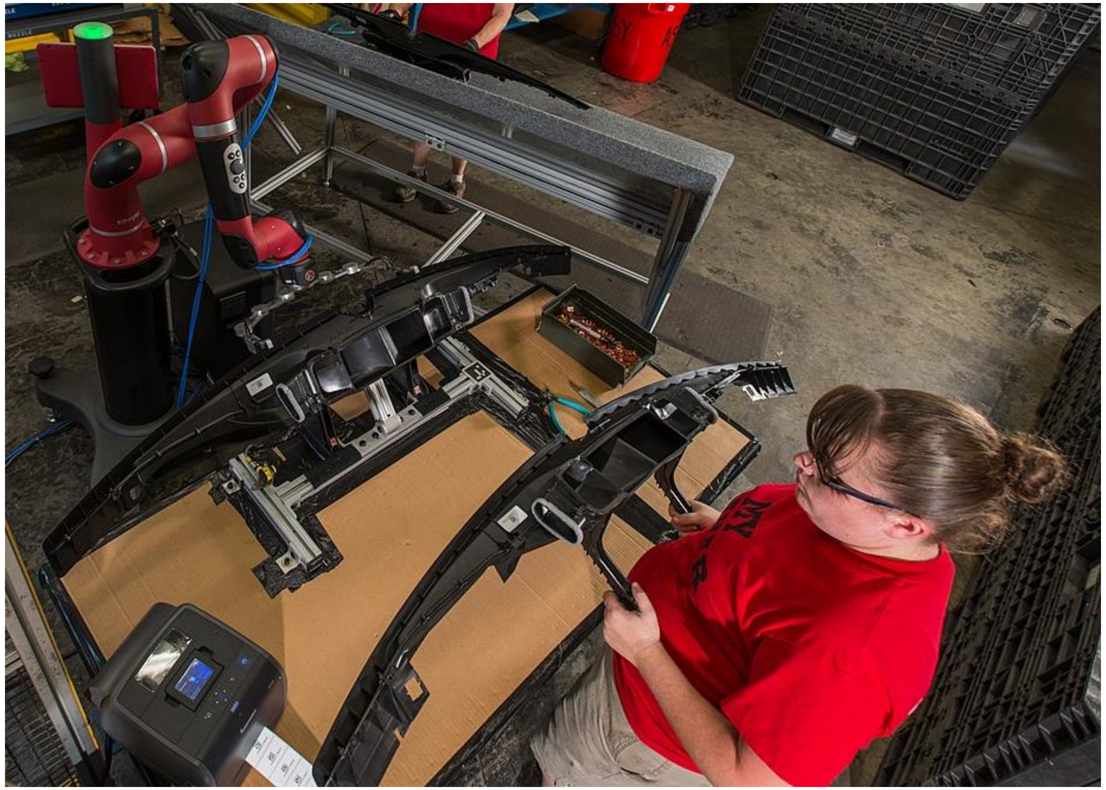
Research Questions

How does embodiment affect performance and trust towards an agent in a teamed physical task?

How does agent reliability interact with embodiment to affect human performance and perception?

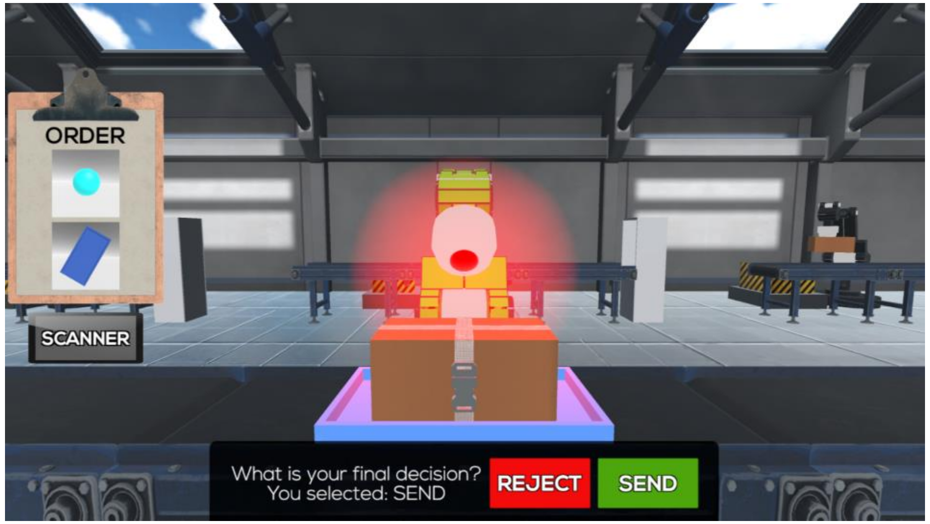
Task Context

- Common use cases for collaborative robotics (cobots)
  - Assembly, processing, packaging, un/loading, inspections
- Expectation: high reliability, human rarely in the loop (mostly automation)
- Introduce collaborative decision-making b/w humans and agents

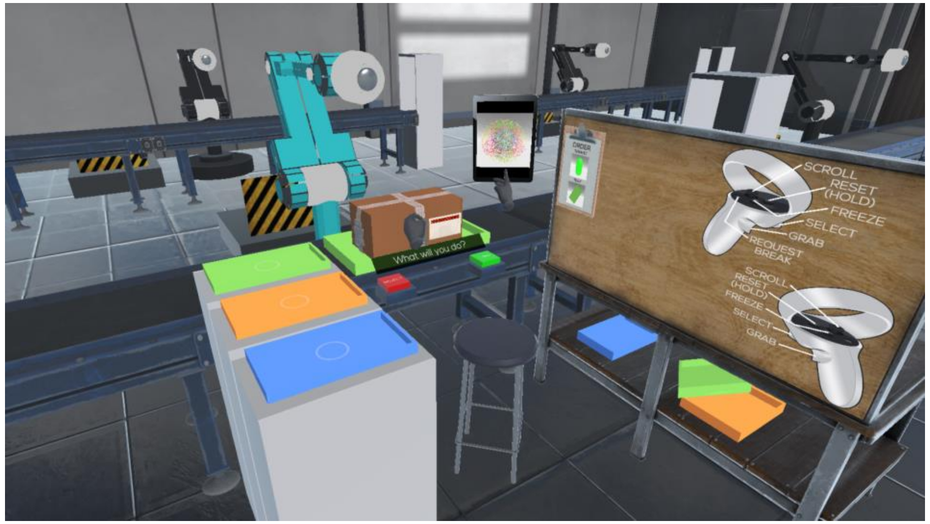


Task Selection: Warehouse

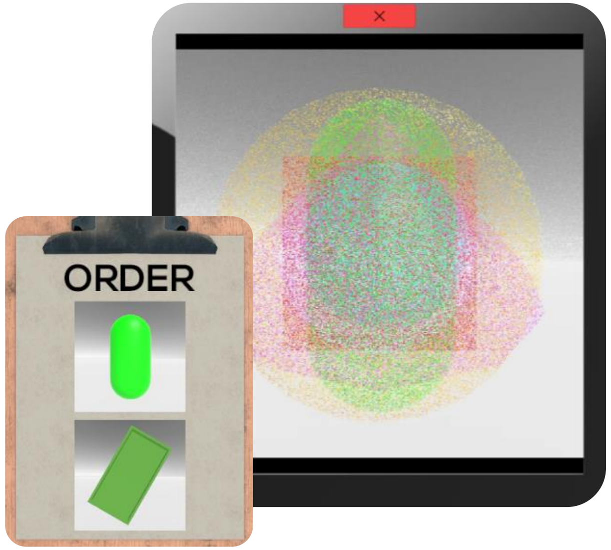
- Human and robot agent collaborate in signal detection theory-based game
  - Goal: Send out good packages, reject bad packages; as quick as possible
  - Robot gives a recommendation (package good or package bad – reliability varies within-subjects)
- 2 versions
  - Screen – telepresent agent
  - Virtual Reality – copresent agent



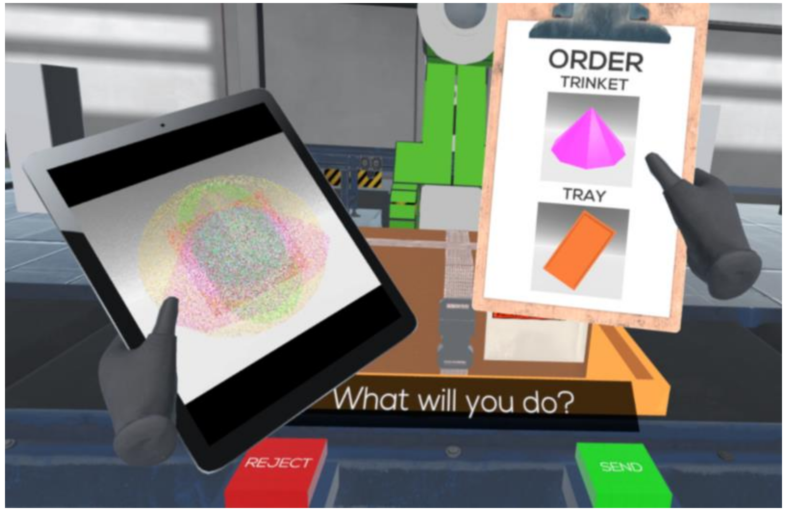
What is your final decision?
You selected: SEND



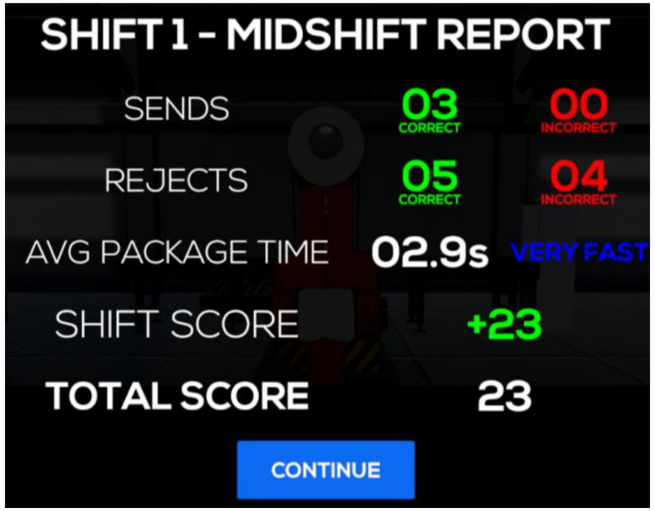
Task Features Explored



Signal Detection Theory
(w/ time pressure)
Given noisy feed, find target object
Adds uncertainty and reliance on robot recommendation
Asymmetric rewards



Embodied Interactions
VR participants could grab and interact with objects in the environment
Different levels of workload can incite distinct trust calibration


SHIFT 1 - MIDSHIFT REPORT

SENDS 03 CORRECT 00 INCORRECT

REJECTS 05 CORRECT 04 INCORRECT

AVG PACKAGE TIME 02.9s VERY FAST

SHIFT SCORE +23

TOTAL SCORE 23

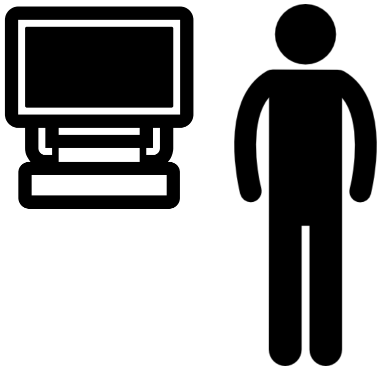
CONTINUE

Sporadic Feedback
Participants would receive performance feedback at middle and end
Allows for discrete calibration points and time-based analyses

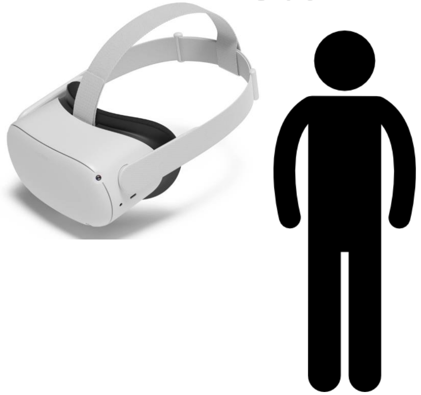
Independent Variables

Representation
(2 levels)

Screen



VR



Reliability
(4 levels)

Perfect
(100%)

Ideal
(91%)

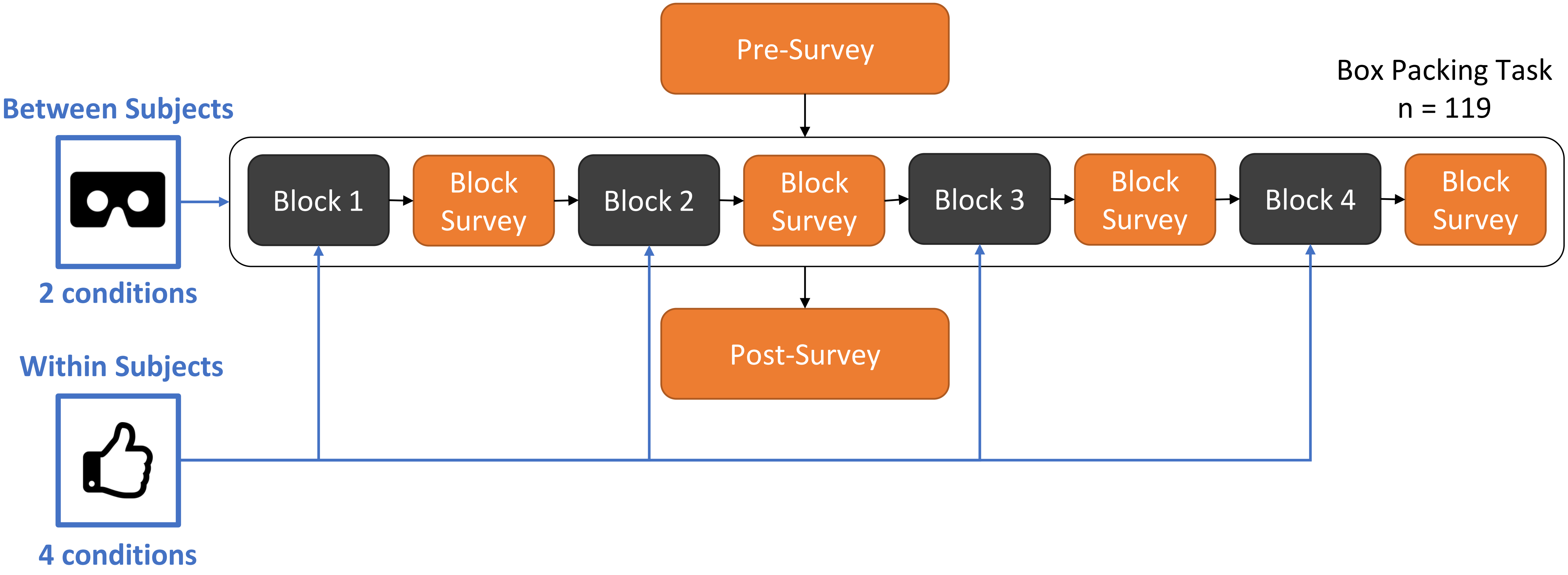
Good Enough
(75%)

No Info
(50%)

Measures

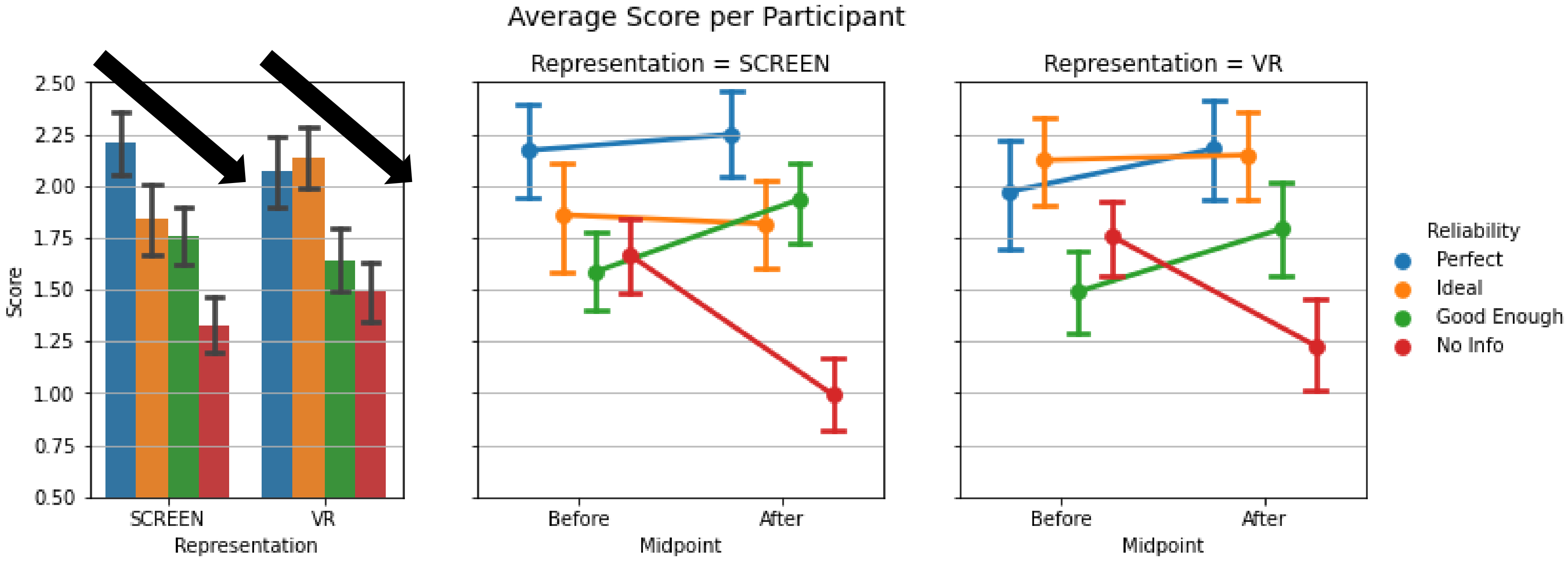
Pre-Survey	**Cognitive Reflection**, **Propensity to Trust**
Task (Behavioral)	**Accuracy**, **Reliance/Compliance**, Adherence, Switches, **Deferrals**
Block Survey	Workload, Information Preference, **Trust**, **Manipulation Check**
Post-Survey	Robot Ranking, Immersion

Study Design



Recruited in-person/through Prolific

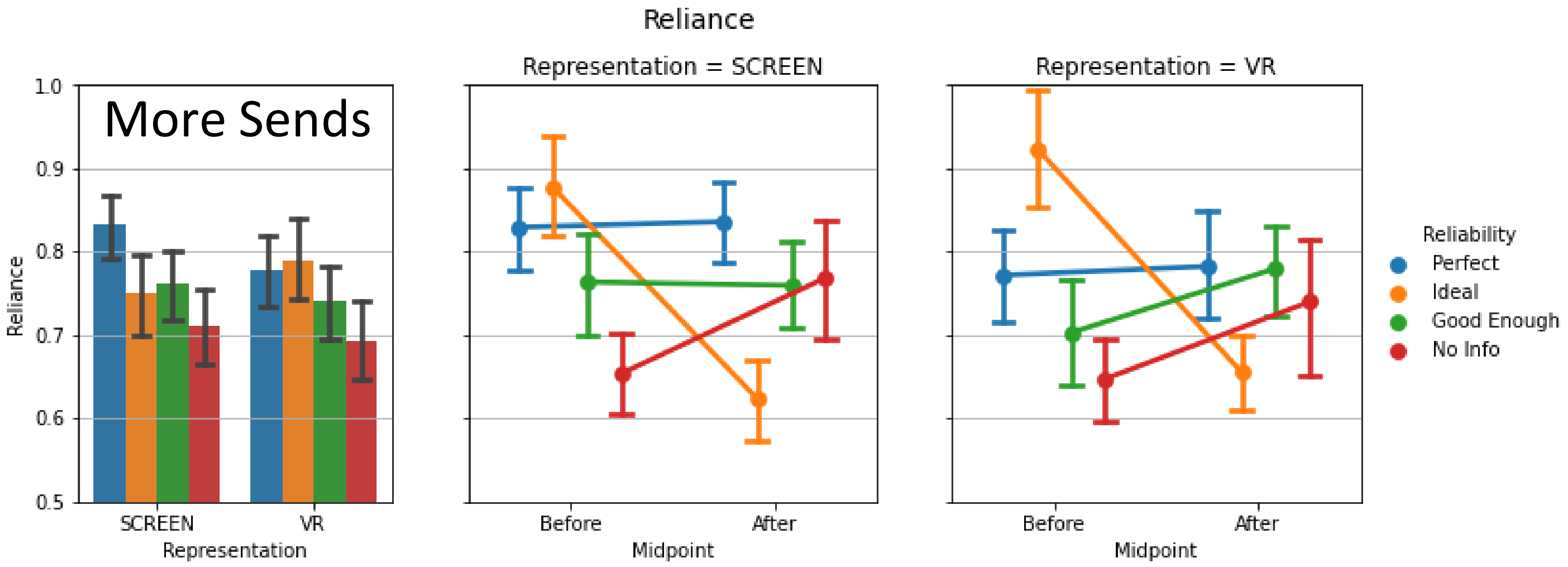
Performance scales with reliability (expected)



Performance expected to be a function of reliability ($p < 0.001$)

Best gains over time in **Good Enough** condition; not enough to outperform **Perfect** or **Ideal** ($p < 0.001$)

Decision making at thresholds

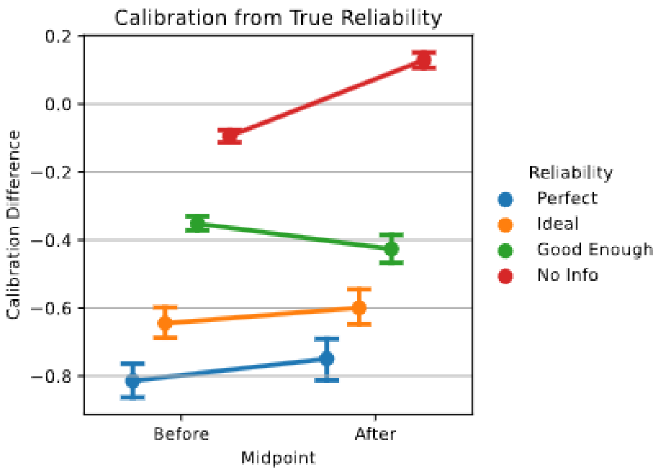


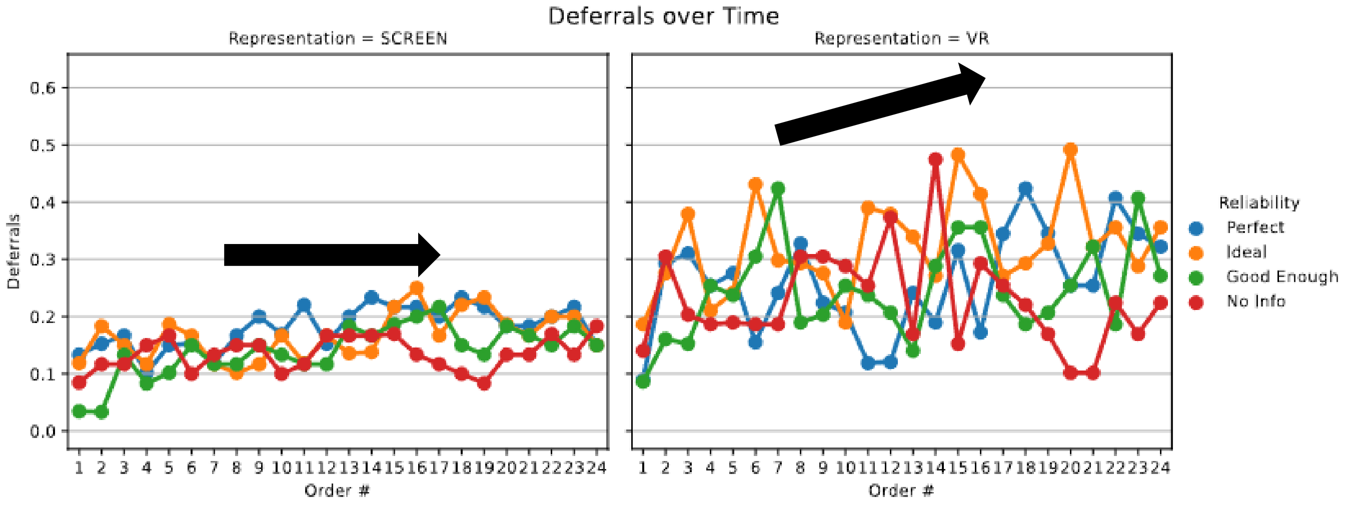
Interesting sweetspots: **1 error** or **50% errors** (p < 0.001)

Ideal: Sends decreased

No Info: Sends increased

Trust calibration improved with embodiment

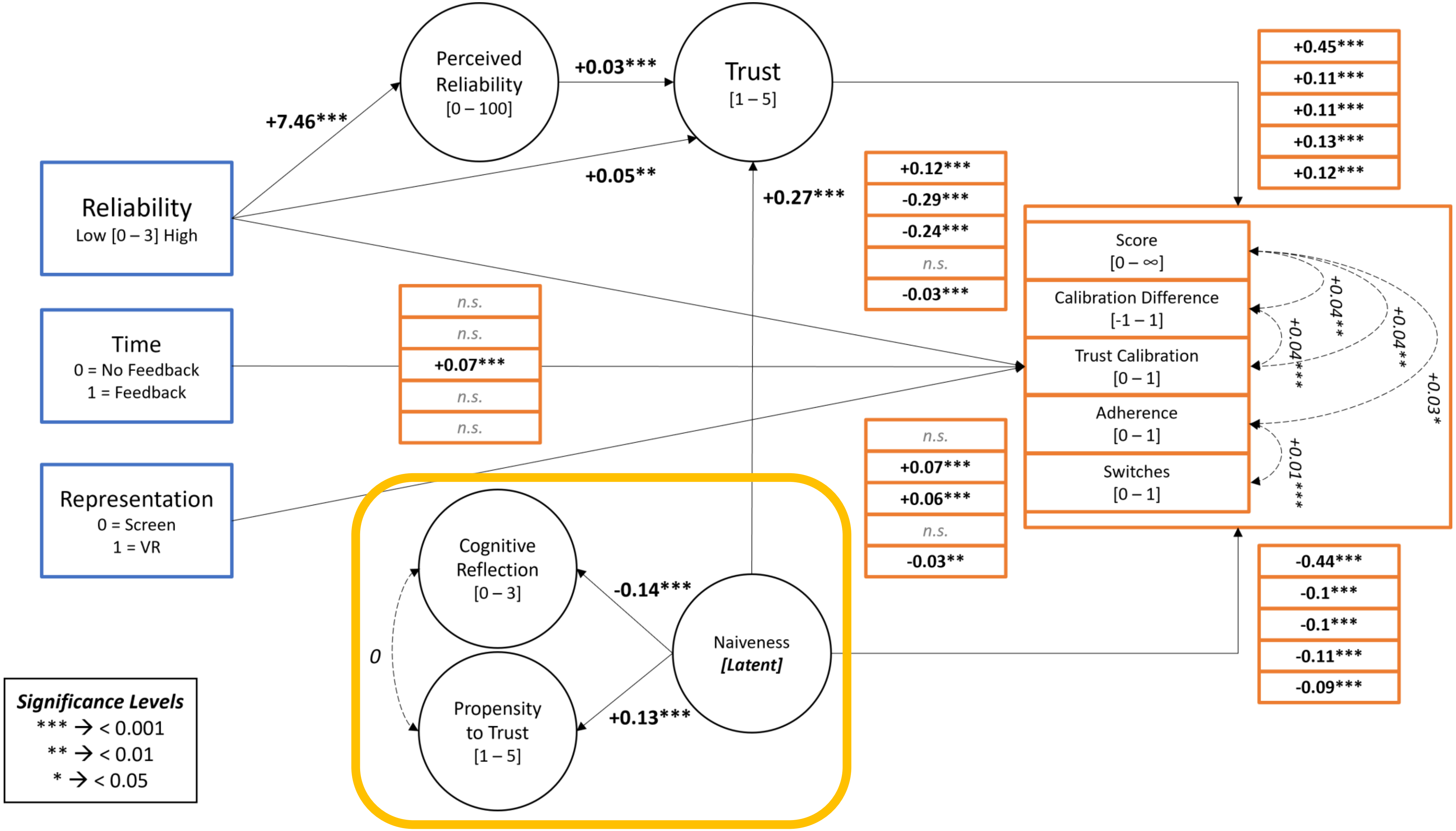




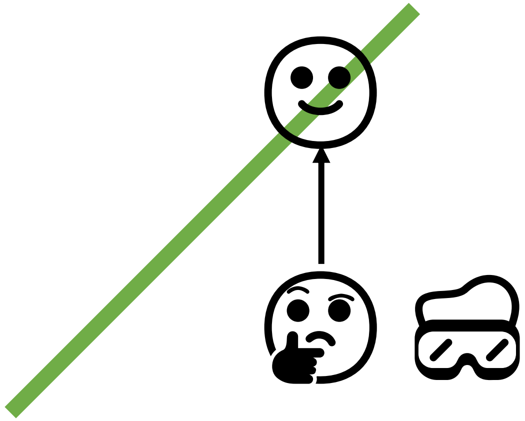
VR had higher deferrals over time ($p < 0.001$),
increasing overall trust calibration

Reliability did not affect deferrals, thus it is a
property of embodied interactions

Final SEM: Embodiment is good, “naiveness”

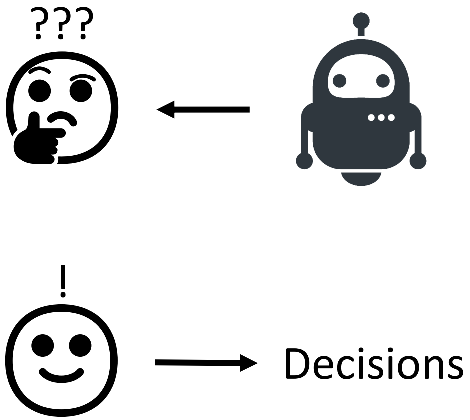


Summary



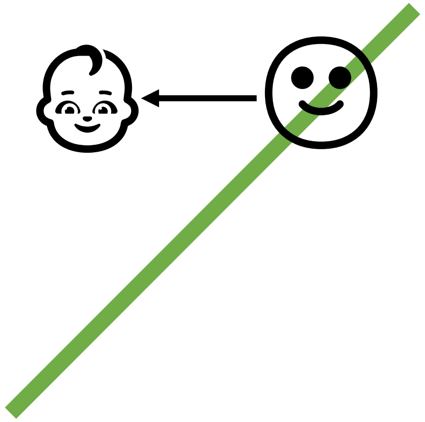
Agent framing is important to establish a baseline of trust allocation

Embodiment allows for improved trust calibration following this allocation



Humans can determine reliability with little to no feedback, even in highly uncertain scenarios

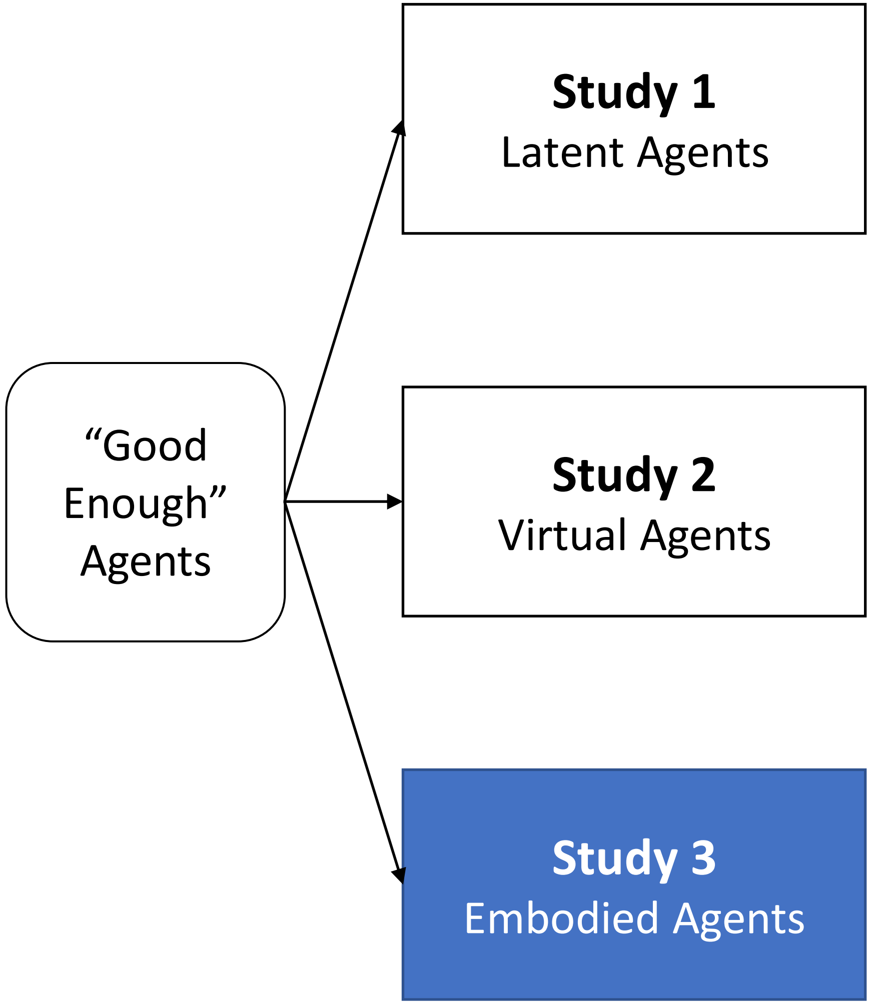
This perceived reliability can then affect decision-making



Independent differences converged into a “naiveness” factor, worsening trust calibration and performance

“Naiveness” is mediated by trust, matching prior models establishing pre-dispositions and outcomes

Dissertation Comparative Analysis



“Good Enough” Agents

Study 1
Latent Agents

- “Good Enough” agents can help users exercise their judgment more often in influence tasks with decision support systems.
- The “gold” standard (explanations, control) could be “silver” instead.

Study 2
Virtual Agents

- “Good Enough” agents can contribute to improved situation awareness for behavioral predictions.
- Trust positively affected situation awareness; a new insight from non-supervisory control tasks.

Study 3
Embodied Agents

- “Good Enough” agents can be a good calibration starting point for negative framing before feedback.
- Embodied interactions allow for more behavioral trust, information acquisition, and improved trust calibration.

Closing Research Questions

How does lowering agent reliability (i.e., “Good Enough” agents) affect task performance and human trust across different domains?

Lower reliability *benefits*: prevent knowledge loss, exploit environmental interactions, and improve trust calibration under framing.

These *benefits* can also describe performance and trust.

Closing Research Questions

What is the relationship between human individual differences, agent reliability, and technology use?

Individual differences **interplay** with SA, motivation, and workload.

These constructs affect outcomes on a domain-by-domain basis.

Categorizing domains may be a viable research approach.

Closing Research Questions

How do different task requirements interact with the human-AI interaction dynamic?

Tasks may require **time pressure, different invocation strategies, agent framing, agent representation, reliability thresholds**, with potential others.

Requirements can change how performance and trust develops.

We should consider *not* looking at IV/DVs in a vacuum (when possible).

“Good Enough” Agents: Investigating Designed Imperfections in Human-AI Teams Across Parallel Domains

Thank you!

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

UC SANTA BARBARA

DEVCOM ARMY RESEARCH LABORATORY

Next: Research Data Scientist @ Meta