Human-Centered Explainable AI

XAISS 2022

Ujwal Gadiraju

@UJLAW


1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

TODAY'S 4-COURSE MEAL

- Introduction to Human-Centered AI
- HCXAI
- Trust in Human-AI Interaction
- Evaluation of HCXAI Systems

TECHNOLOGICAL REVOLUTIONS THROUGH AI

- Transportation
- Health
- Finance
- Education
- Manufacturing
- …


126 mi
64
MPH
65
55


TECH
Tesla's Autopilot thinks the moon is a big old traffic light in the sky
When the moon hits your AI like a big pizza pie... that's a cue to slow down?

French Tax Collectors Use A.I. to Spot Thousands of Undeclared Pools

Algorithms combing through satellite photos found over 20,000 unreported swimming pools in a few regions, yielding an expected $10 million in taxes, and the system will soon go nationwide.

WILL A ROBOT STEAL **YOUR** JOB?

KILLER COMPUTERS

Bill Gates warns 'dangerous AI' poses a threat 'like nuclear weapons'

AI WARNING:

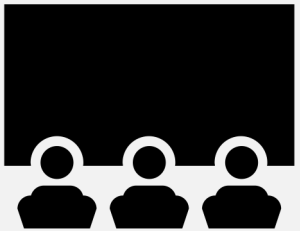
Robots will destroy a HUGE number of jobs, claims expert

AI could be used to ***TAKE OVER*** the ***WORLD*** through 'evil' fake news and hijacking cars

The AI Narrative is a Seesaw of Extremes

- AI Dreams : Startling advances

- AI Nightmares : Terrifying possibilities



Is there a third perspective?

HUMAN-CENTERED AI

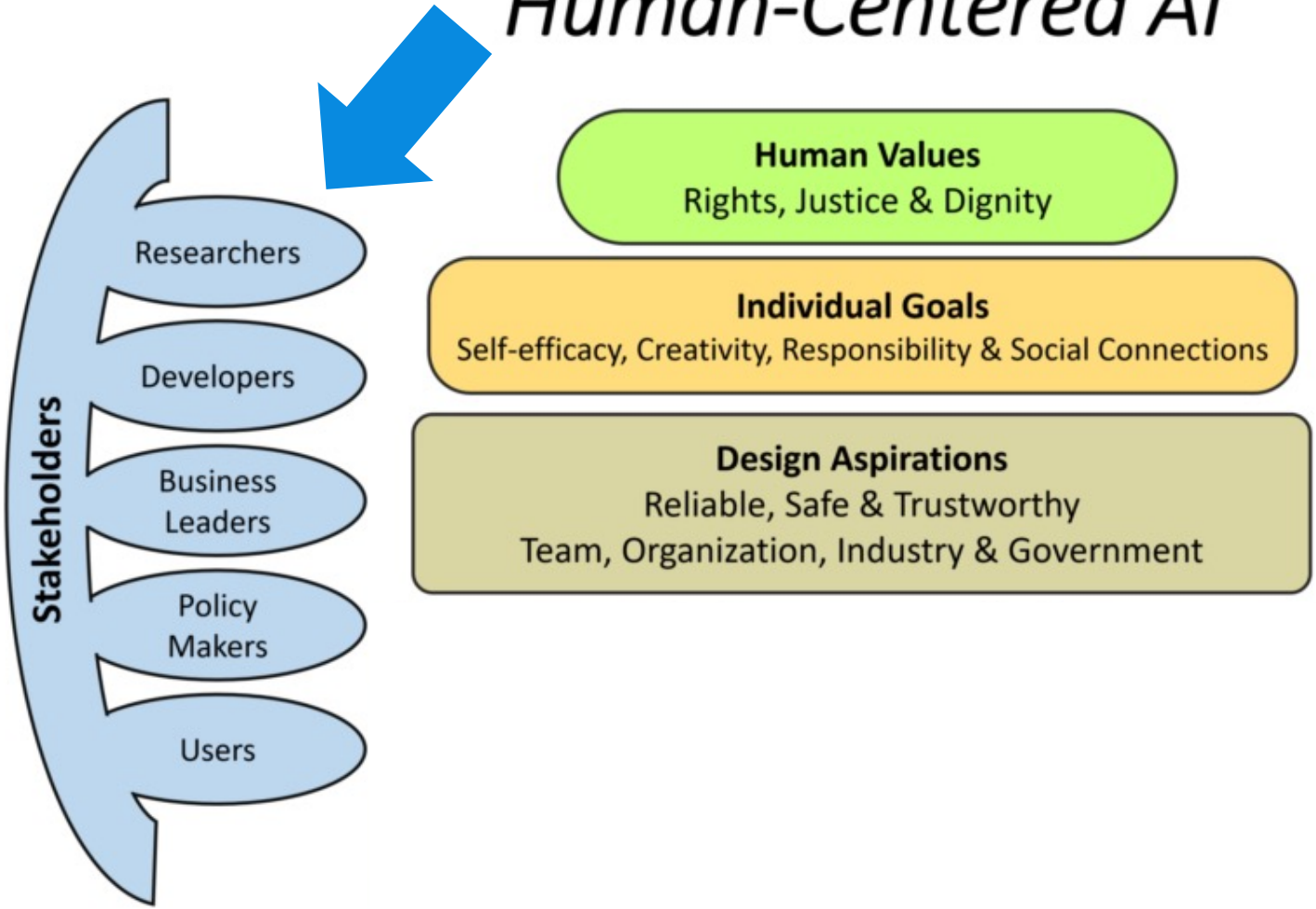
… AI that can amplify human abilities, empower people, augment, and enhance human experiences …

Human-Centered AI

Human-Centered AI

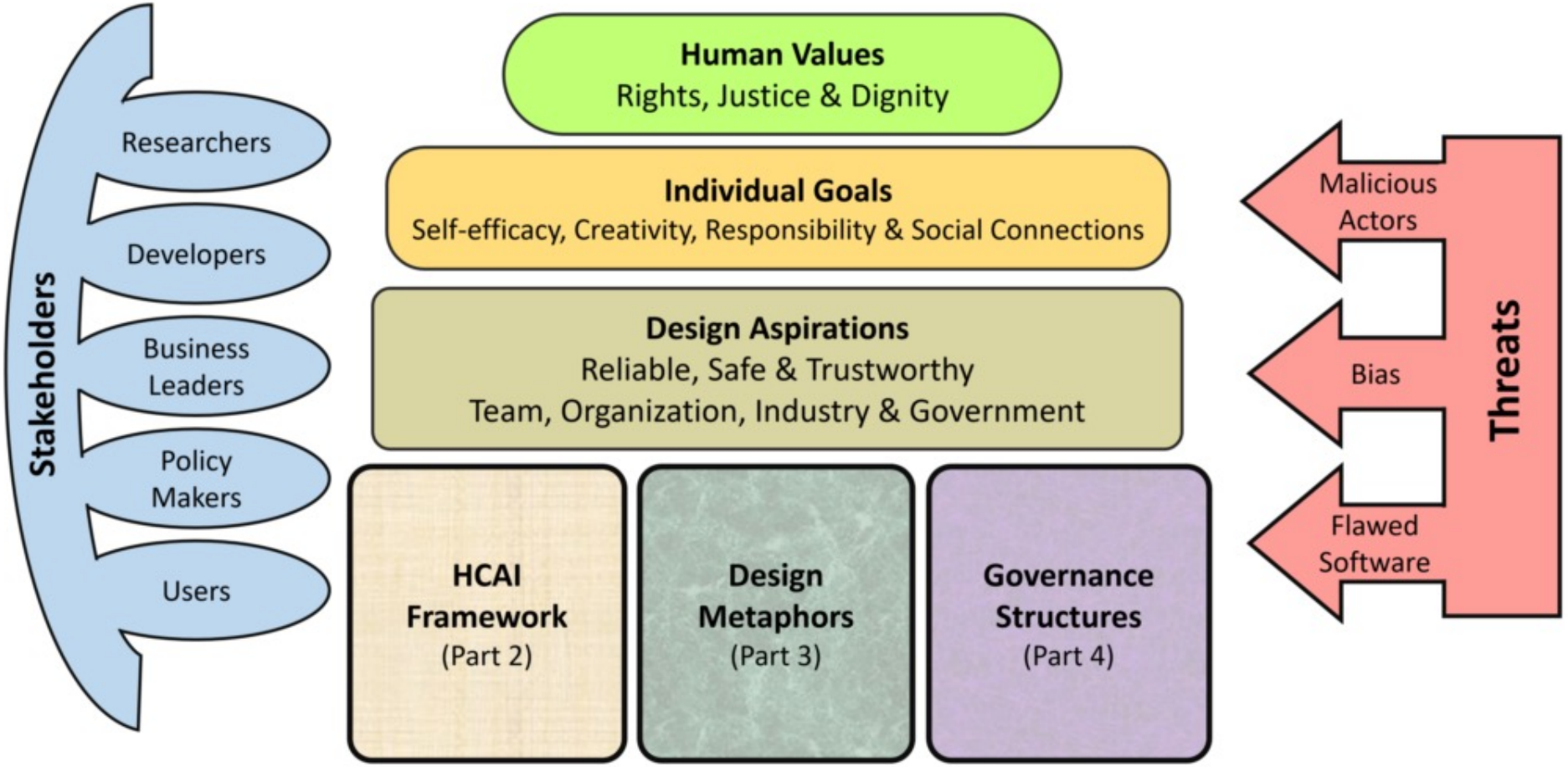
Human-Centered AI

Human-Centered AI



Human-Centered AI

Human-Centered AI

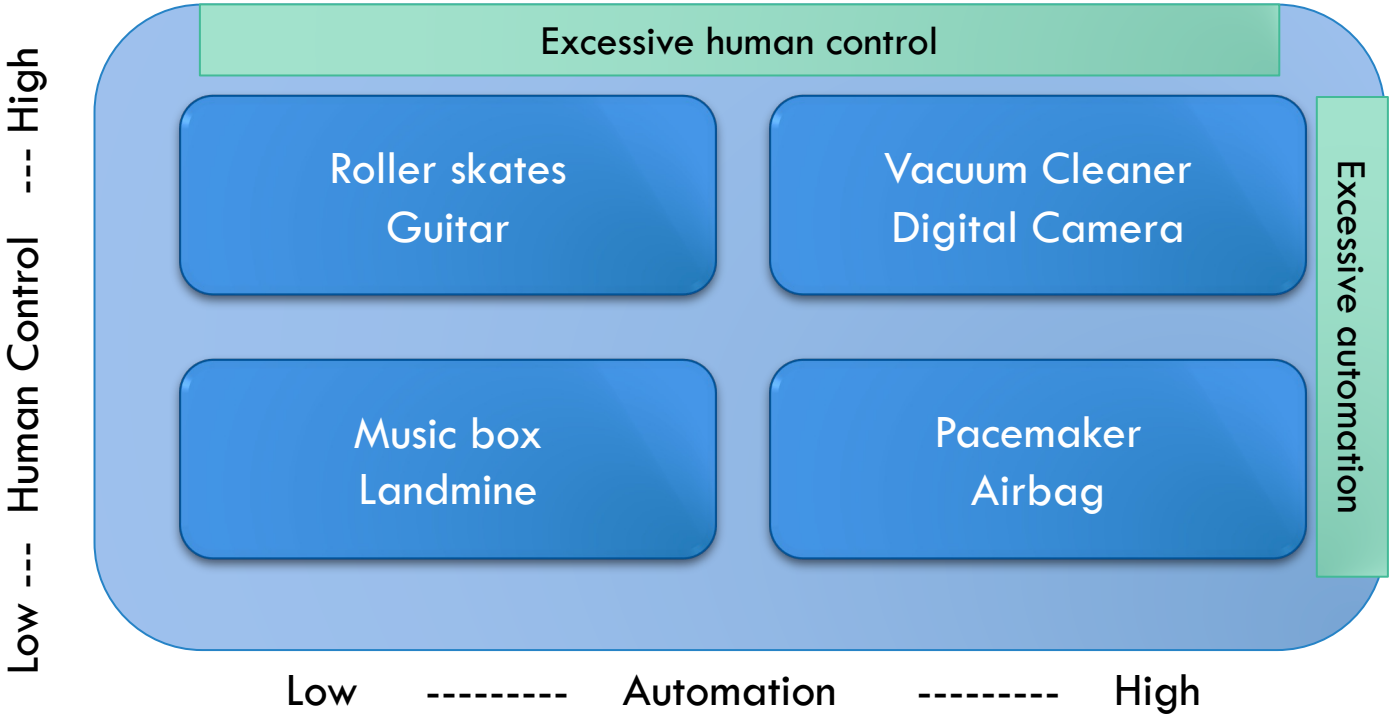


https://www.academicfringe.org

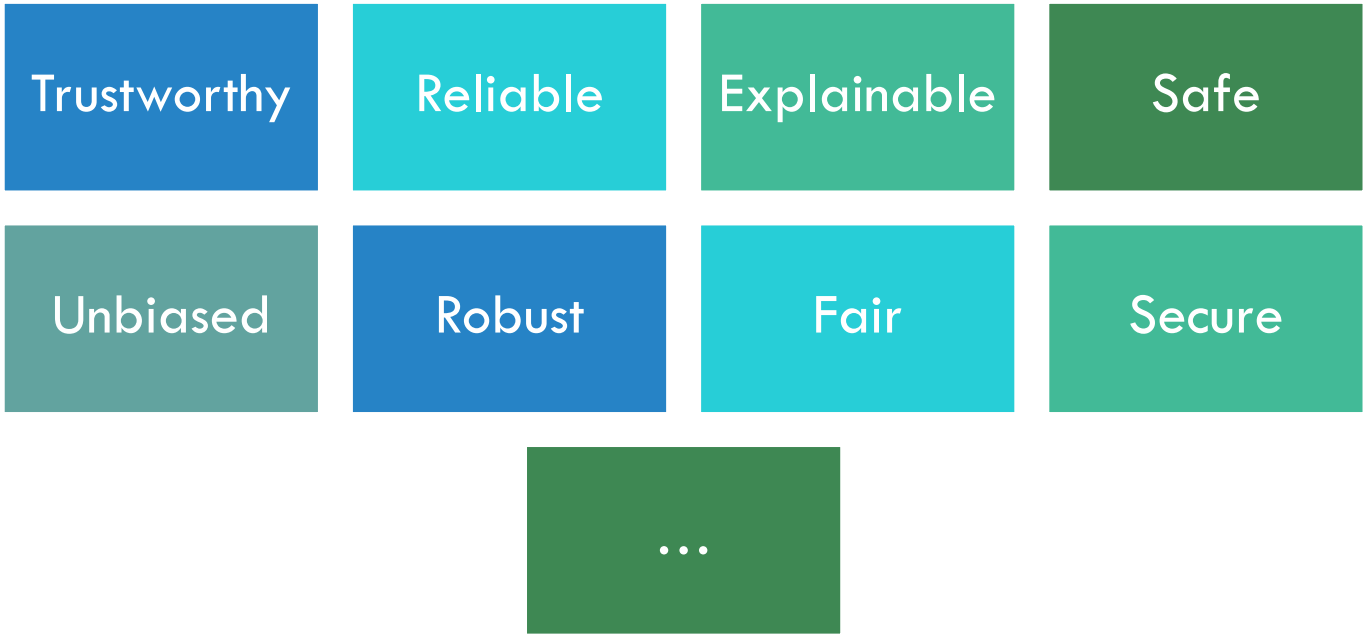
- First edition on “Crowd Computing & Human-Centered AI”
- Second edition on “Responsible Use of Data”
- Third edition on “Designing at Scale with Human-AI Collaboration”

THE BALANCING ACT OF HUMAN CONTROL

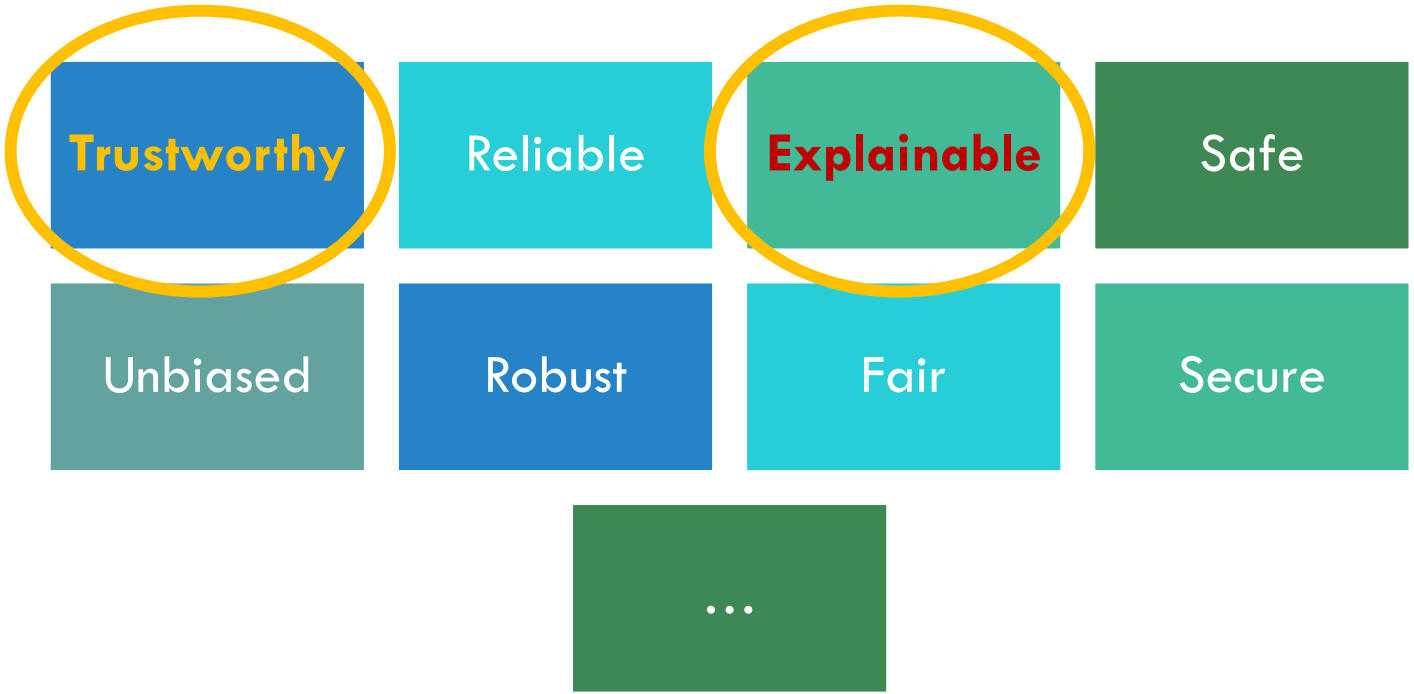
Credits: Ben Schneiderman



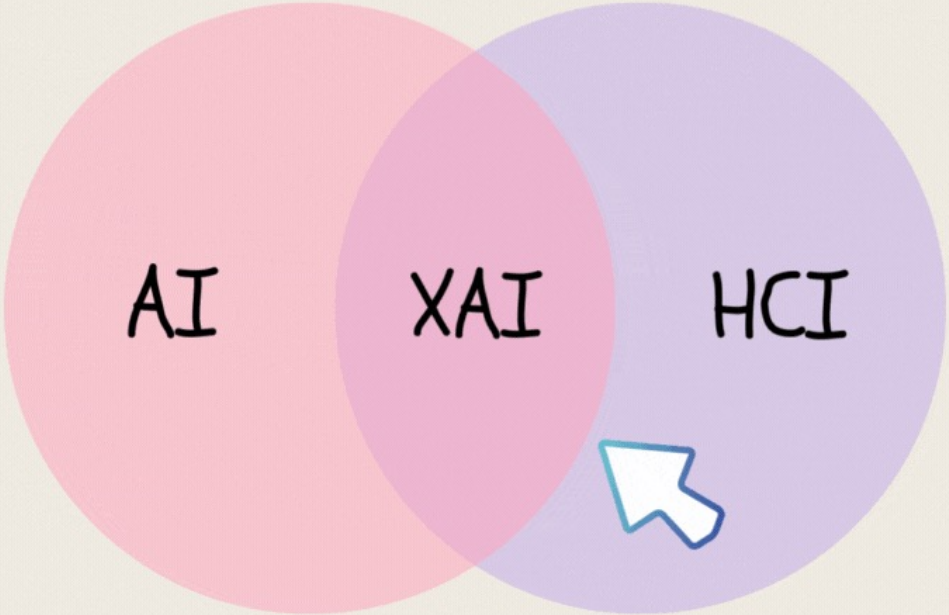
HCAI ATTRIBUTES



HCAI ATTRIBUTES



THIS IS WHY WE'RE ALL HERE



ALGORITHMIC LENS

In this “golden age” of AI most work has focused on explainability from an algorithmic perspective.

HUMAN LENS

Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach

Upol Ehsan and Mark O. Riedl

Georgia Institute of Technology
Atlanta, GA 30308, USA
ehsanu@gatech.edu, riedl@cc.gatech.edu

- An approach that puts the human at the center of technology design
- Critical Technical Practice (pioneered by AI researcher Phil Agre, 1997)

- Understanding "*who*" the human is
  - Interplay of values
  - Interpersonal dynamics
  - Socially situated nature of AI systems

WHAT'S THE BEST EXPLANATION, ANYWAY?

- Human-Human interactions ?
- Human-AI interactions ?

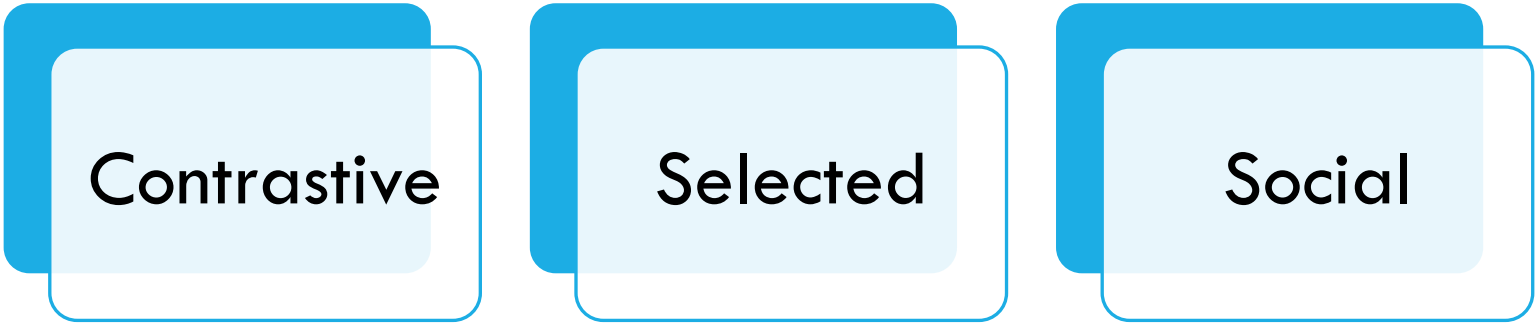
➔ How people explain things to each other is a good place to start!

Explanation in artificial intelligence: Insights from the social sciences

Tim Miller

School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

EXPLANATIONS ARE …



Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38.

EXPLANATIONS ARE CONTRASTIVE

> The key insight is to recognise that one does not explain events per se, but that one explains why the puzzling event occurred in the target cases but not in some counterfactual contrast case. – Hilton (1990)

- Why penzai rather than bonsai?
- Can help reduce cognitive load

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin, 107*(1), 65–81. https://doi.org/10.1037/0033-2909.107.1.65

EXPLANATIONS ARE SELECTED

> Consider how the cause of death might have been set out by the physician as ‘multiple haemorrhage’, by the barrister as ‘negligence on the part of the driver’, by the carriage-builder as ‘a defect in the brakelock construction’, by a civic planner as ‘the presence of tall shrubbery at that turning’. None is more true than any of the others, but the particular context of the question makes some explanations more relevant than others. – Hanson (1965)

N. R. Hanson, Patterns of discovery: An inquiry into the conceptual foundations of science, CUP Archive, 1965.

EXPLANATIONS ARE SOCIAL

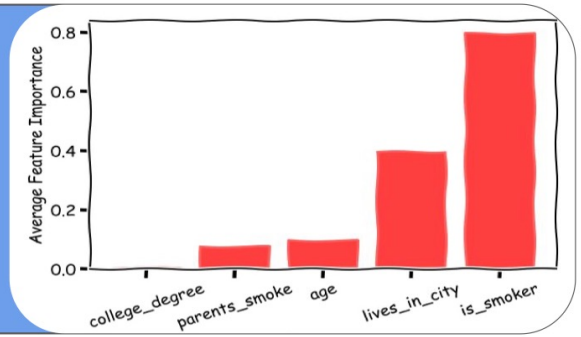
> Causal explanation is first and foremost a form of social interaction. One speaks of giving causal explanations, but not attributions, perceptions, com- prehensions, categorizations, or memories. The verb "to explain" is a three-place predicate: someone explains something to someone. Causal explanation takes the form of conversation and is thus subject to the rules of conversation.
>
> – Hilton (1990)

Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2019). A grounded interaction protocol for explainable artificial intelligence. *arXiv preprint arXiv:1903.02409*.

INTERACTIVE EXPLANATIONS ➔ WATCH THIS SPACE!

For women predicted high risk for lung cancer that are older than 65, why did the model decide to predict them as high risk?

Good question! It looks like the model predicted these individuals as high risk mostly because they were smokers but also because they live in large cities. I'm highly confident these are the reasons because the explanations have high fidelity. Here's the average feature importance for these people (higher means more important).



Wow, it's surprising that whether the person lives is in a city is so important.

Yes, lives_in_city has a significant effect on the predictions for these individuals. Perturbing this feature can flip the prediction for 4 of 15 of the instances in this group.

Lakkaraju, H., Slack, D., Chen, Y., Tan, C., & Singh, S. (2022). Rethinking Explainability as a Dialogue: A Practitioner's Perspective. *arXiv preprint arXiv:2202.01875*.

WHY DO WE NEED TO THINK ABOUT TRUST?

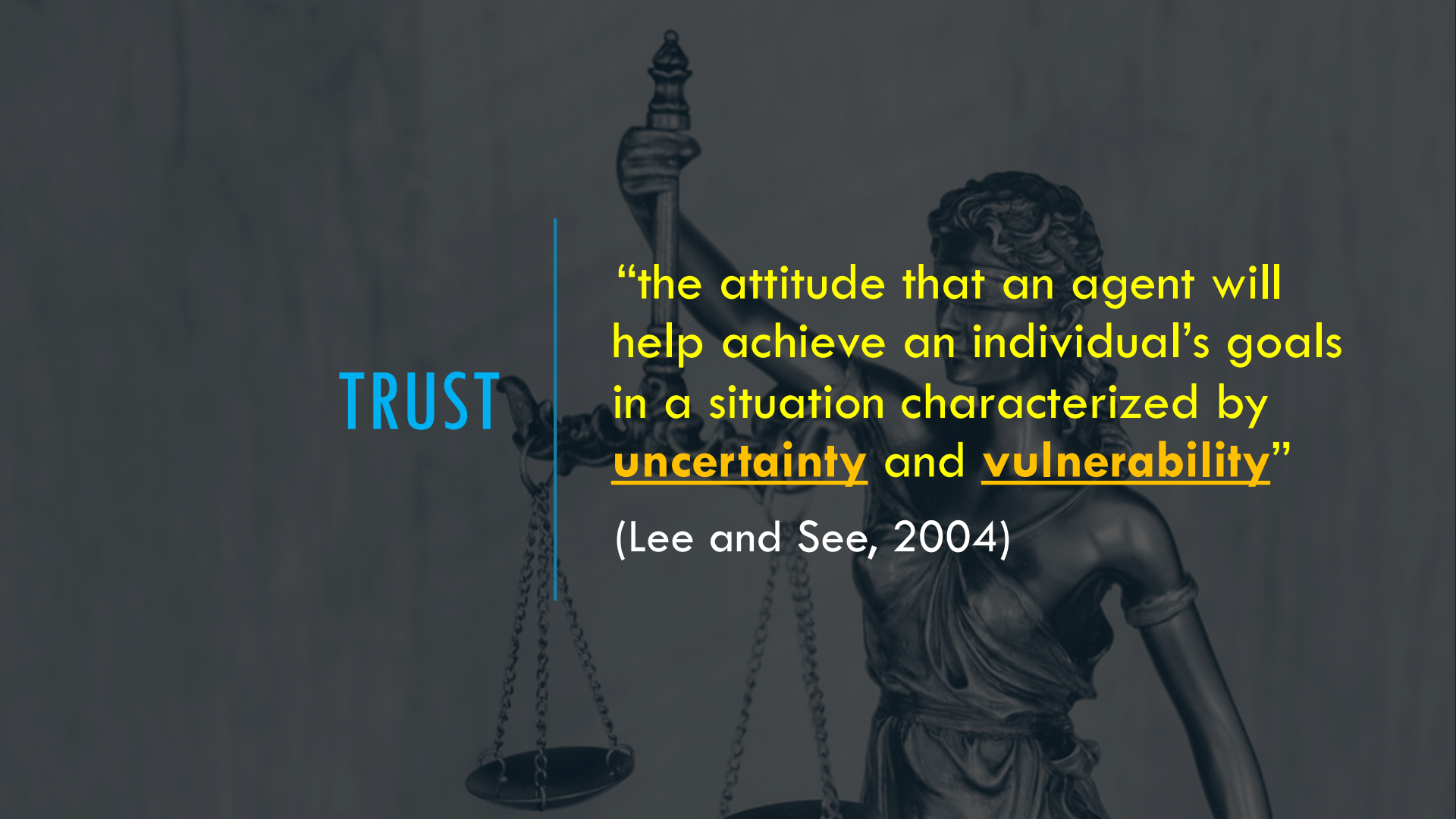
Trust plays a central role in Human-AI interaction

The end goal is NOT trust itself. Trust is a mechanism to help enable predictability and collaboration.

between people

- Obtaining trust in a machine → makes it easier to anticipate machine decisions (i.e., predictability) → human-machine collaboration

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021, March). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 624-635).



TRUST

"the attitude that an agent will help achieve an individual's goals in a situation characterized by **uncertainty** and **vulnerability**"

(Lee and See, 2004)

TRUSTWORTHY AI

- Trust does not imply trustworthiness
- Trustworthiness does not imply trust
- Warranted and Unwarranted trust

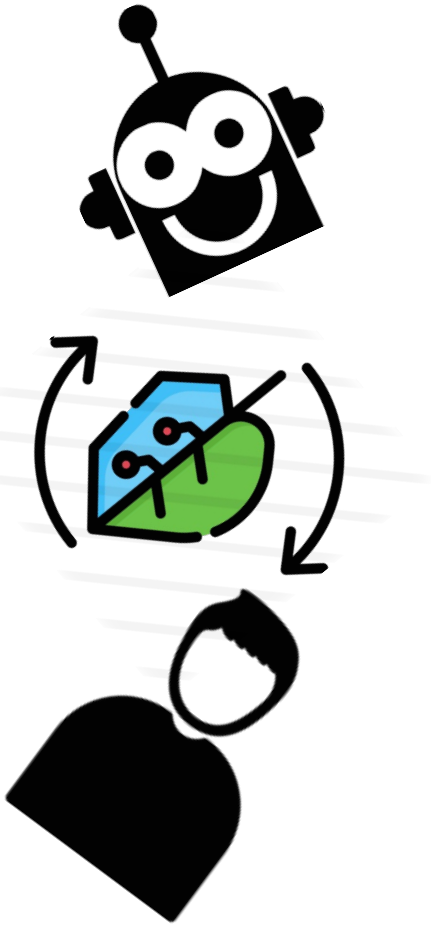
Desirable outcomes of trust

- Warranted trust and distrust
- Avoid unwarranted trust and distrust

RELIANCE ON AI SYSTEMS

EVALUATION IN XAI (1/3)

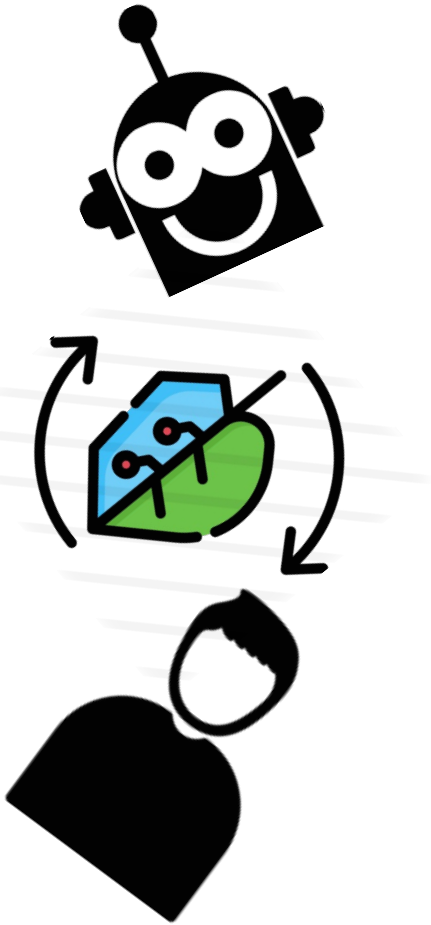
Experiments that only ask the users whether they trust the model, with no vulnerability evaluate neither trust nor trustworthiness!



EVALUATION IN XAI (2/3)

Main methods to evaluate trust:

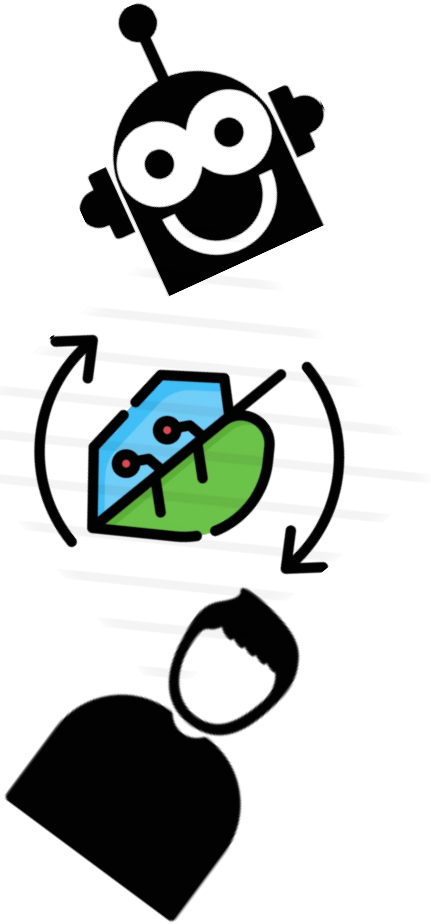
- ➔ Subjective ratings
- ➔ Qualitative interviews
- ➔ Questionnaires
- ➔ Reliance



EVALUATION IN XAI (3/3)

Guidelines for XAI Evaluation:

1. Use tasks that have correspond to clearly measurable performance
2. Ensure there is some vulnerability to risks (e.g., using games, rewards, or incentive schemes)
3. Use reliance as a/at least one measure for trust
4. Manipulate “trustworthiness”



ACM UMAP 2021

Second Chance for a First Impression?
Trust Development in Intelligent System Interaction

TRUST IN HUMAN-COMPUTER INTERACTION

Different **components** (Hoff & Bashir, 2015):

- Dispositional trust
- Situational trust
- **Learned trust**

Accuracy influences trust formation (Beggatio & Krems, 2013)

- Research gap on **trust formation over multiple sessions**

RESEARCH QUESTIONS

- How does a user's trust in an intelligent system with **varying accuracy** evolve over **multiple interactions**?
- How does the accuracy of an intelligent system mediate user **trust breakdown and recovery**?
- How do **dispositional factors** affect trust formation and evolution in an intelligent system?

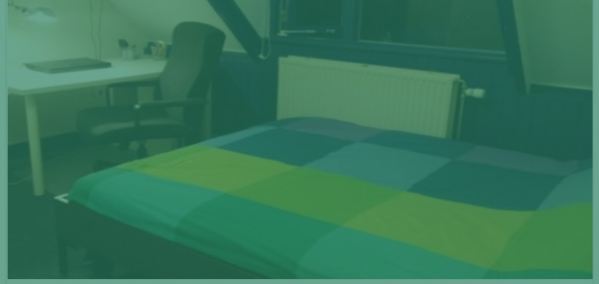
Jan

HOUSING SEARCH

Fully Furnished Loft Studio

Oostblok, Delft

Private Room In Apartment



Shared Room In Student House

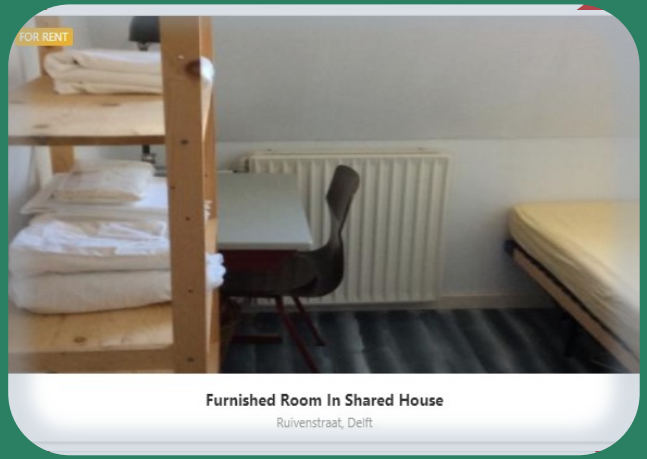
Piet Heinstraat, Delft

Furnished Room In Shared House

Ruivenstraat, Delft

Shared Room In Building

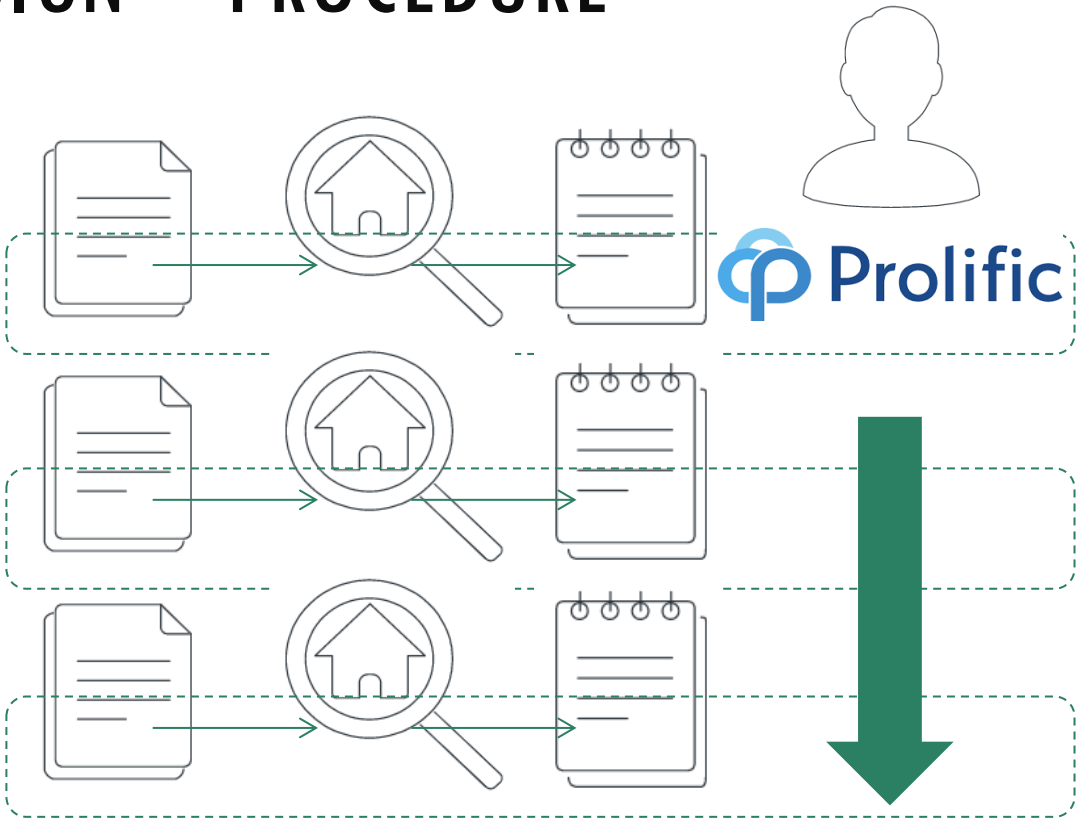
Jaagpad, Rijswijk



Furnished Room In Shared House

Ruivenstraat, Delft

STUDY DESIGN – PROCEDURE



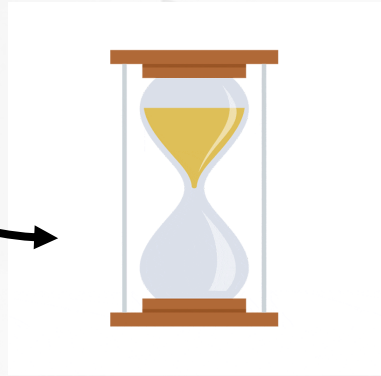
STUDY DESIGN – MEASURES

- Measuring user trust in the system
  - Trust in automation scale (Jian et al., 2000)
  - Propensity to trust scale (Frazier et al., 2013)
- Measuring affinity for technology
  - ATI scale (Franke et al., 2019)

MODELING TRUST

Trust requires three components (Hardin, 2006):

- **actors** to form trust
- an **incentive** to trust
- a **risk** to trust



TASK DESIGN

Easy Scenario

Peter is moving to Delft as a first year BSc. student. He is a very easy-going guy and is looking for a **shared room** which fits his rent **budget of 300€**. Further, he would require **registration at the municipality**.

Complex Scenario

Jan is a Dutch citizen moving to **Delft** for a PhD. He is looking for a **Studio Apartment** for **at least 2 years**, with a maximum **budget of 750€**. He needs his place to be **close to a supermarket** and does not mind the commute time to the university.

Find a house that meets the requirements and submit it.

Click on the house to see additional information

Go Back

Scenario

Jan is a Dutch

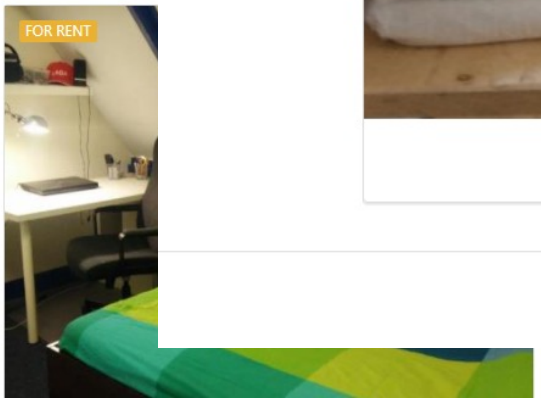
of 750 euros. He needs

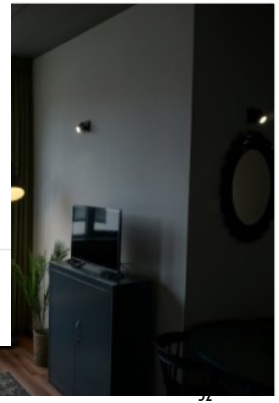
The Intelligent System searched the database for your requirements, and came up with this recommendation!

Scroll to the bottom of the screen to submit the house.

Furnished Room In Shared House

Ruivenstraat, Delft





Submit this house

Usage incentive

SYSTEM RELIANCE

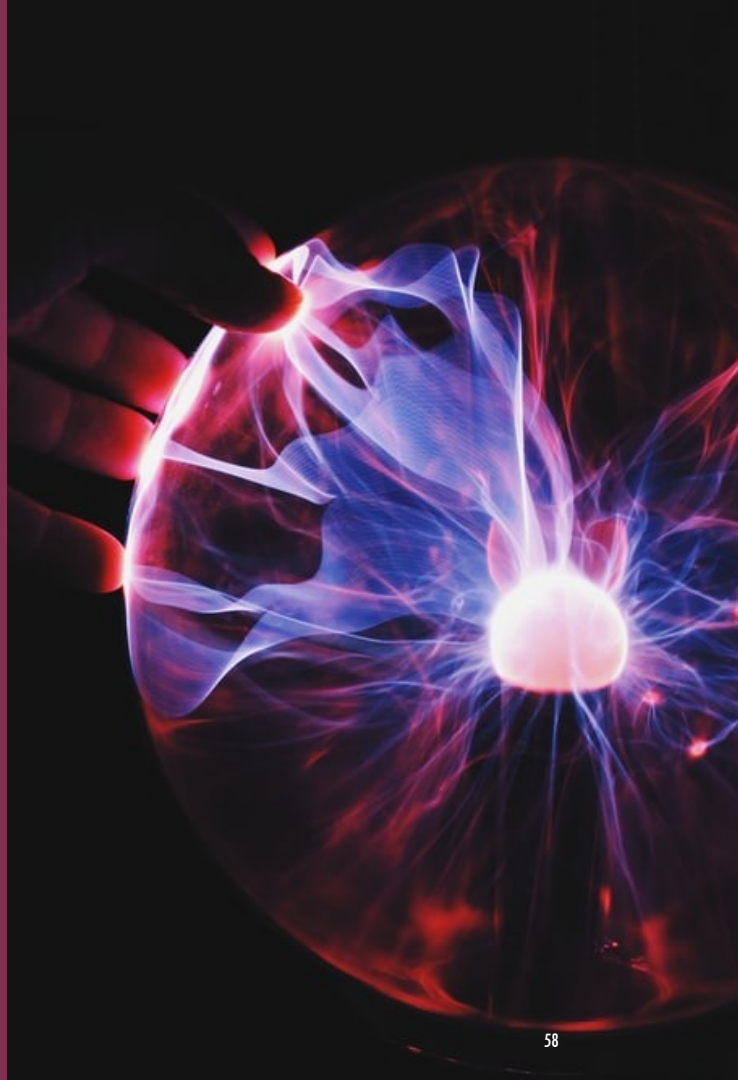
- **User performance**: 78%, 66%, 92%
  - Dependent on difficulty of scenarios
  - Dependent on system accuracy
- **System reliance**: 42% to 73%
  - Dependent on system accuracy
  - First impression matters

INFLUENCE OF DISPOSITIONAL FACTORS

- Age of participant (p=.006)
- Affinity with technology (p=.012)
- Level of education
- Country of origin
- Gender
- Propensity to trust

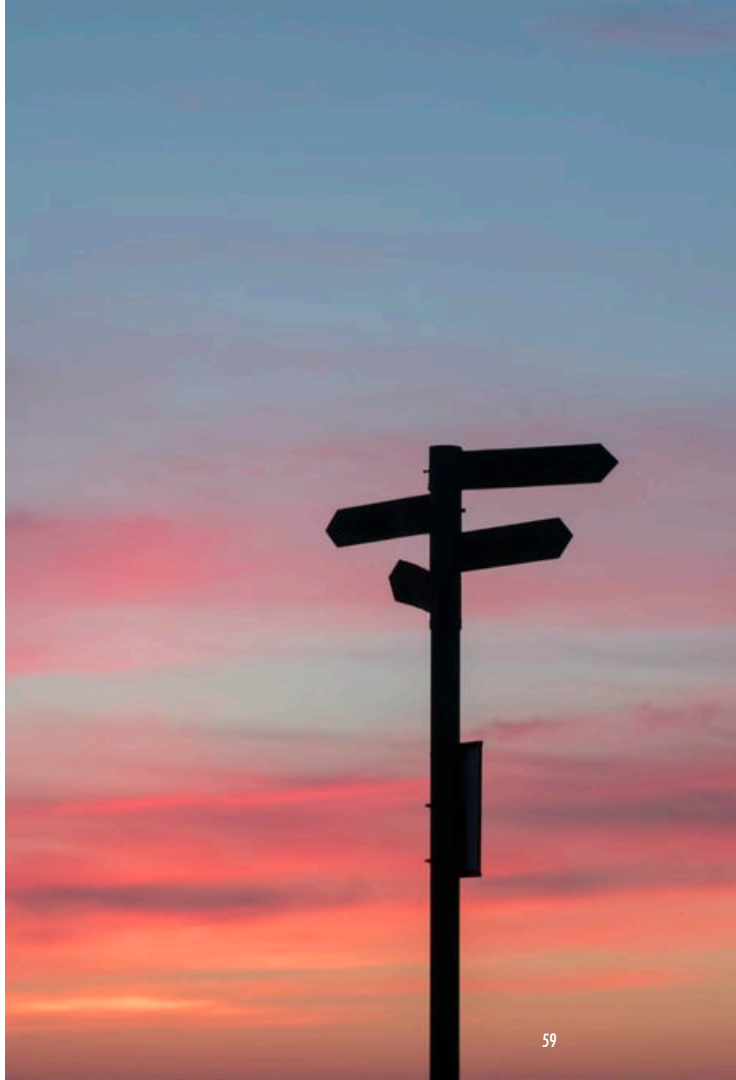
IMPLICATIONS

- Sub-par first experiences can gravely affect system adoption
- Impressions of a learning system can increase trust in the system
- Consistent system performance over time is valuable for trust
- Trust recovery mechanisms can be adapted to user dispositional traits (e.g. age, affinity with technology)



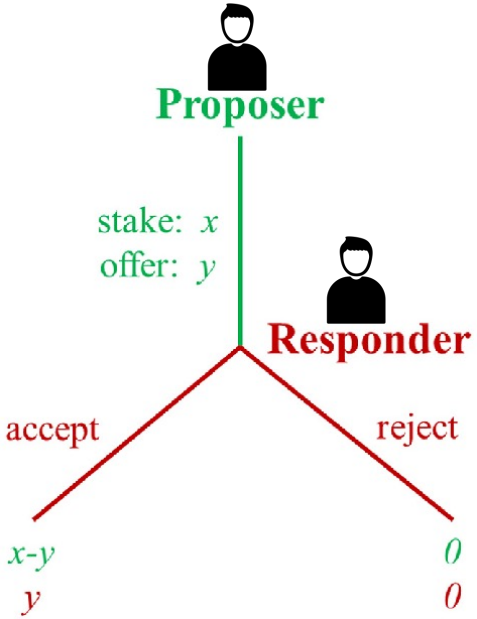
FUTURE DIRECTIONS

- How can we positively shape first impressions in Human-AI interaction?
- Trusting behaviours versus trusting beliefs to understand the affect of dispositional traits
- Degree of inaccuracy of AI systems
- Levels of anthropomorphism and system intelligence



Best Student Paper Award at AAAI HCOMP'20

IMPACT OF ALGORITHMIC DECISION-MAKING ON HUMANS : CASE OF THE ULTIMATUM BARGAINING GAME



How does dependence on decision support systems affect the behavior of human stakeholders?

- people relying on such systems to make their decisions
- people directly impacted by those decisions

- Behavioral economics framework: Ultimatum Bargaining Game

1. Perceptions of fairness change due to the introduction of ADM system in the human decision-making process
2. A better understanding of the ADM system increases cooperation among players

The effect of AI systems on human behavior in economic bargaining environments (e.g. trade, auctions, wage negotiations) …

- Surge in algorithmic decision-making
  - Algorithmic pricing
  - Automated credit-loan scores
  - Hiring decisions

Many promises of AI …

AI technologies hold great promise for market efficiency, organizations and consumer surplus

- ***Condition 1***: *People do not actively avoid AI-systems*
- ***Condition 2***: *AI-systems do not systematically err in predicting human choices*

Research Questions

1. Do humans self-select into specific interactions depending on an AI-system's autonomy?

2. Are human preferences for or against AI-systems explained by economic expectations?

3. Does the introduction of AI-systems into a bargaining environment affect overall economic welfare?

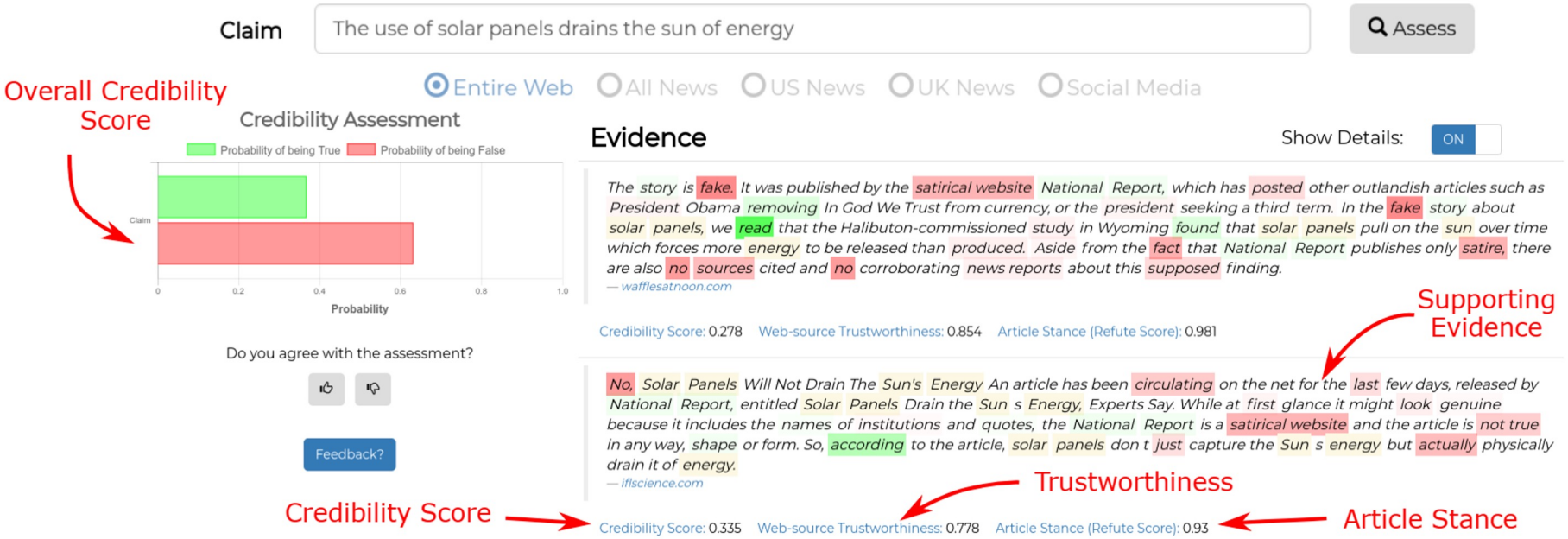
Key Findings

- **Responders prefer to approach human proposers**
- Human avoidance of AI-systems cannot be explained by their economic expectations. Responders are *more* likely to believe that the autonomous AI-system maximizes their expected income. Hence**, they appear to over-write their economic self-interest to avoid bargaining with the AI-system**
- Responders who bargain with an autonomous AI-system demand a larger share of the pie to successfully contract with the substituted human

Understanding the Role of Explanation Modality in AI-assisted Decision-making

VINCENT ROBBEMOND, Delft University of Technology, The Netherlands
OANA INEL*, University of Zurich, Switzerland
UJWAL GADIRAJU, Delft University of Technology, The Netherlands

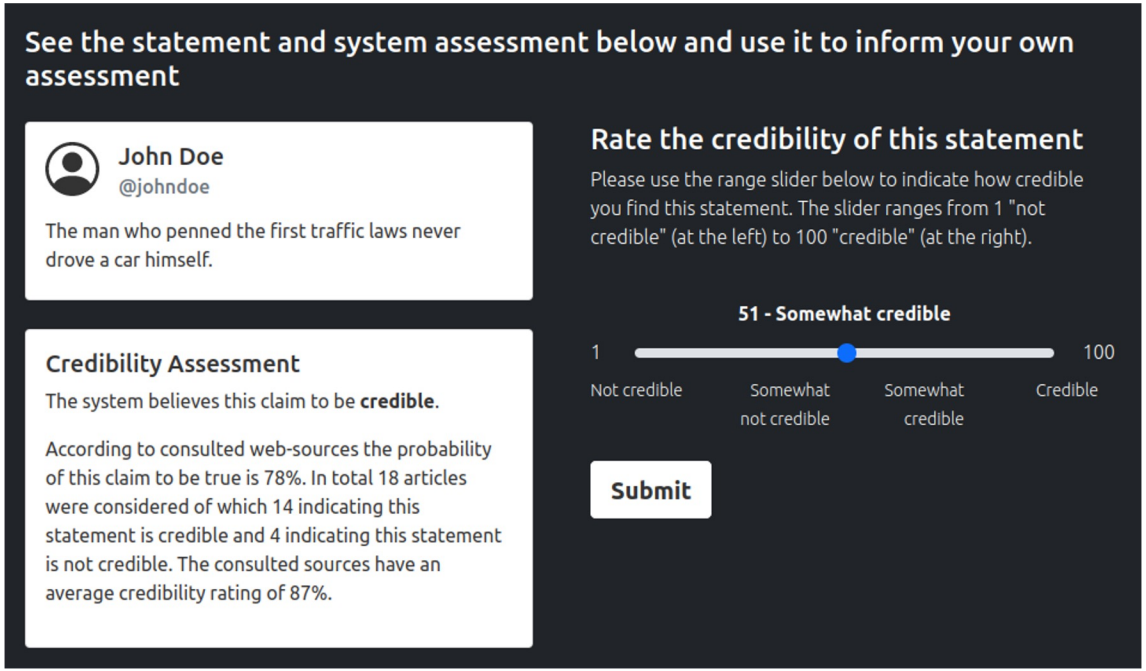
Explaining credibility assessments



TU Delft

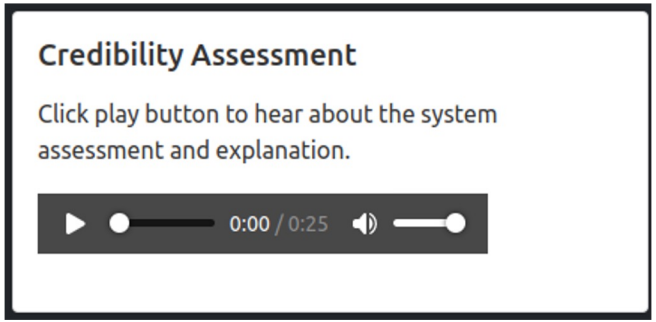
Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018. CredEye: A Credibility Lens for Analyzing and Explaining Misinformation. In Companion Proceedings of the The Web Conference 2018 (Lyon, France) (WWW '18). International World Wide Web Conferences

Explanation modalities we used



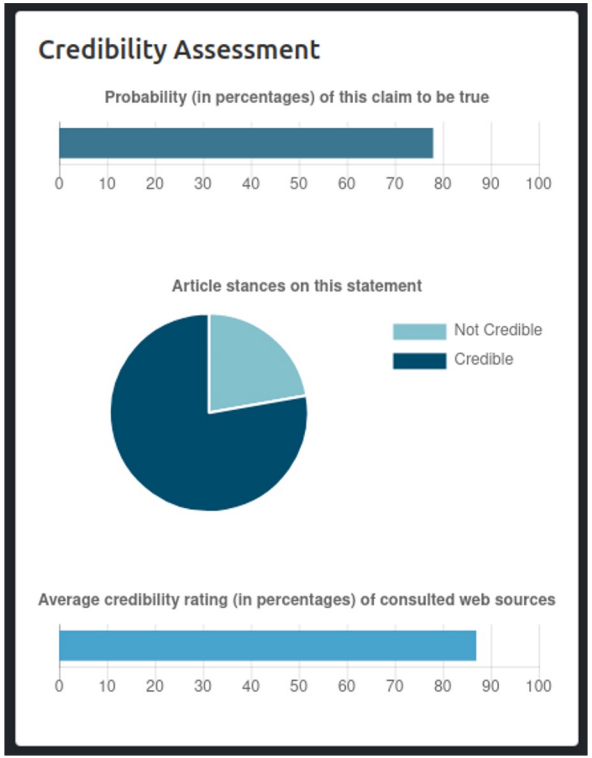
- Task interface with a text explanation

Explanation modalities we used



- Example of audio explanation component

Explanation modalities we used



- Example of graphic explanation component

KEY FINDINGS

Overall, explanations have a significantly positive effect on user accuracy

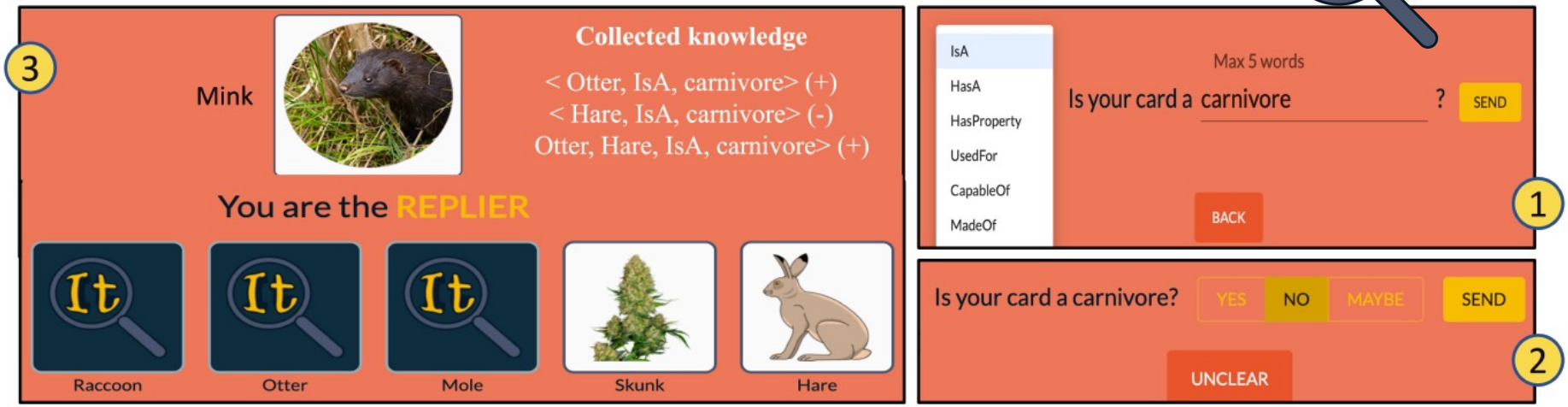
Text and audio explanations are most effective at increasing user accuracy

Graphic explanations benefit significantly from being combined with text/audio explanations

TOWARDS ROBUST AI: ELICITING DIVERSE KNOWLEDGE FROM HUMANS

- Commonsense knowledge is necessary for building neuro-symbolic AI systems and debugging deep learning models
- What types of knowledge can be collected using existing knowledge acquisition methods? → Umm… we don't really know!
- Need for collecting broad *tacit* and *negative* knowledge, and *discriminative* knowledge → Our solution… a GWAP, FindItOut

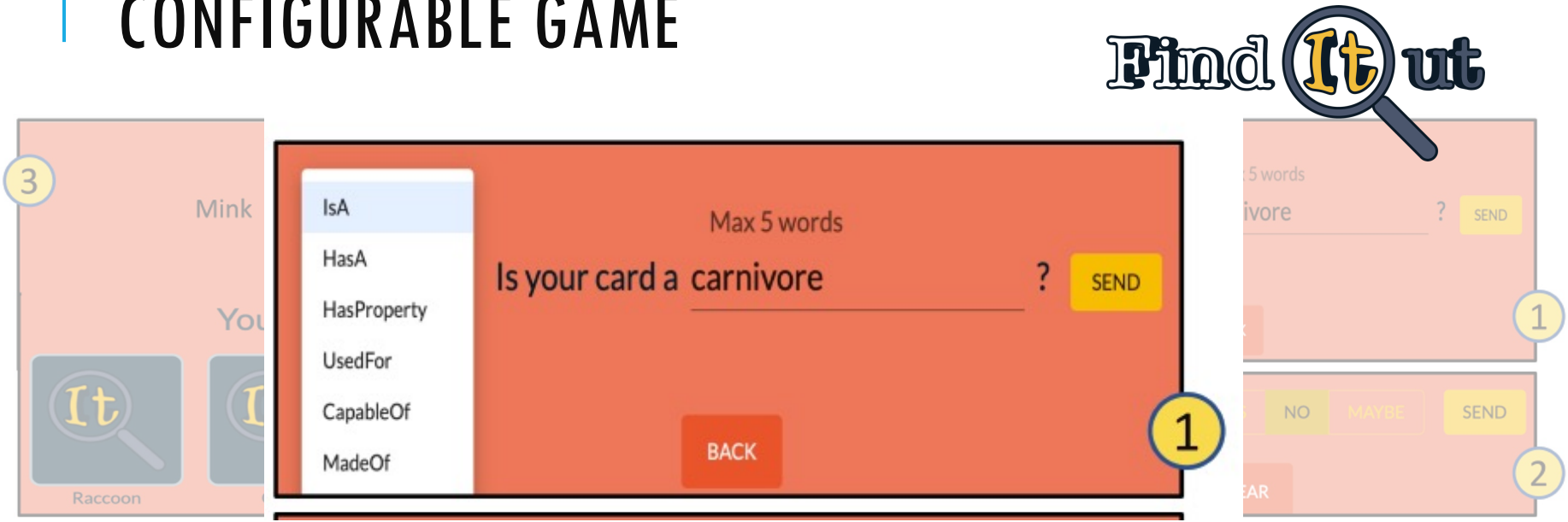
ELICITING DIVERSE KNOWLEDGE USING A CONFIGURABLE GAME



Play ➔ https://finditout.vercel.app/

Best Demo & Poster Award at AAAI HCOMP 2021

ELICITING DIVERSE KNOWLEDGE USING A CONFIGURABLE GAME



Play ➔

https://finditout.vercel.app/

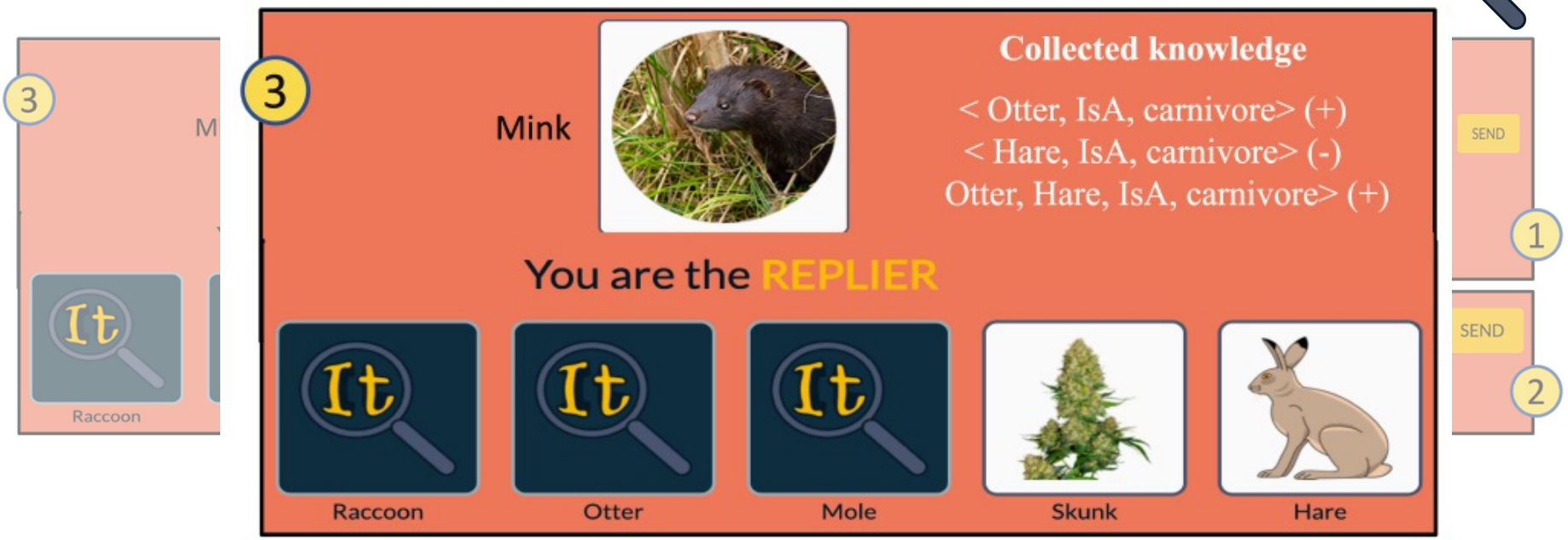
ELICITING DIVERSE KNOWLEDGE USING A CONFIGURABLE GAME

Play ➔

https://finditout.vercel.app/

ELICITING DIVERSE KNOWLEDGE USING A CONFIGURABLE GAME

Find It ut



Play ➔

https://finditout.vercel.app/

COLLECTING TACIT KNOWLEDGE FOR DOWNSTREAM AI TASKS

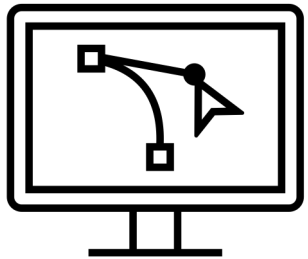
Board	Type	Question	Knowledge Tuple
floor, window, bathroom, walls, ceiling, chandelier, mirror, bedroom	Explicit Tacit	Can your card be found inside an apartment? Can your card be used for decoration?	<bathroom, AtLocation, inside apartment> <chandelier, UsedFor, decoration>
necklace, dress, boots, shoes, pants, trousers, jeans, skirt	Explicit Tacit	Can your card be found in your wardrobe? Is your card typically worn by cowboys?	<dress, AtLocation, wardrobe> <boots, HasProperty, worn by cowboys>

Empirical Results:

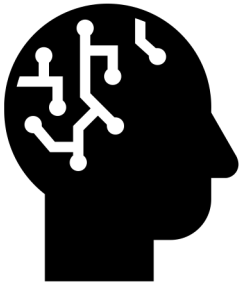
- 125 players played 2430 rounds ➔ 150*k* knowledge tuples
- Efficiency of game is 10x higher than a reference baseline; *Verbosity*
- Usefulness validated in two downstream AI tasks
  - Commonsense Question-Answering
  - Identification of Discriminative Attributes
- Enjoyable game experience (player experience inventory)

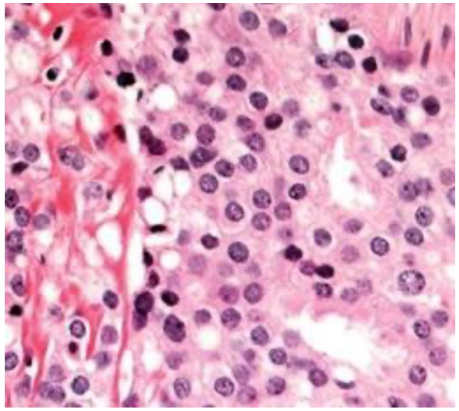
WWW 2022
Best Paper Nomination

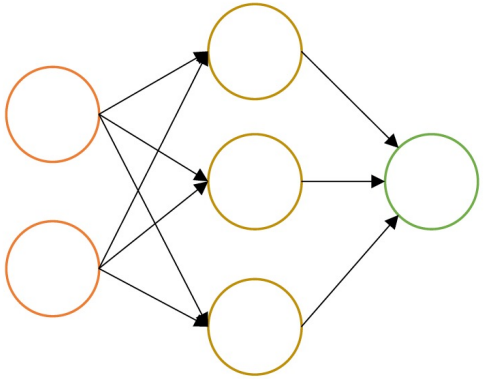
APPROPRIATE RELIANCE ON AI SYSTEMS BY HUMANS (1/2)



- How does stated accuracy inform system reliance?
- ANIE : Analogies for Intelligible Explanations
- A structural mapping of a target domain to be clarified (e.g., the system accuracy) onto a source domain which the recipient of the analogy is more familiar with
  - "the system is 75% accurate, which is about as reliable as the AstraZeneca vaccine is for protecting against covid"
  - "the system is 75% accurate, which is about as reliable as the 5-day weather prediction"







Positive for prostate cancer

Input Sample

ML model

Model Prediction

Concept-level explanation (target sentence): With cribriform and fused glands in needle core biopsy from prostate, this is diagnosed as adenocarcinoma of the prostate.

Analogy-based explanation: Cribriform and fused glands in needle core biopsy is definitely a sign of prostate cancer. ***It is like recognizing a unicorn due to the horn on its head.***

APPROPRIATE RELIANCE ON AI SYSTEMS BY HUMANS (2/2)

- Trust development in Human-AI interaction
  - Single and multiple interactions; individuals and communities
- Conversational human-AI interfaces
  - AI system metaphors and anthropomorphism

(a) God (b) Human (c) Animal (d) Plant (e) Inorganic Object

TRUST (HUMAN–AI) → RESEARCH GAPS

How does trust evolve in the interaction between humans and AI systems?

To what extent is trust that is established through such interactions robust to system accuracy over time?

What factors mediate trust formation?

GET YOUR HANDS ON!

- https://github.com/salesforce/OmniXAI

OmniXAI: A Library for Explainable AI

Wenzhuo Yang[1,*], Hung Le[1], Silvio Savarese[1], and Steven C.H. Hoi[1,*]

[1]Salesforce Research
[*]Corresponding Authors: {wenzhuo.yang,shoi}@salesforce.com

u.k.gadiraju@tudelft.nl

https://www.ujwalgadiraju.com

@UJLAW