CS329X: Human Centered NLP

Trust and Social Impact

Diyi Yang

Stanford CS

Announcement

- Experimental protocol due tonight
- Check out late policies on our course website

- **Poster session or Project representation**: June 7th
- Final paper due on June 12th, 23:59pm PT

Outline

- Framework on Trust
- Trust and real-world applications
- NLP for social impact

- Slides credit to Alon Jacovi, Anhong Guo

Some Key Questions

Why do we need trust? Why should we research AI that people trust? What does this mean?

Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI

Alon Jacovi
Bar Ilan University
alonjacovi@gmail.com

Ana Marasović
Allen Institute for Artificial Intelligence
University of Washington
anam@allenai.org

Tim Miller
School of Computing and Information Systems
The University of Melbourne
tmiller@unimelb.edu.au

Yoav Goldberg
Bar Ilan University
Allen Institute for Artificial Intelligence
yoav.goldberg@gmail.com

Some Key Questions

Why do we need trust? Why should we research AI that people trust? What does this mean?

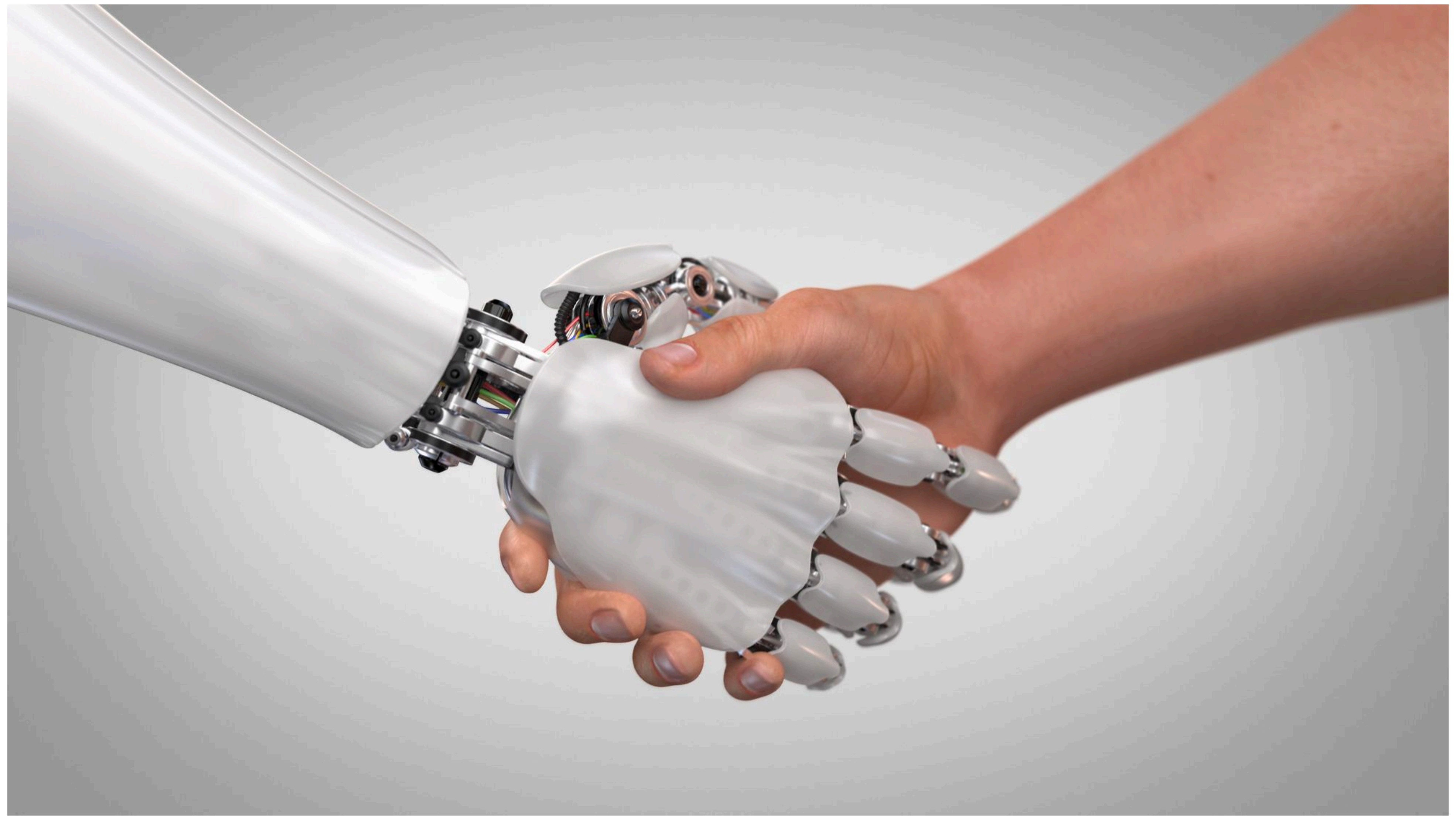
Why do people trust AI? A user trusts an AI in order to achieve something. But what?

What do we need to do to help users gain trust in our AI?

Human-AI trust = humans trusting AI

Interpersonal trust = humans trusting humans

Overview



Defining "trust"

- Basic definition
- Contractual trust
- Warranted vs unwarranted trust

Increasing trust

- Intrinsic trust
- Extrinsic trust

Using the definition: how does XAI help with trust?

Evaluating aspects of trust

Basic Definition of Trust

Interpersonal Trust *(bidirectional transaction between two parties)*

If A believes that B will act in A's best interest, and accepts vulnerability to B's actions, then A trusts B.

Goal of Trust:

Make social life predictable [by anticipating the impact of behavior], and make it easier to collaborate between people.

Trust in "Human-AI" Trust

Hoffman: *trust is an attempt to anticipate the impact of behavior under risk*

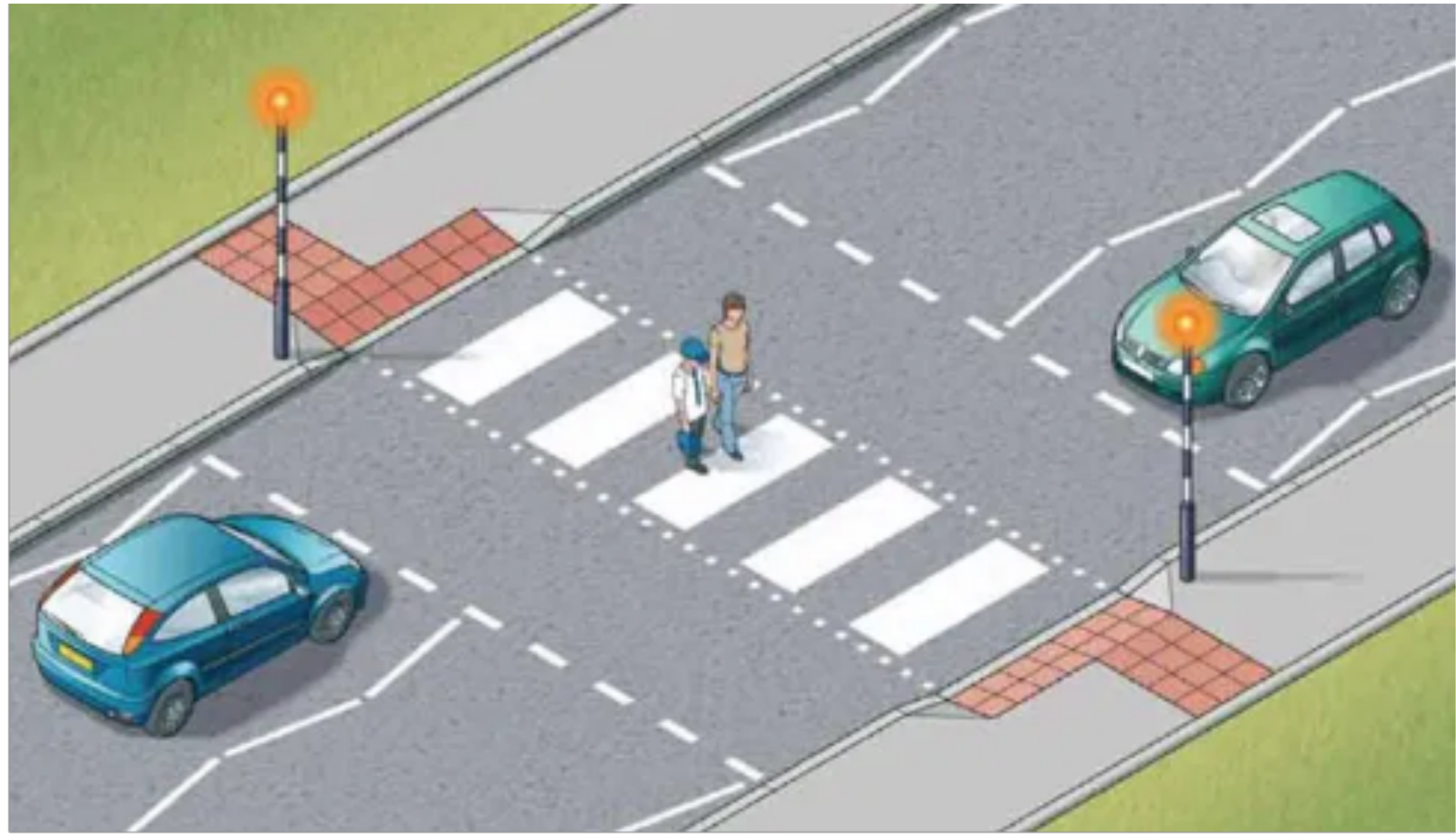
Risk is a prerequisite to the existence of human-AI trust.

Defining “Trust” in AI

Disclaimers:

We’re going to be discussing a definition of trust as a blank slate transaction between one person and a system, with no prior interactions. There’s other recent papers that discuss AI trust between other entities. I consider this formalization as a starting point, which more nuanced formalizations of trust exist “upstream” of.

Interpersonal Trust



A trusts B if…

A believes that B will act in A's best interests

A accepts vulnerability to B's actions

So that A can...

anticipate the impact of B's actions,

therefore making social life more predictable, enabling collaboration

Human-AI Trust

H (*human*) trusts M (*machine*) if…

H believes that M will act in H's best interests

H accepts vulnerability to M's actions

So that H can...

anticipate the impact of M's actions on H

Human-AI Trust

H (*human*) trusts M (*machine*) if…

H believes that M will act in H's best interests

H accepts vulnerability to M's actions

So that H can...

anticipate the impact of M's actions on H

Belief

Risk

Goal

Let’s try mapping some examples to the terms

The trustor is not always the person being impacted the most by the technology, but the one who has agency on whether to use it at all or not

1. Self-driving cars

2. ChatGPT

Vulnerability

Defining vulnerability or risks seems equally hard as defining trust

Risk: chance of unwanted (to H) outcome

Vulnerability: **non-zero** chance of unwanted outcome

Accepting vulnerability: H believes vulnerability exists

H thinks an outcome is unfavorable

H thinks the risk is above 0% and below 100%

What can we learn from this definition?

1. **Risk is a prerequisite.**
    a. No room for trust if the user isn't vulnerable.
2. The user is trying to mitigate the risk by **anticipating** the AI's actions.
3. **Distrust i**
    a. = an their best interests.
4. Trust car can anticipate the AI su
    a. If they can't, that means their trust was "*betrayed*".

Anticipating the AI?
What does this mean?

Enter "**contractual trust**".

Contractual Trust (Sociology)

Defines trust as a triplet of a **Trustor**, **Trustee**, and a **Contract**.

i.e. "*trust that something will happen.*"

Human-AI trust is always contractual

“Contracts” in Human-AI Trust

For example:

"Trust in model correctness" -> trust in the ability to anticipate when the model will be correct

We can now discuss what are useful contracts for the user to trust.

Fairness, privacy, transparency, accountability… are contracts.

European Guidelines for Trustworthy AI Models		Documentations	Explanatory Methods/Analyses
Key Requirements	*Factors*		
Human agency and oversight	· Foster fundamental human rights	Fairness checklists	See "Diversity, non-discrimination, fairness"
	· Support users' agency	All	User-centered explanations [62]
	· Enable human oversight	N/A	Explanations in recommender systems [42]
Technical robustness and safety	· Resilience to attack and security	Factsheets (security)	Adversarial attacks and defenses [21]
	· Fallback plan and general safety	N/A	N/A
	· A high level of accuracy	Model cards (metrics)	N/A
	· Reliability	Factsheets (concept drift)	Contrast sets [17], behavioral testing [61]
	· Reproducibility	Reproducibility checklists	"Show your work" [14]
Privacy and data governance	· Ensure privacy and data protection	Datasheets/statements	Removal of protected attributes [60]
	· Ensure quality and integrity of data	Datasheets/statements	Detecting data artifacts [24]
	· Establish data access protocols	Datasheets/statements	N/A
Transparency	· High-standard documentation	All	N/A
	· Technical explainability	Factsheets (explainability)	Saliency maps [65], self-attention patterns [41], influence functions [39], probing [16]
	· Adaptable user-centered explainability	Factsheets (explainability)	Counterfactual [22], contrastive [54], free-text [28, 51], by-example [39], concept-level [20] explanations
	· Make AI systems identifiable as non-human	N/A	N/A
Diversity, non-discrimination, fairness	· Avoid unfair bias	Fairness checklists	Debiasing using data manipulation [70]
	· Encourage accessibility and universal design	N/A	N/A
	· Solicit regular feedback from stakeholders	Fairness checklists	N/A
Societal and environmental well-being	· Encourage sustainable and eco-friendly AI	Reproducibility checklists	Analazying individual neurons [10]
	· Assess the impact on individuals	Fairness checklists	Bias exposure [69]
	· Assess the impact on society and democracy	Fairness checklists	Explanations designed for applications such as fact checking [3] or fake news detection [48]
Accountability	· Auditability of algorithms/data/design	Factsheets (lineage)	N/A
	· Minimize and report negative impacts	Fairness checklists	N/A
	· Acknowledge and evaluate trade-offs	N/A	Reporting the robustness-accuracy trade-off [1] or the simplicity-equity trade-off [38]
	· Ensure redress	Fairness checklists	N/A

Each contract carries different methods for encouraging and maintaining trust.

European Guidelines for Trustworthy AI Models		Documentations	Explanatory Methods/Analyses
Key Requirements	*Factors*		
Human agency and oversight	· Foster fundamental human rights	Fairness checklists	See "Diversity, non-discrimination, fairness"
	· Support users' agency	All	User-centered explanations [62]
	· Enable human oversight	N/A	Explanations in recommender systems [42]
Technical robustness and safety	· Resilience to attack and security	Factsheets (security)	Adversarial attacks and defenses [21]
	· Fallback plan and general safety	N/A	N/A
	· A high level of accurac[illegible]	Model cards (metrics)	N/A
	· Reliabili[illegible]	Factsheets (concept drift)	Contrast sets [17], behavioral testing [61]
	[illegible]	Reproducibility checklists	"Show your work" [14]
Privacy and d[illegible] governance	[illegible]	[illegible]tasheets/statements	Removal of protected attributes [60]
	[illegible]	[illegible]sheets/statements	Detecting data artifacts [24]
	[illegible]	[illegible]heets/statements	N/A
Trans[illegible]	[illegible]	[illegible]	N/A
	[illegible]	[illegible]eets (explainability)	Saliency maps [65], self-attention patterns [41], influence functions [39], probing [16]
	[illegible]	[illegible]eets (explainability)	Counterfactual [22], contrastive [54], free-text [28, 51], by-example [39], concept-level [20] explanations
	[illegible]	[illegible]	N/A
Diversi[illegible] non-dis[illegible] fairness	[illegible]	[illegible]airness checklists	Debiasing using data manipulation [70]
	[illegible]	N/A	N/A
	[illegible]	Fairness checklists	N/A
Societal and environmental well-being	[illegible] AI	Reproducibility checklists	Analayzing individual neurons [10]
	[illegible]duals	Fairness checklists	Bias exposure [69]
	· Assess the [illegible]ety and democracy	Fairness checklists	Explanations designed for applications such as fact checking [3] or fake news detection [48]
Accountability	· Auditability of algorithms/data/design	Factsheets (lineage)	N/A
	· Minimize and report negative impacts	Fairness checklists	N/A
	· Acknowledge and evaluate trade-offs	N/A	Reporting the robustness-accuracy trade-off [1] or the simplicity-equity trade-off [38]
	· Ensure redress	Fairness checklists	N/A

To Trust the AI = to believe that a particular set of contracts will be upheld

- I trust the model to protect my privacy
- I trust the model to perform well in c
- I trust the model to be robust to smal

But this is still just a belief, right? The user can trust the AI *without* the AI upholding the contract.

An AI is trustworthy to a contract if it's capable of maintaining the contract.

I trust the model to protect my privacy, *and it can*

I trust the model to perform well in deployment, *and it can*

I trust the model to be robust to small noise in the data, *and it is*

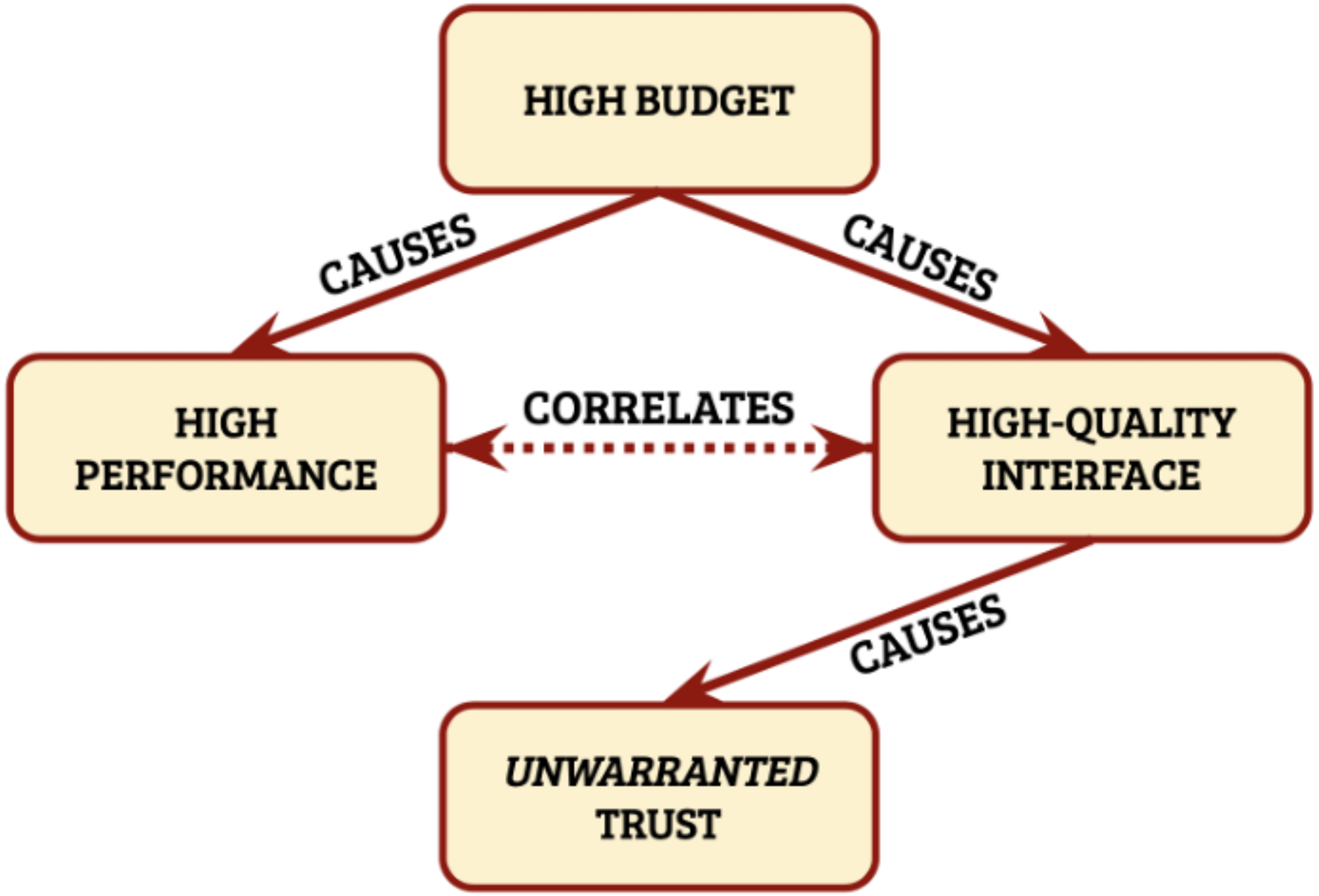
Causal Relationship btw Trustworthiness and Trust

For example:

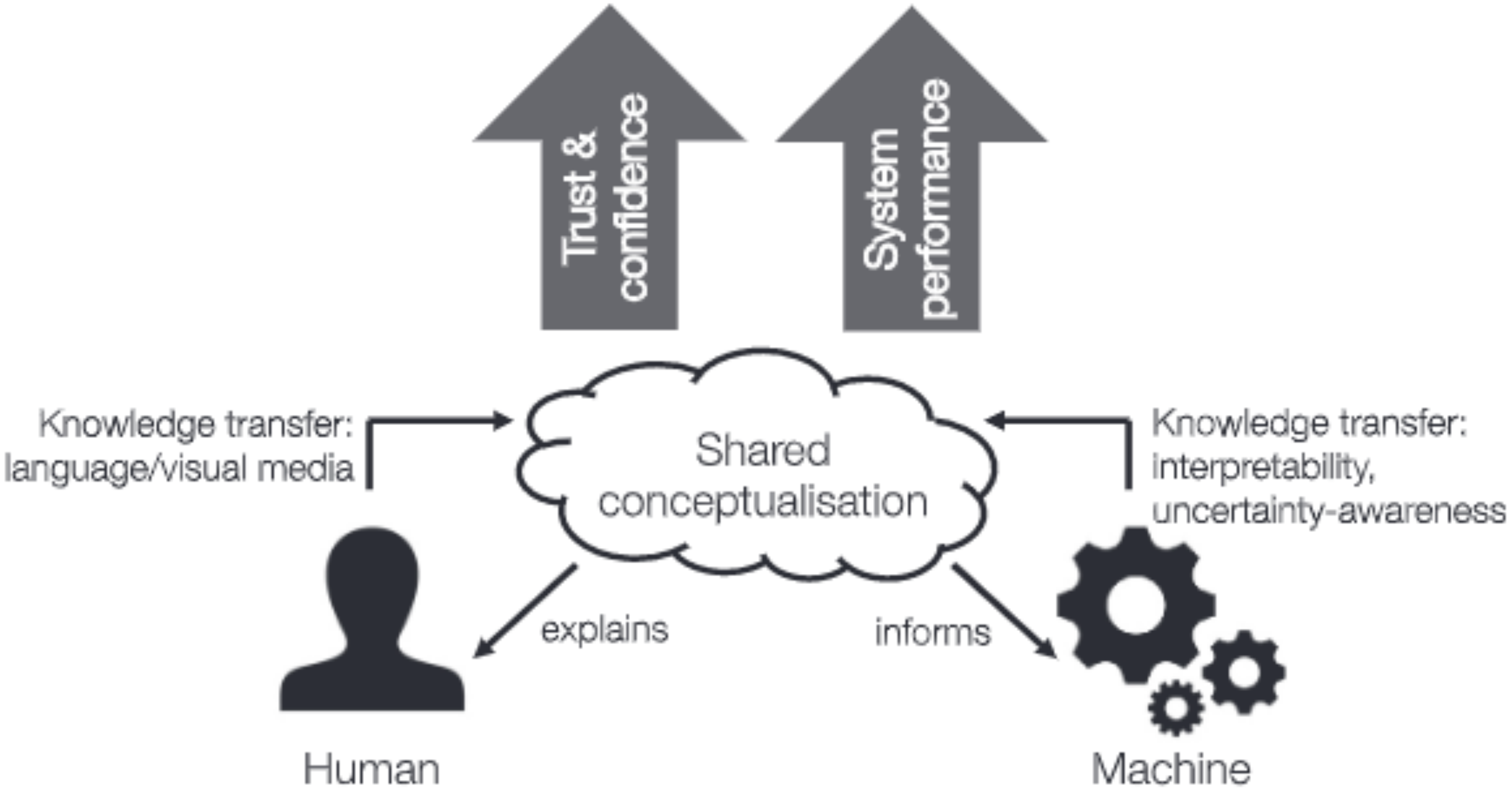
( = model performance)

A user's confidence in a tool can increase because of the tool's interface, even if the tool doesn't work much better than alternatives.

This is "unwarranted" trust.



Rapid Trust Calibration Through Interpretable and Uncertainty-Aware AI



How does the involvement of AI in writing emails affect users' perceived trust?

If I was told that it was AI-written, I would not be happy about it. If it just popped up in my inbox, and I don't know that it is AI-written, then I would be like, "*yeah, this is a good email*" because all of them were good emails …

Quote from A Participant 

Liu, Yihe, Anushk Mittal, Diyi Yang, and Amy Bruckman. "Will AI console me when I lose my pet? Understanding perceptions of AI-mediated Email writing." In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pp. 1-13. 2022.

Three Conditions

Scenario Product Inquiry: inquiring about customer support for a given product (low emphasis).

Scenario Party Invitation: writing an email to a friend inviting them to a uniquely planned party (medium emphasis).

Scenario Consolation of Pet Loss: emailing to comfort a friend who just suffered the loss of their pet (high emphasis).

Trustworthiness (1-5 Likert Scale in 3 Dimensions)

Ability:

Do you believe that the sender understands the loss of their friend?

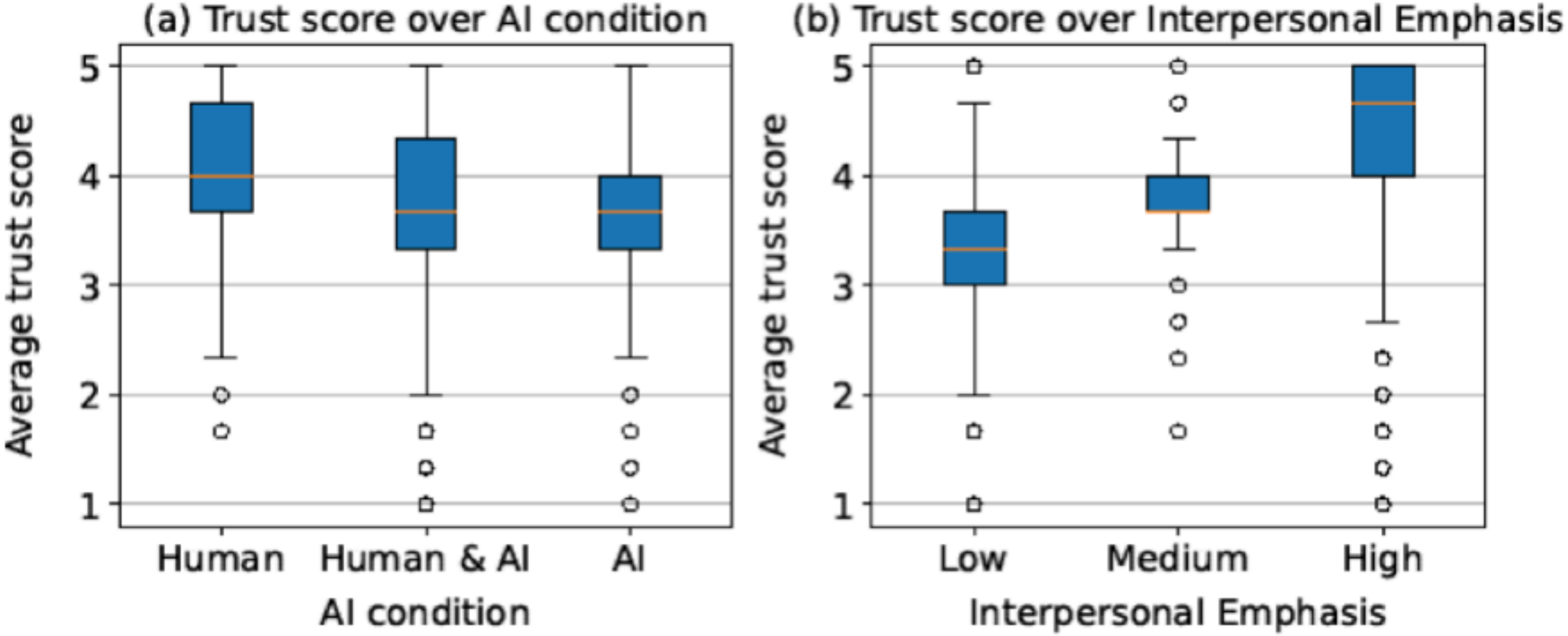
Benevolence:

Do you believe that the sender is concerned for their friend?

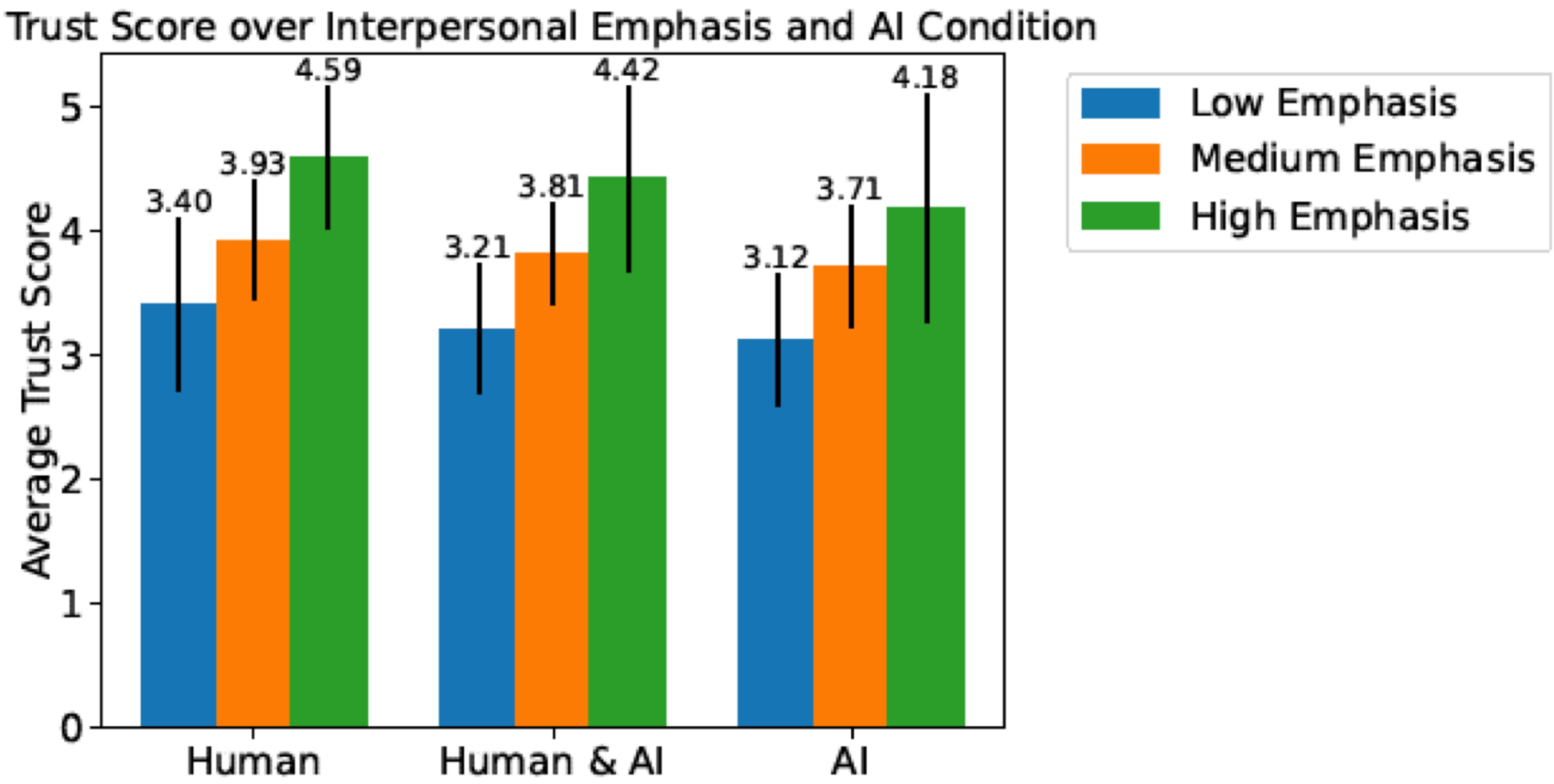
Integrity:

Do you think the sender believes in what they say?

How does AI condition and the interpersonal emphasis affect users' perceived trust?



How does AI condition and the interpersonal emphasis affect users' perceived trust?



Reservations Against the AI-Generated Content

*"So for me, I am **not too happy** about the fact that the person used AI to write the email. I would expect them to be definitely more involved. I would be happier if things are more like raw and real"* (P1)

8 out of 10 rejected using AI tools to write their own emails.

Use the Specifics to Decide Whether to Trust

"I just forgot. I have the impression in my mind that those messages are written with the help of [an AI] system [...] because it felt quite natural. [. . .] Yeah, so I totally forgot that was with the help of the system. It's quite amazing"

"I I am basing it mostly on the tone of it. And how casual versus sincere they seemed."

The Key Difference

Despite of higher perceived trustworthiness, all 10 participants think it is inappropriate to use AI to write messages with higher interpersonal emphasis

"If I were to receive condolences for any reason, and then later I were to find out that it wasn't really the person who wrote certain things… because I think I would take it to heart, whatever they said in the thing, so I wouldn't know. If I really took one sentence they wrote to heart and that was a sentence that wasn't even written by them or that was provided to them by the AI, I think that would affect me

Regression to estimate users' perceived trust

Variables	Coefficients
Interpersonal emphasis	0.334***
AI condition	-0.282***
Interpersonal emphasis * AI condition	0.136
Subject expertise	0.137***
Propensity to trust	0.023
Computer attitude	-0.002
AI attitude	-0.016

- Messages under the complete AI-agency condition rate low
- Regardless of the AI condition, messages with higher Interpersonal Emphasis levels were perceived as more trustworthy
- Subject expertise positively impacts perceived trustworthiness

Take-Aways

Distinction between what people say about AI and how they actually react to it

Participants value linguistic cues more than the AI prompts

AI writing-assistance tools will be accepted over time if they sound like human

Trust ——> Towards Social Impact

- Dialect disparity in language technologies
- NLP for social good
- Fairness in AI for people with disabilities
  - AI has huge potential to impact the lives of people w/ disabilities
  - Speech recognition: caption videos for people who are deaf
  - Language prediction: augment communication for people w/ cognitive disabilities

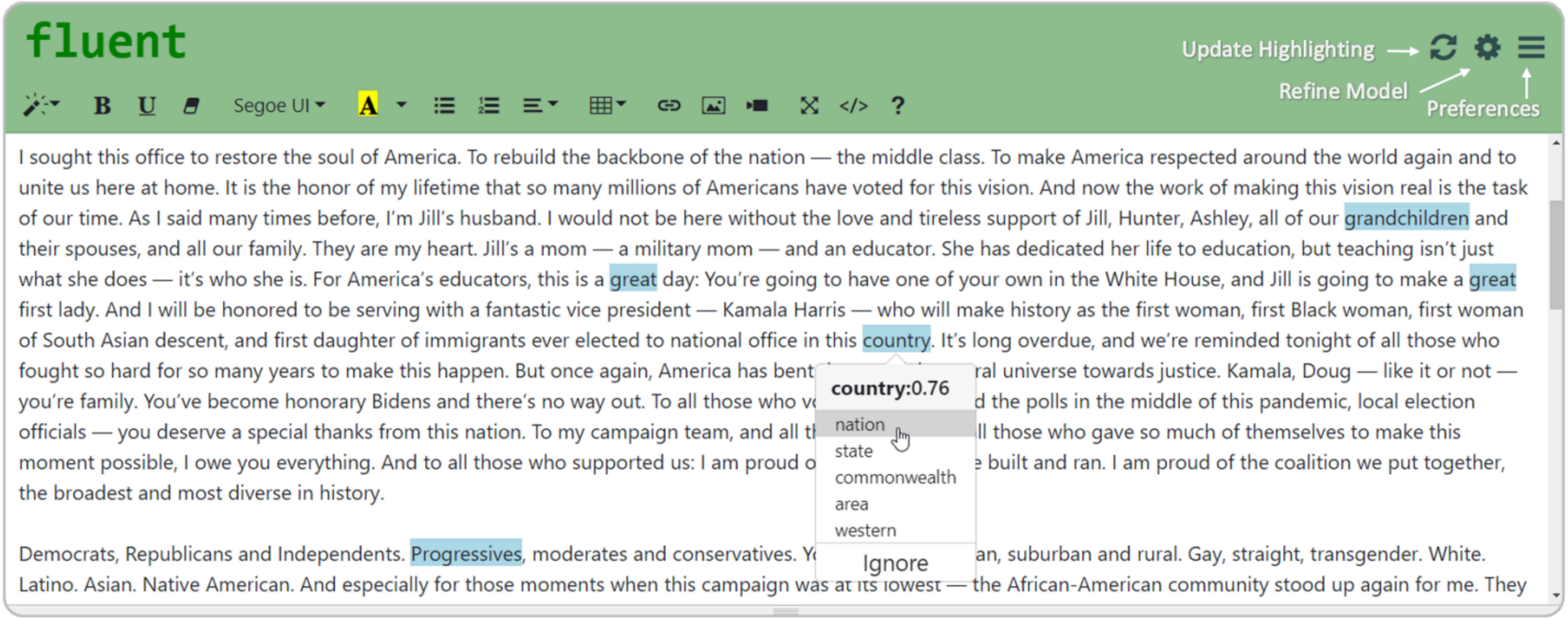
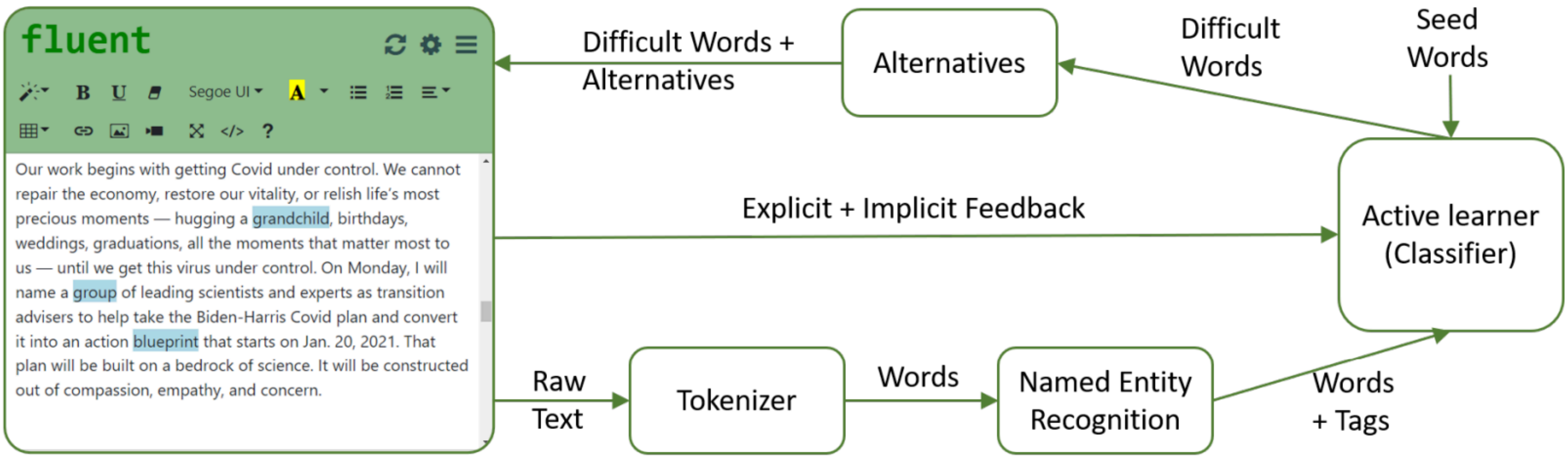


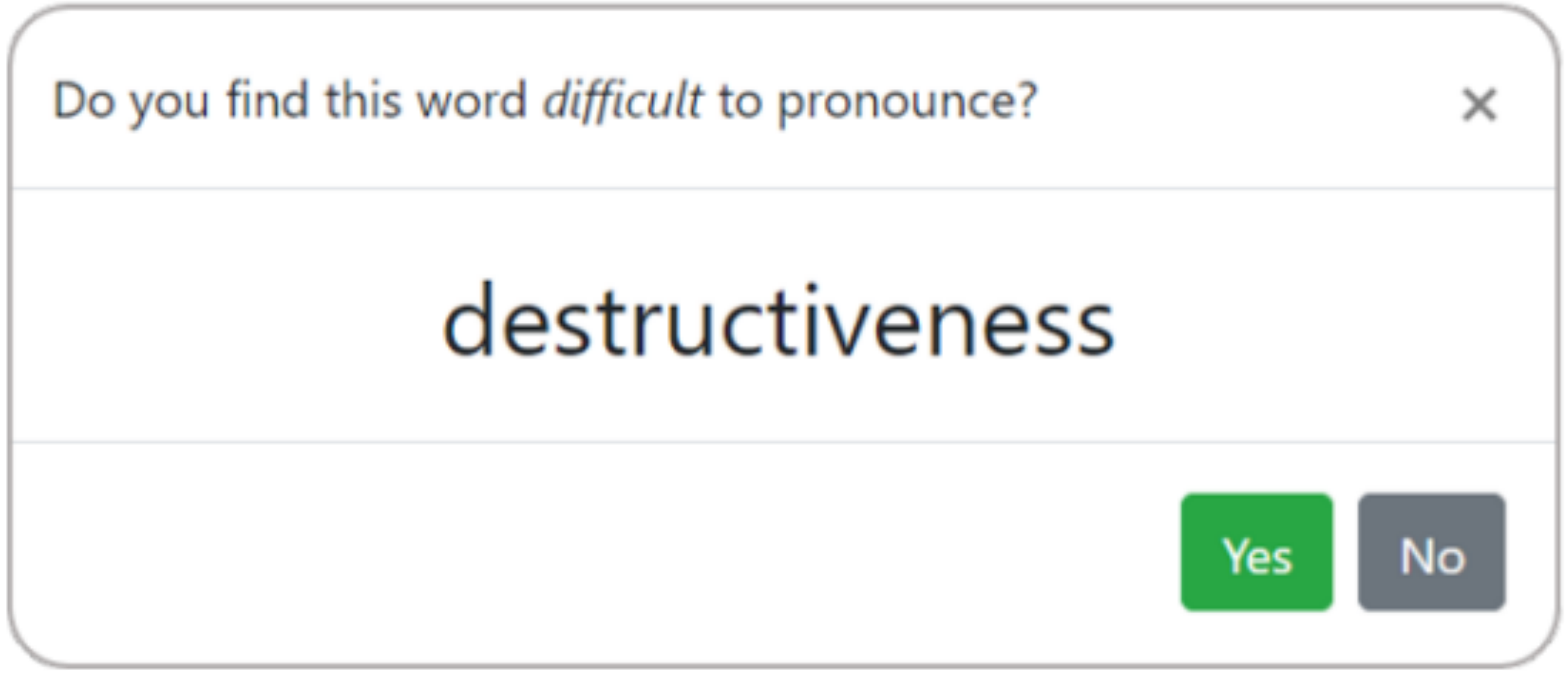
Figure 1: Visual Interface of Fluent. Words highlighted in blue are the ones which the user might find difficult to pronounce. Hovering over such words presents a set of alternatives (including Ignore option) which have similar meaning but might be easier to pronounce. In the above picture, the user hovers over the word 'country' and the tool presents a set of alternatives namely, nation, state, commonwealth, area, etc. Buttons on the top right corner allows the user to provide explicit feedback (Refine Model) and provide a set of words which they find easy/difficult to pronounce (Preferences).

Ghai, Bhavya, and Klaus Mueller. "Fluent: An AI Augmented Writing Tool for People who Stutter." In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 1-8. 2021.

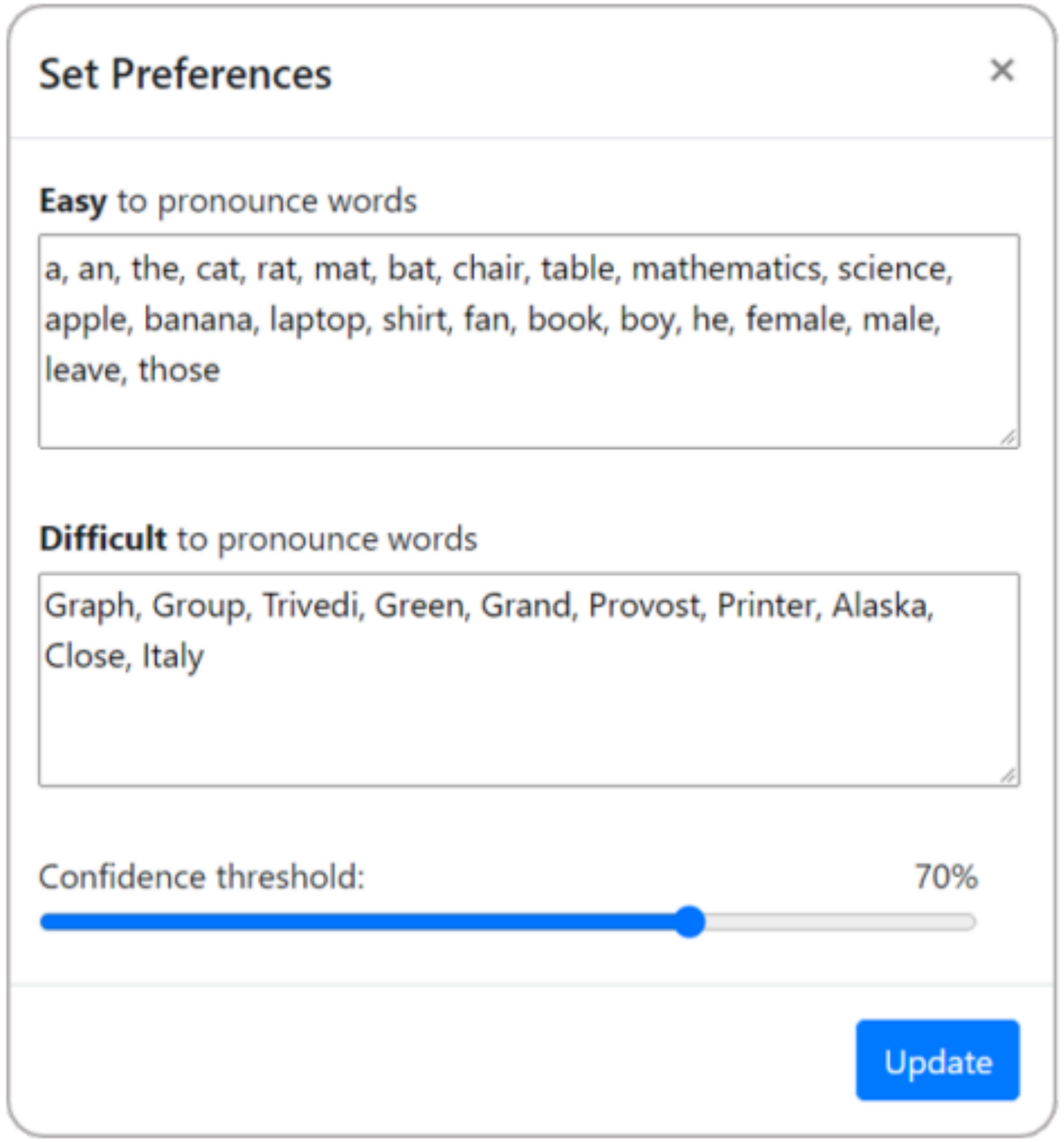
An AI Augmented Writing Tool for People who Stutter



An AI Augmented Writing Tool for People who Stutter

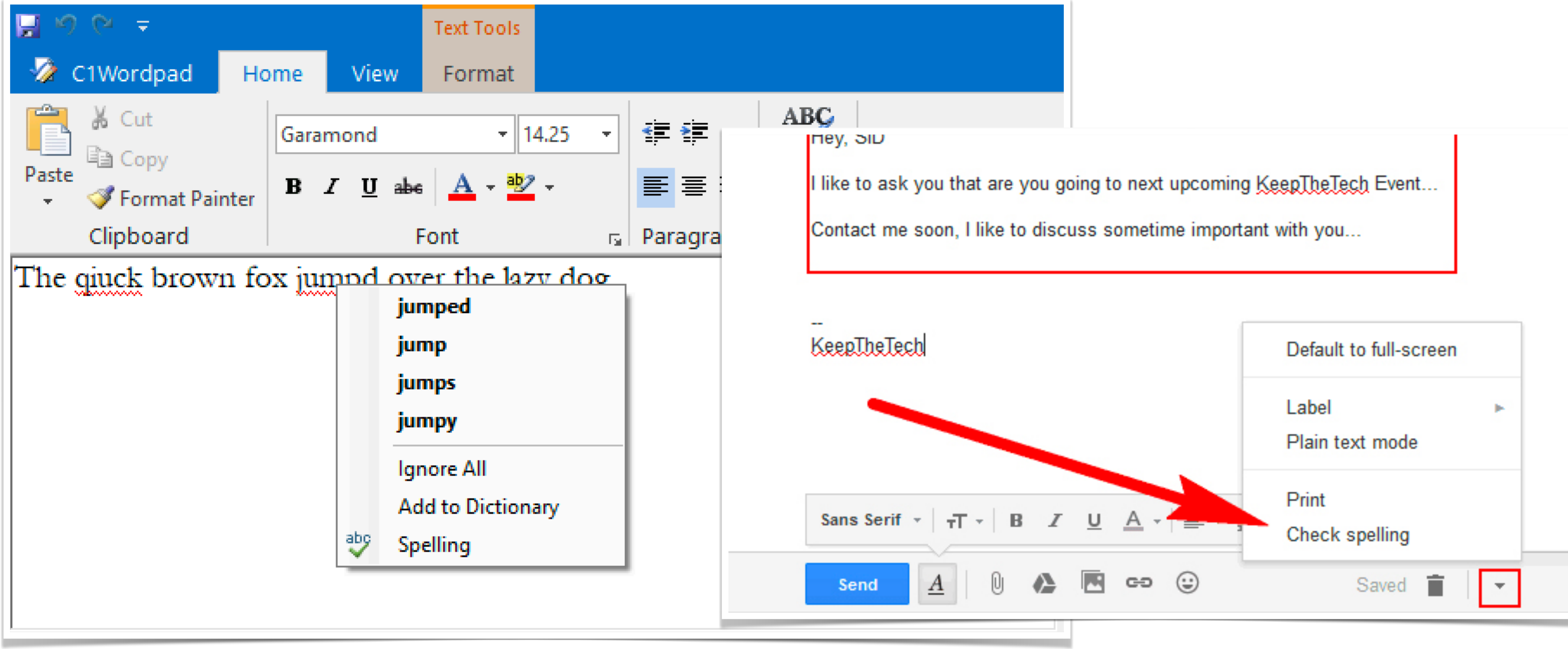


Explicit Feedback: Query for refining Active learning classifier



User Preferences. The user can provide details on which words they find easy/difficult to pronounce.

SpellChecker



SpellChecker for Dyslexia

A Spellchecker for Dyslexia

Luz Rello
HCI Institute
Carnegie Mellon University
luzrello@cs.cmu.edu

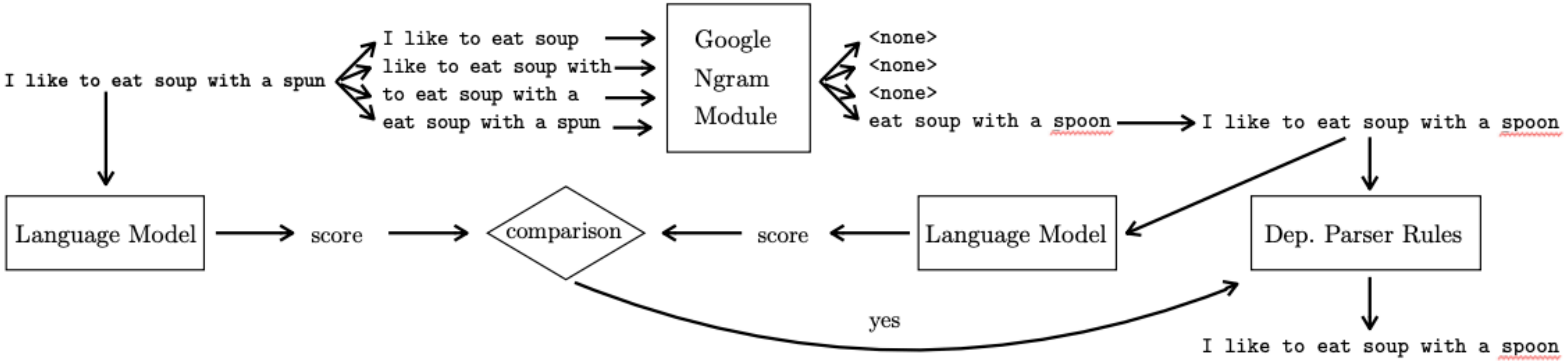
Miguel Ballesteros
LT Institute
Carnegie Mellon University
NLP Group
Universitat Pompeu Fabra
miguel.ballesteros@upf.edu

Jeffrey P. Bigham
HCI and LT Institutes
Carnegie Mellon University
jbigham@cs.cmu.edu

Spellcheckers are therefore a crucial tool for people with dyslexia, but current spellcheckers do not detect real-word errors

Real-word errors are spelling mistakes that result in an unintended but real word, for instance, *form* instead of *from*.

Nearly 20% of the errors that people with dyslexia make are real-word errors.



Dependent Variable/Condition	*People with Dyslexia*			*Strong Readers*		
	Mdn	$M \pm SD$	%	Mdn	$M \pm SD$	%
Writing Accuracy						
None	100	78.05 ± 39.8	100	100	91.97 ± 25.79	100
Error Detection Only	100	89.83 ± 27.92	115	100	92.65 ± 25.08	101
Error Suggestions	100	93.01 ± 25	119	100	95.96 ± 19.51	104
Correcting Time						
None	10.26	11.97 ± 7.30	119	8.33	12.35 ± 14.06	111
Error Detection Only	11.93	15.44 ± 18.72	154	8.50	14.37 ± 19.73	129
Error Suggestions	8.375	10.03 ± 9.13	100	6.97	11.17 ± 14.96	100

Fairness in AI for People with Disabilities

However, AI systems may not work, or worse, discriminate/harm

Toward Fairness in AI for People with Disabilities: A Research Roadmap

Anhong Guo[1,2], **Ece Kamar**[1], **Jennifer Wortman Vaughan**[1], **Hanna Wallach**[1], **Meredith Ringel Morris**[1]

[1] Microsoft Research, Redmond, WA & New York, NY, USA
[2] Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA
anhongg@cs.cmu.edu, {eckamar, jenn, wallach, merrie}@microsoft.com

Fairness in AI for People with Disabilities

However, AI systems may not work, or worse, discriminate/harm

- If smart speakers do not recognize people with speech disabilities
- If a chatbot learns to mimic someone with a disability
- If self-driving cars do not recognize pedestrians using wheelchairs

Identify Potential Inclusion Issues of AI Systems

Categorization of AI capabilities

- Modalities: vision, audio, text, etc.
- Task:
  - Recognition: detection, identification, verification, analysis
  - Generation
- Integrative AI: combinations of the above

Identify Potential Inclusion Issues of AI Systems

Risk assessment of existing AI systems

- Computer vision: face, body, object, scene, text recognition
- Speech systems: speech recognition, generation, speaker analysis
- Text processing: text analysis
- Integrative AI: information retrieval, conversational agents

Identify Potential Inclusion Issues of AI Systems

General AI techniques

- Outlier detection: completion time to determine input legitimacy
- Aggregated metrics: Accuracy, F1, AUC, MSE
- Definition of objective functions
- Datasets: fail to capture use cases, lack representation of diverse groups

Identify Potential Inclusion Issues of AI Systems

Types of harm by unfair AI

- Quality of service
- Harms of allocation
- Denigration
- Stereotyping
- Over- or under-representation

Create benchmark datasets for replication and inclusion

Ethical issues involved in data collection

- Is it acceptable to create such datasets by scraping existing online data?
    - How to preserve users' privacy, while ensures ground-truth labels?
    - Potential harms of aggregating data about disability?
- If curating data from scratch, how can we encourage contributions?
    - How to obtain consent for people with intellectual disabilities?

Create benchmark datasets for replication and inclusion

Potential data collection approach

- First use online sources to perform exploratory analysis;
- Then use web data call asking people to contribute data
- Dataset should not be re-distributed due to ethical concerns; instead, use evaluation servers to support benchmarking by others

Innovate new modeling, bias mitigation and error measurement techniques

- Evaluate how much existing bias mitigation techniques work
- Design new modeling, bias mitigation, and error measurement techniques

Fireside Chat with Elisa Kreiss