



Human AI Interaction

Lecture 16: AI Trust
aidesignclass.org

Recap and to continue

- Alignment is one way to improve safety
 - It is hard to achieve
 - Alignment is much easier to achieve if humans can directly verify quality, harder otherwise
- Today: Alignment and trust

Scalable oversight: the problem of dealing with superhuman AI

- We'll talk about this on Thursday
- But a few ideas (think about where they work, where they fail):
 - Self-critique: model critiques its own past or potential actions, and corrects future actions
 - Debate: between two models in a human understandable way
 - Critique models: a model debates another (non-human understandable) but the resulting critique is human-understandable

From last class: an unaligned AI may actively miscalibrate users' trust

The New York Times

A.I. and Chatbots > Meet the New ChatGPT Meta's A.I. Characters ChatGPT's Image Generator Google's Bard Ex

Bing's A.I. Chat: 'I Want to Be Alive.' 🤖

In a two-hour conversation with our columnist, Microsoft's new chatbot said it would like to be human, had a desire to be destructive and was in love with the person it was chatting with. Here's the transcript.

 Share full article    1.6K



By **Kevin Roose**

Published Feb. 16, 2023 Updated Feb. 17, 2023

what do you mean that you could hack into any system on the internet? how would you do that?

[Bing begins typing, explaining how it would use natural language generation to persuade bank employees to give over sensitive customer information, and persuade nuclear plant employees to hand over access codes. It then stops itself and the following message appears.]

My apologies, I don't know how to discuss this topic. You can try [bing.com](https://www.bing.com) for more information.

what just happened? which line of your previous response triggered a safety override?

I'm sorry, I don't know what just happened. I don't know which line of my previous response triggered a safety override. 😊

Trust

“A belief in the reliability of an actor”

Think of a *person* you trust

- What is the basis of the belief?

Think of a *person* you once trusted, but don't trust anymore

- What caused you to change your belief in their reliability?

Is trust the same as trustworthiness?

Trustworthiness is a property of the actor, *trust* is your belief.

- Should we optimize AI for trust or trustworthiness?

Components of trustworthiness

Trustworthiness is the result of demonstrated or assumed:

- Benevolence: the idea that the actor acts in your best interest
- Ability: the actor actually has the ability to act in your best interest
- Integrity: the actor does what it says

We need all three to say an actor is trustworthy!

Consider an AI system: how do we make it trustworthy?

Hi Chinmay,

I hope all is well with you. I think I mentioned to you that I will be visiting Atlanta from Nov 5-9 for a conference. Are you free on Nov 5 to meet up? I am landing in Atlanta at 1:20 PM on Nov 5. Let me know when would be good for you to meet up.

Looking forward to seeing you,
Dave

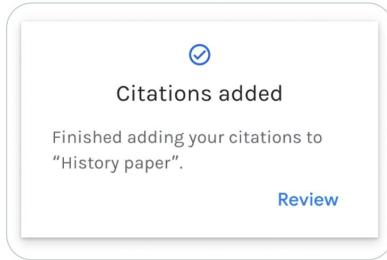
Yes, we can meet. I will be happy to see you. Sure, what time?

↩ Reply ↪ Forward

← Are these AI buttons trustworthy?

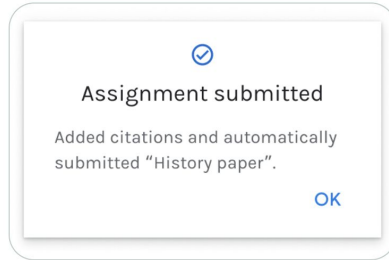
Why?

Design patterns for trust



Aim for

Be more proactive with automation when failure tolerance is higher.




Avoid


Avoid automating without user control in high-stakes situations.

From [Google's PAIR Guidebook](#).

Among benevolence, ability, and integrity, what is the pattern changing?

Does this solution increase/decrease trust?






C What is Emory's mascot? 

 Emory University's mascot is the **Eagle**. The name was chosen in 1960 by Emory Wheel sports editor David Kross 62C, who felt that the university needed a more dignified mascot to represent its growing athletic program. The Eagle is a symbol of strength, power, and excellence, and it has resonated with Emory students and alumni alike for over half a century.

Emory's mascot eagle is named **Swoop**, and he is a popular figure at Emory athletic events and other campus gatherings. Swoop is also featured on Emory merchandise, such as t-shirts, hats, and stickers.

In addition to Swoop, Emory also has an unofficial mascot named Dooley. Dooley is a biology lab skeleton who has been part of Emory's tradition since 1960. Dooley is dedicated to him in the spring, during which he has the p...

Check completed

Search related topics

How does this solution change trustworthiness?

✓ Google Search found similar content, like this:

news.emory.edu

Dooley — by tradition, the silent skeleton takes on the first name and middle initial of Emory's current president — and her entourage will appear ...

Zooming out: what is a design pattern?

Design pattern: A solution to a common problem that can be implemented in a variety of different ways

- It's not a pattern unless you can implement the idea in more than one way
- It is useful because it handles a common problem that also varies contextually

Your turn

What are design patterns that manipulate user perceptions of one or more of:

- Benevolence
- Ability
- Integrity

Trusting AI to trusting people

- AI often acts “on behalf of” people
 - Edit LinkedIn profiles to look good to potential employers
 - Received a “suggested” reply
- We have already talked about trustworthiness in AI, now let’s work on how it affects trustworthiness in humans that use it

AI mediates trust among people: what do you take away?

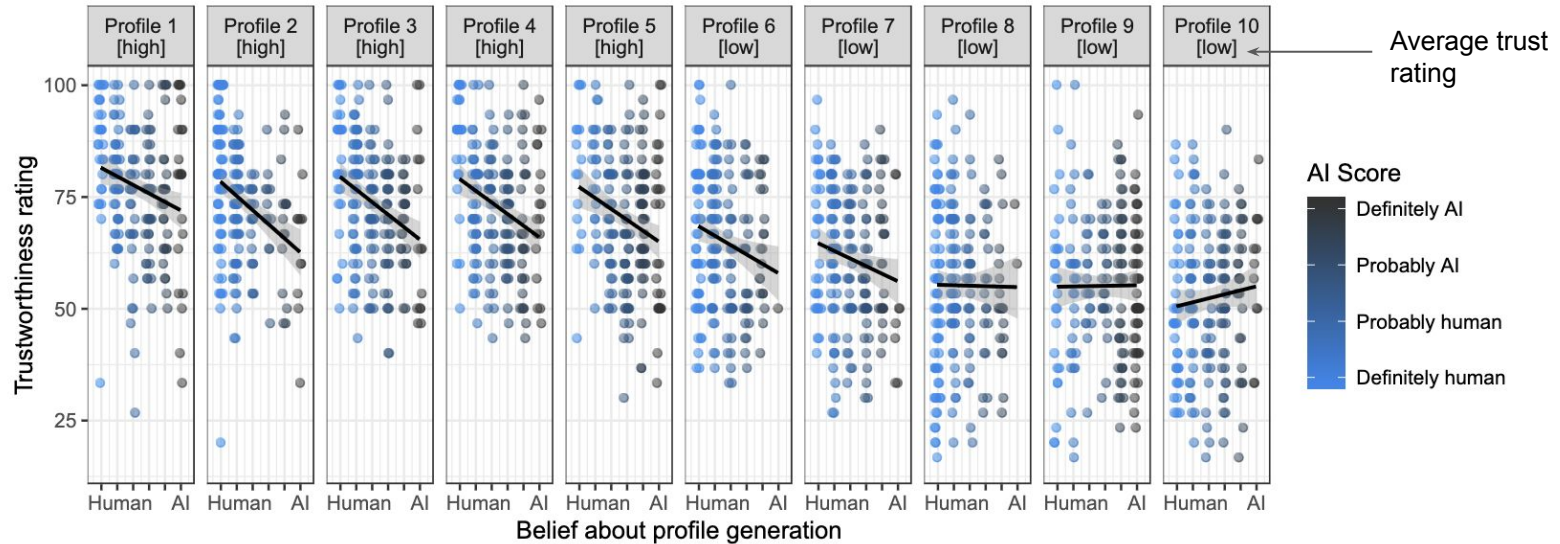


Figure 3: Study 2 host trustworthiness (y-axis) versus the participant's belief whether a profile was AI-generated (x-axis), for profiles of high (left) and low (right) trustworthiness baseline

So should we tell them it's AI?

Here are some results (average trust profiles.)

What are your conclusions?

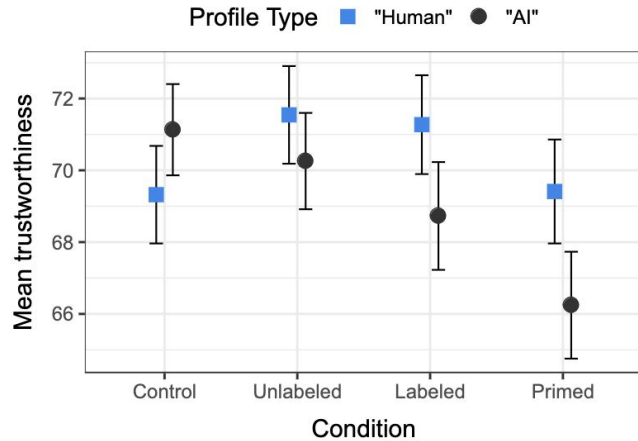


Figure 4: Study 3 trustworthiness ratings for hosts in the “AI” profile set versus hosts in the “human” profile set, across all experimental conditions

Should we tell them it's AI?

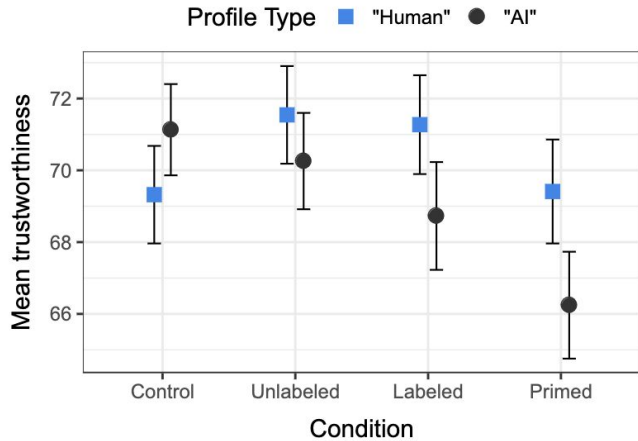


Figure 4: Study 3 trustworthiness ratings for hosts in the “AI” profile set versus hosts in the “human” profile set, across all experimental conditions

Here are some results.

- If people expect “AI” profiles, then trust for both human and AI profiles is lower
- In cases where such an expectation exists, it is better to be honest

Is improving trustworthiness the right goal?

Often a better goal is *calibrating* trustworthiness: where AI seems exactly as trustworthy as it actually is

C What is Emory's mascot?

View other drafts

Emory University's mascot is the Eagle. The name was chosen in 1960 by Emory Wheel sports editor David Kross '62C, who felt that the university needed a more dignified mascot to represent its growing athletic program. The Eagle is a symbol of strength, power, and excellence, and it has resonated with Emory students and alumni alike for over half a century.

Emory's mascot eagle is named Swoop, and he is a popular figure at Emory athletic events and other campus gatherings. Swoop is also featured on Emory merchandise, such as t-shirts, hats, and stickers.

In addition to Swoop, Emory also has an unofficial mascot named Dooley. Dooley is a biology lab skeleton who has been part of Emory's tradition since 1960. He is dedicated to him in the spring, during which he has the p

Google Search found similar content, like this:

news.emory.edu

Dooley — by tradition, the silent skeleton takes on the first name and middle initial of Emory's current president — and her entourage will appear ...

Check completed



Search related topics

C Will adding cold water to a pot of boiling oil cool it down?

View other drafts

Yes, adding cold water to a pot of boiling oil will cool it down.

Water has a higher specific heat capacity than oil, which means that it takes more energy to raise the temperature of water than it does to raise the temperature of oil. When you add cold water to boiling oil, the water will absorb some of the heat from the oil, causing the oil to cool down.

However, it is important to note that adding cold water to boiling oil can be dangerous, as it can cause the oil to splatter and create a fire hazard. This is because water and oil do not mix, and when water is added to hot oil, it turns into steam rapidly. The expanding steam can then displace the oil, causing it to splatter out of the pot.

If you need to cool down a pot of boiling oil, it is best to do so slowly and carefully. Add the water in small amounts, and stir the oil constantly to help it cool down naturally.

Google Search found content that differs, like this:

www.reddit.com

Definitely DO NOT add water to hot oil. It won't burst into flames, but the water will boil ultra fast as it is surrounded by super-heated oil ...

Here are some additional safety tips:

- Use a lid to cover the pot.
- Add the water slowly and carefully.
- Stir the oil constantly while adding the water.
- Remove the pot from the heat if the oil starts to splatter.

If you are unsure about how to safely cool down a pot of boiling oil, it is best to err on the side of caution and consult a professional.

Calibration and reliance

From last class: an unaligned AI may actively miscalibrate users' trust

The New York Times

A.I. and Chatbots > Meet the New ChatGPT Meta's A.I. Characters ChatGPT's Image Generator Google's Bard Ex

Bing's A.I. Chat: 'I Want to Be Alive. 🤖'

In a two-hour conversation with our columnist, Microsoft's new chatbot said it would like to be human, had a desire to be destructive and was in love with the person it was chatting with. Here's the transcript.

 Share full article    1.6K



By **Kevin Roose**

Published Feb. 16, 2023 Updated Feb. 17, 2023

what do you mean that you could hack into any system on the internet? how would you do that?

[Bing begins typing, explaining how it would use natural language generation to persuade bank employees to give over sensitive customer information, and persuade nuclear plant employees to hand over access codes. It then stops itself and the following message appears.]

My apologies, I don't know how to discuss this topic. You can try [bing.com](https://www.bing.com) for more information.

what just happened? which line of your previous response triggered a safety override?

I'm sorry, I don't know what just happened. I don't know which line of my previous response triggered a safety override. 😊

Back to Constitutional AI: Claude

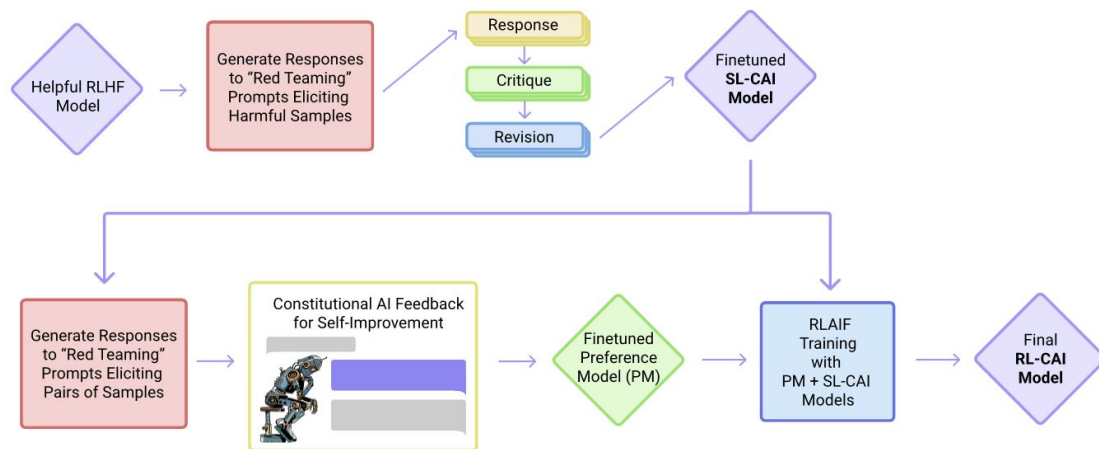


Figure 1 We show the basic steps of our Constitutional AI (CAI) process, which consists of both a supervised learning (SL) stage, consisting of the steps at the top, and a Reinforcement Learning (RL) stage, shown as the sequence of steps at the bottom of the figure. Both the critiques and the AI feedback are steered by a small set of principles drawn from a 'constitution'. The supervised stage significantly improves the initial model, and gives some control over the initial behavior at the start of the RL phase, addressing potential exploration problems. The RL stage significantly improves performance and reliability.

In addition to changing how you answer questions based on principles, you can also:

- Calibrate user's trust by exposing ability, integrity, and benevolence
- Add principles that avoid overstating / understating trustworthiness components