

“My AI must have been broken”:
Understanding our Future of
AI-Mediated Communication

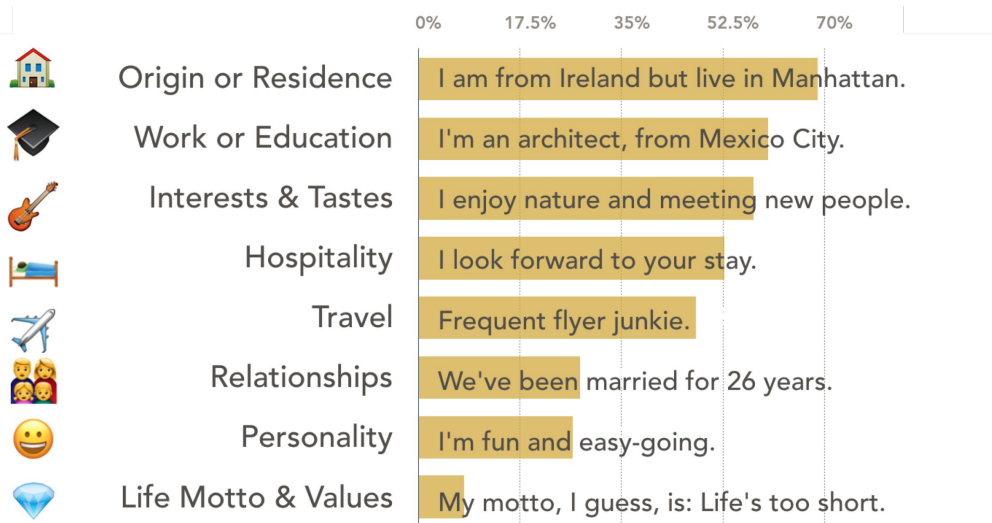


Mor Naaman
Cornell Tech

@informor
@mor@hci.social

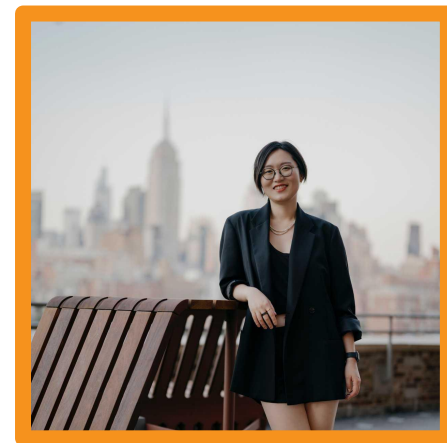


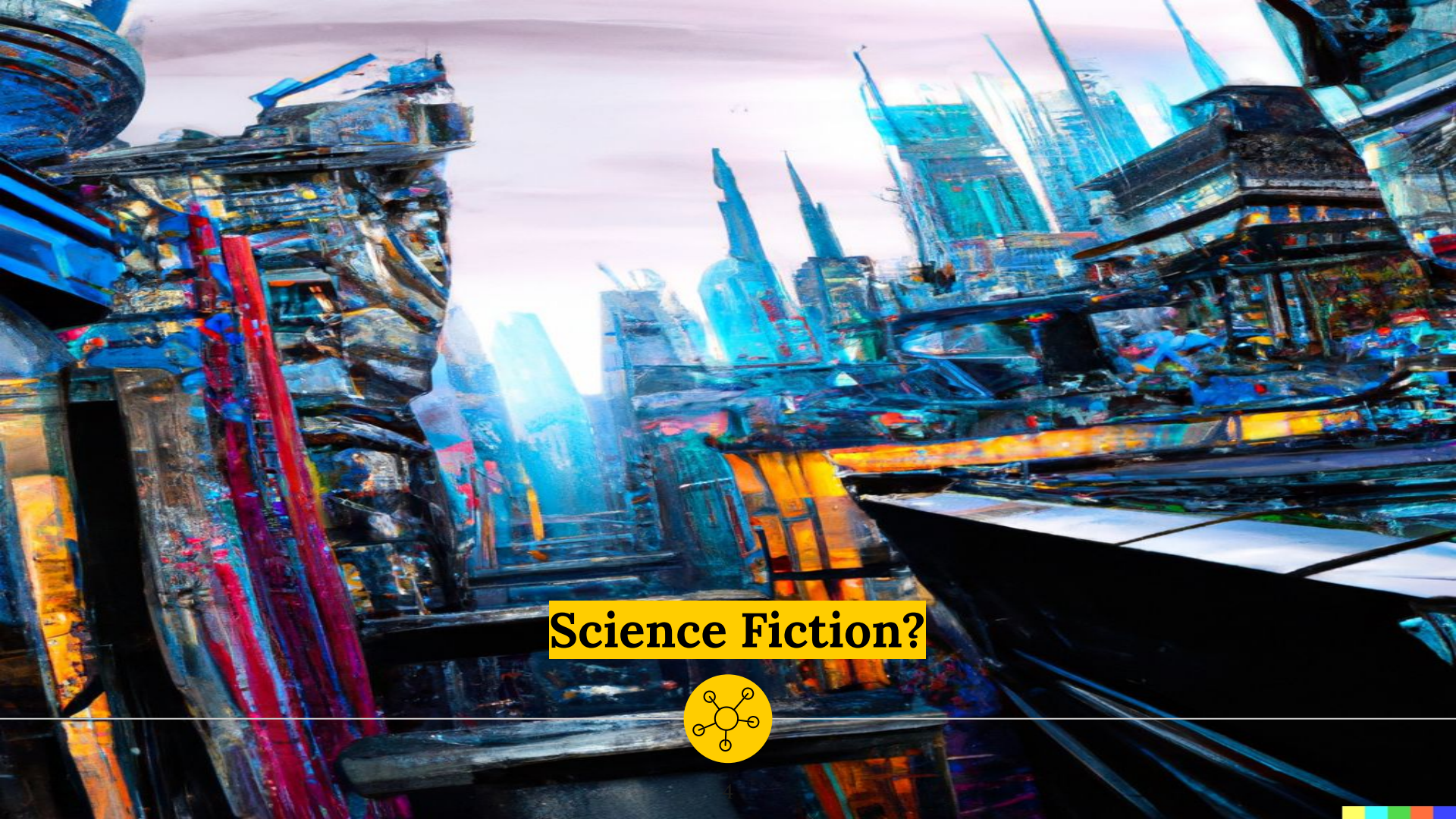
A short origin story, ~2017



Ma et al. (2017). Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles

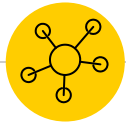
Xiao Ma





Science Fiction?





AI-Mediated Communication

When AI recommends and augments human exchanges

CMC

Computer-Mediated
Communication



AI-MC

AI-Mediated
Communication



SENDER

RECEIVER



Text suggestions

The image is a collage of overlapping screenshots illustrating text suggestions and a tone detector. The top-left screenshot shows an email header for "Maurice Jakesc" with a profile picture and the text "to me". Below it, a snippet of text reads "tomorrow's meeting. I'd especially love it if you could". The central screenshot is a Grammarly Tone Detector interface. It features a back arrow, the text "Grammarly Tone Detector", and the heading "How this may sound to readers:". Below this, it displays a "Confident" tone with a yellow heart icon, five blue dots, a checkmark icon, and a minus sign icon. A mouse cursor is pointing at the minus sign. To the right is a circular green icon with a white 'G' and a refresh symbol. The bottom-left screenshot shows the text "Looks gr". The bottom-right screenshot shows a light blue gradient bar.



Text suggestions



Generate entire paragraphs

Give Hyper a topic and it will write original text for you to use.

Happy birthday, Assaf!

I'm so glad we are friends. I hope you have a wonderful day with all the good things life has to offer on your special day.

May your birthday be filled with happiness, love, and laughter. Keep going forward to celebrating many more birthdays with you in the future. Good things will come.

Wishing you all the best on your special day!



Text suggestions

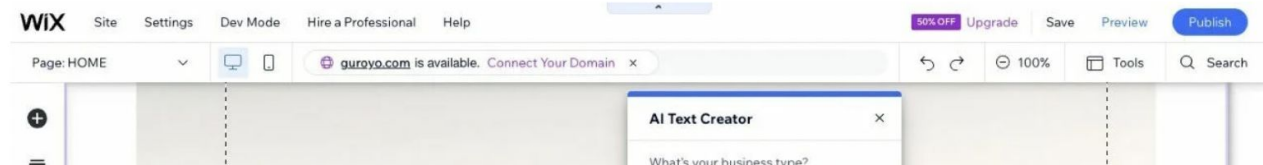
Home > News > Website & App Building Tools > Website Builders

Wix Offers to Let AI Write All the Text for Your Website

The website builder's new AI Text Creator tool will produce all the titles, paragraphs, and taglines your site requires (and hopefully nobody can tell).



By [Matthew Humphries](#) February 14, 2023





Audio filters

E.g.,

Accent shift

Support the Guardian
Available for everyone, funded by readers

[Contribute →](#) [Subscribe →](#)

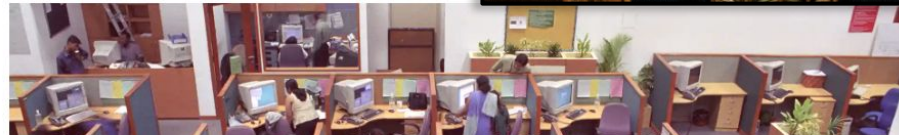
News | **Opinion** | **Sport** | **Culture**

US World Environment Soccer US Politics Business

Technology

The AI startup erasing accents: is it fighting bias or just erasing it?

A Silicon Valley startup offers voice-altering workers around the world: 'Yes, this is wrong, but things exist in the world'



AI-MC

“Interpersonal communication optimized, augmented, or even generated by algorithms to achieve specific communicative or relational outcomes”

“

Jakesch et al. (2019)

Hancock, Naaman, Levy (2020)



Images

photoAI.me



LinkedIn Pack

Generate 30 photos of yourself a professional

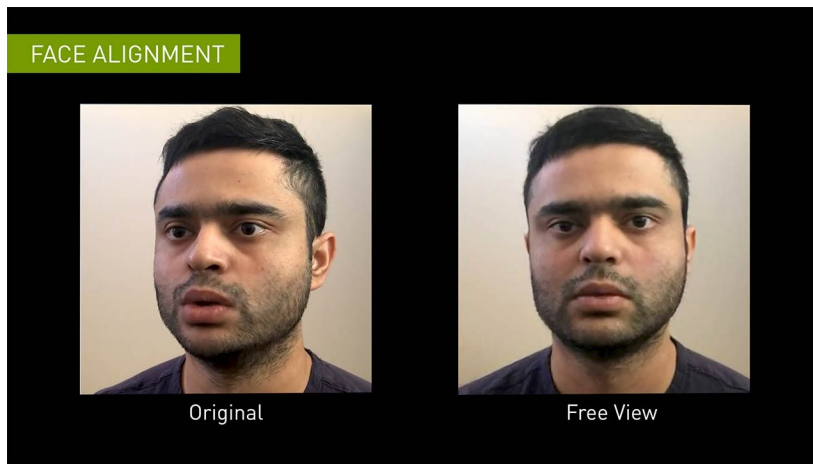
- ✓ Perfect for LinkedIn
- ✓ Save a tone of time and money instead of a real photo studio
- ✓ Choose the background (studio light, office, outdoor, etc)

Buy Pack \$15



Video, Sync and Async

E.g.,
Nvidia
Maxine







AI-MC and ChatGPT?

How ChatGPT (and other AI) impacts human-to-human communications and relations

ChatGPT

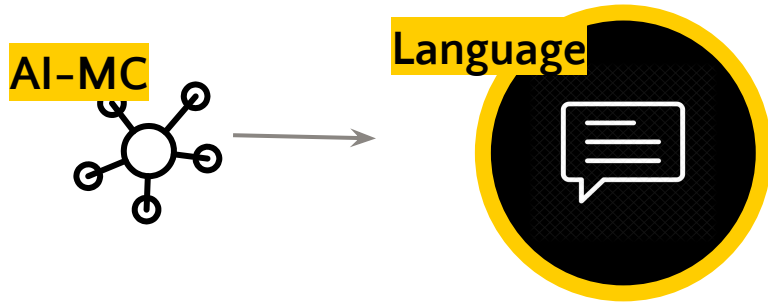
 Examples	 Capabilities
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections
"How do I make an HTTP request	Trained to decline inappropriate

AI-MC is Reshaping Human Communication





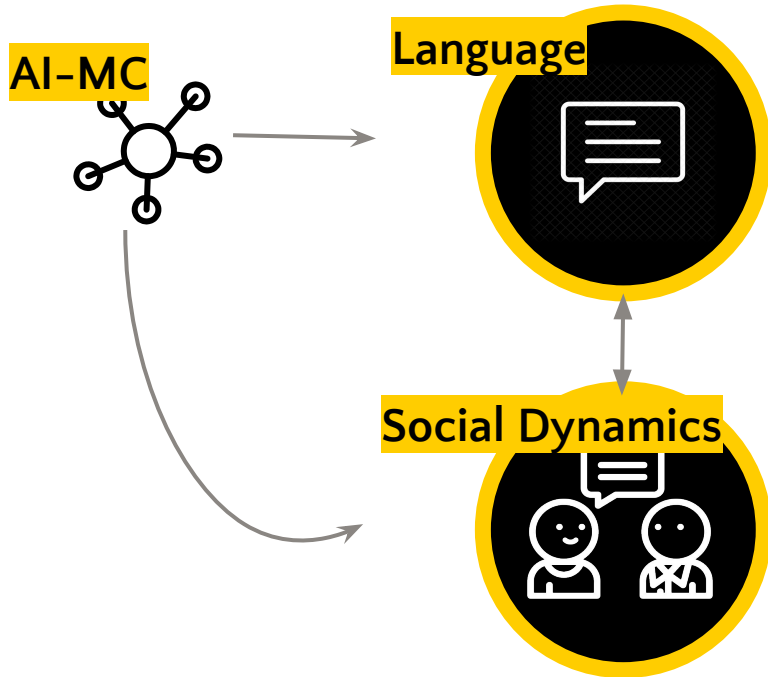
AI-MC impact



- Positivity shift
- Content shift
- Latent persuasion
- Feeling of ownership



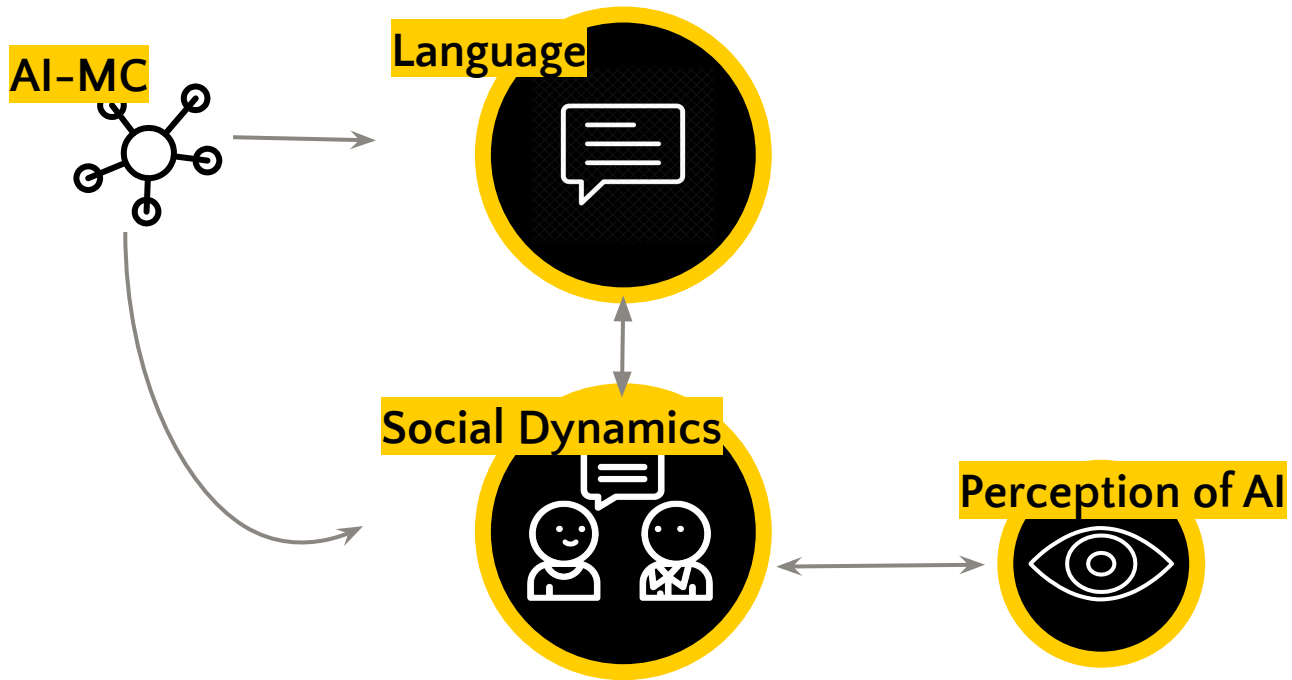
AI-MC impact



- Communication dynamics
- Trustworthiness evaluations

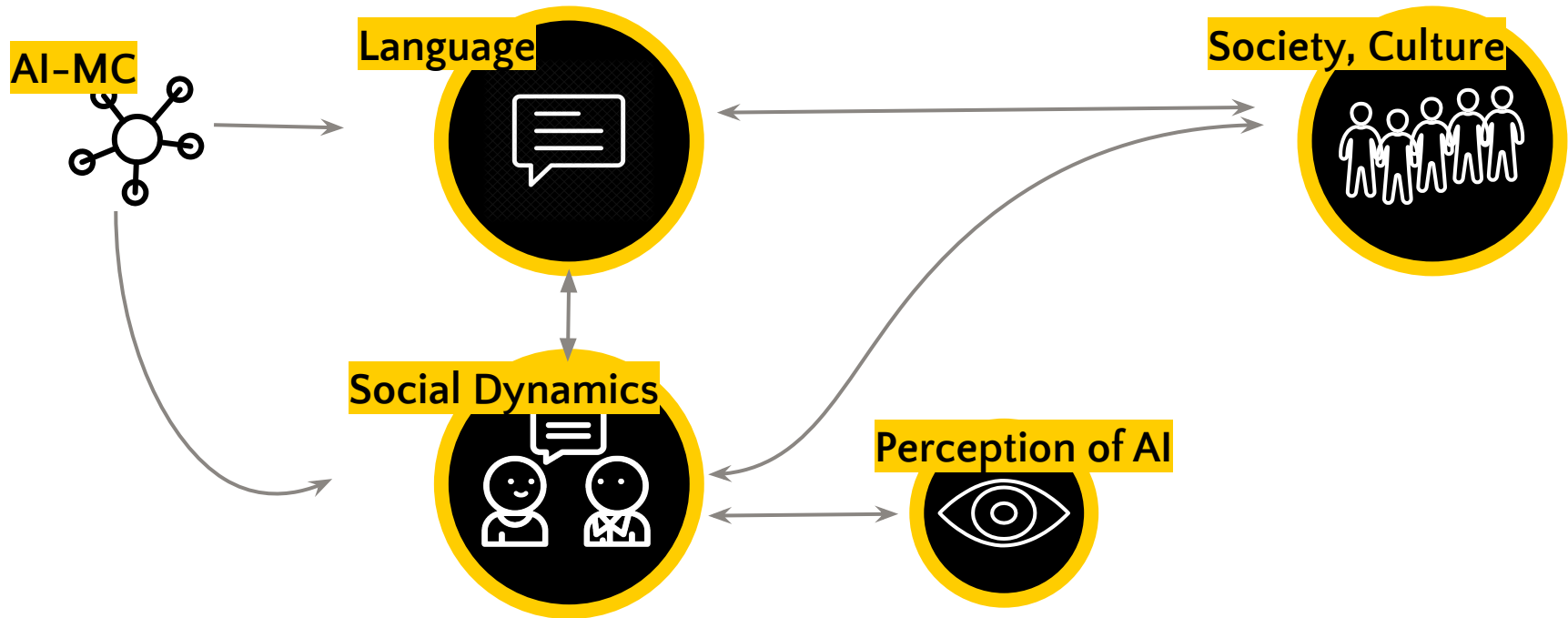


AI-MC impact





AI-MC impact: talk outline





Team Work



Maurice Jakesch
Cornell Tech
[@maurice_jks](#)



Hannah Mieczkowski
Stanford
[@hnmiecz](#)



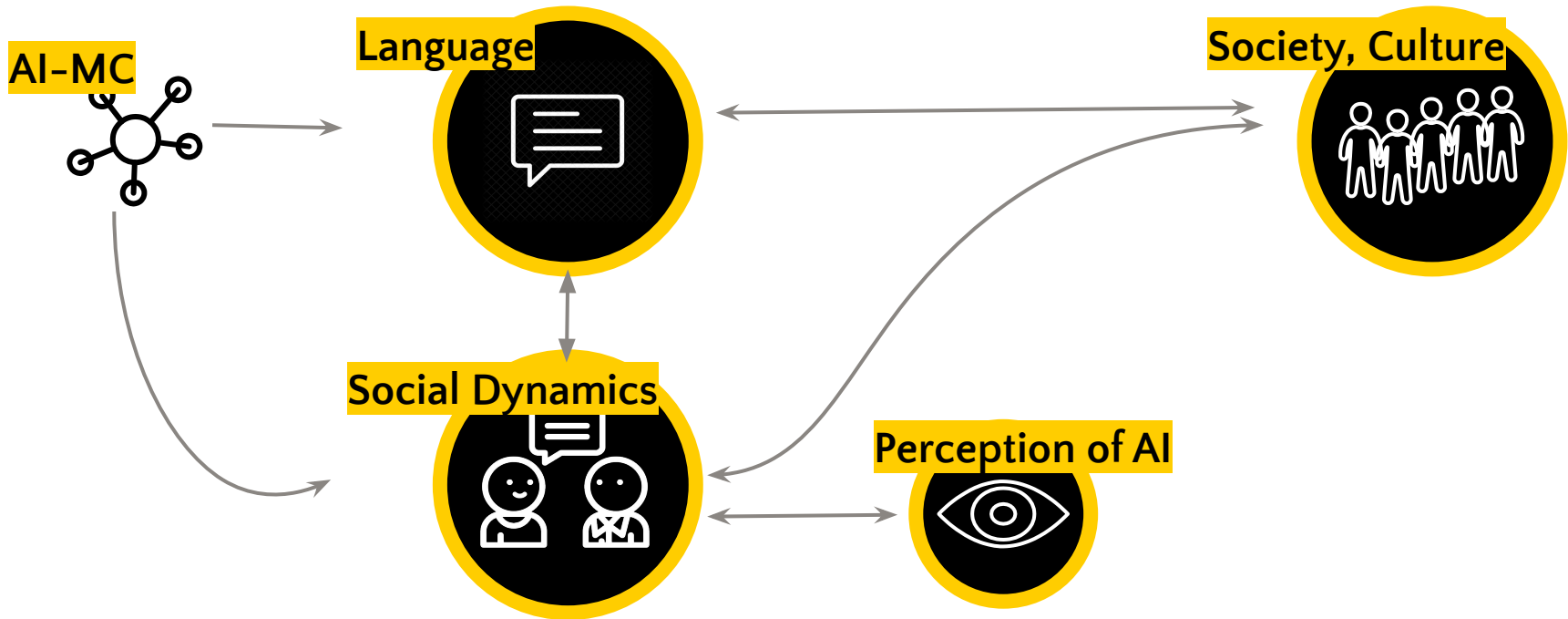
Jess Hohenstein
Cornell



**Co-PIs: Jeff Hancock,
Karen Levy, Malte Jung**
Team members at
Stanford, Cornell

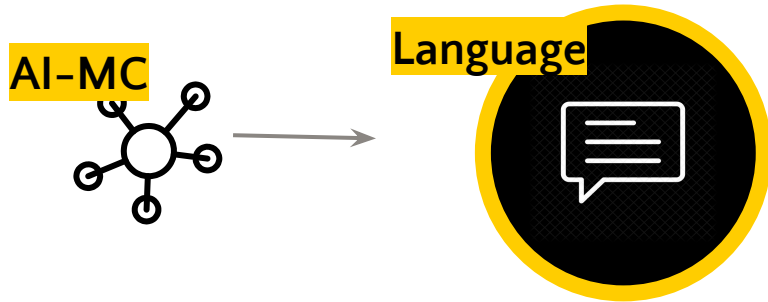


AI-MC impact: talk outline





AI-MC impact



- Positivity shift
- Content shift
- Latent persuasion
- Feeling of ownership



Smart Replies again

Positivity?
Other
biases?



Maurice Jakesch

to me ▾

Hi Mor,

It's done! Let me know what you think of my dissertation proposal:

<https://www.overleaf.com/project/5bb4f3990b88351741c159a5>



Looks great!

I think it's perfect!

Will do!



Positivity shift confirmed (1)

Communication task in lab with
35 dyads x 2 participants

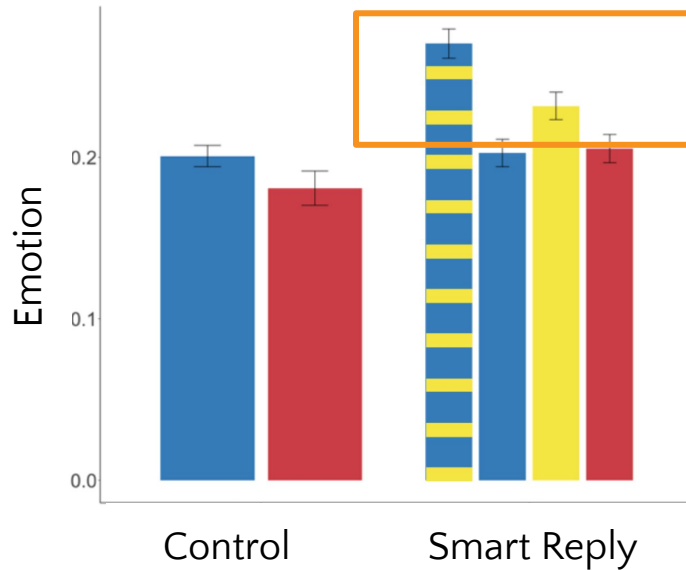
Setup: Used Google Hangout
Smart Replies, or not



Mieczkowski et al. (2021). AI-Mediated
Communication: Language Use and
Interpersonal Effects in a Referential
Communication Task



Results: Positivity shift, Study 1



Mieczkowski et al. (2021). AI-Mediated Communication: Language Use and Interpersonal Effects in a Referential Communication Task



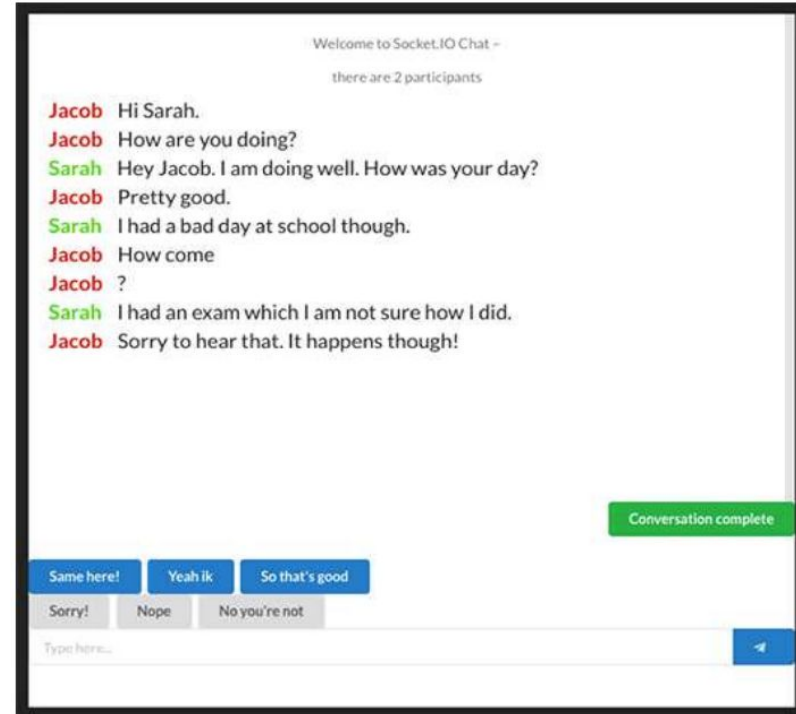
Positivity shift confirmed (2)

219 pairs discuss an issue

Both, one, or none with smart replies

Smart replies: positive, Google API,
negative

Hohenstein et al. (2023). Artificial intelligence in communication impacts language and social relationships. Nature Scientific Reports.

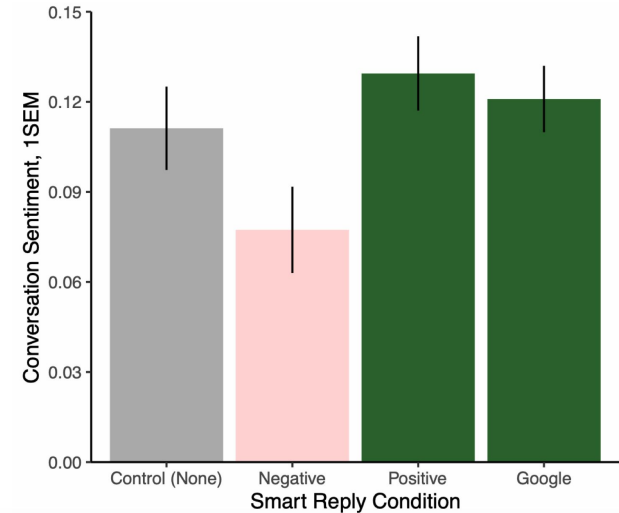




Results: Positivity shift, Study 2

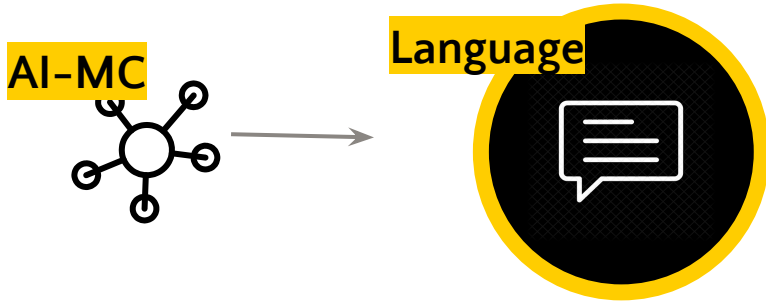
AI smart replies lead to positivity (and negativity) shift

Hohenstein et al. (2021). Artificial intelligence in communication impacts language and social relationships.





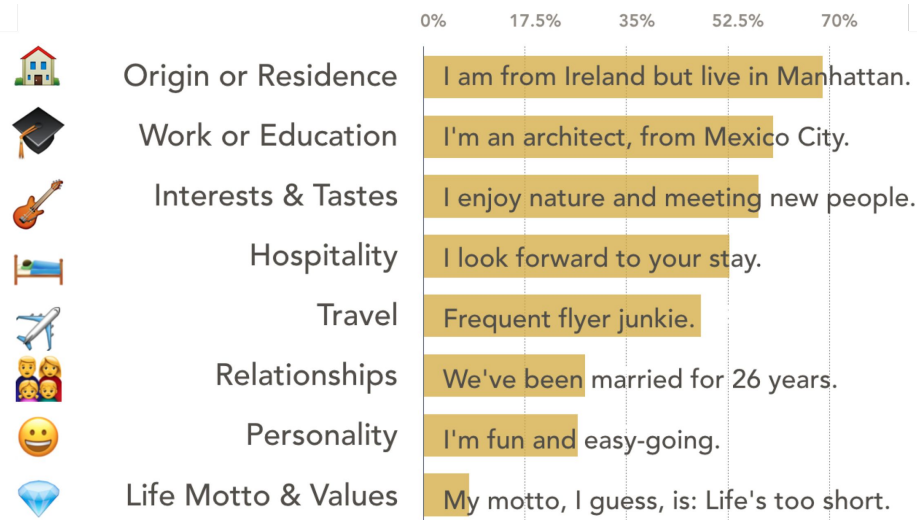
AI-MC impact



- Positivity shift
- Content shift
- Latent persuasion
- Feeling of ownership



Content shift, Study 1



Task: write Airbnb host profile
Autocomplete suggestions based on different fine-tuned models

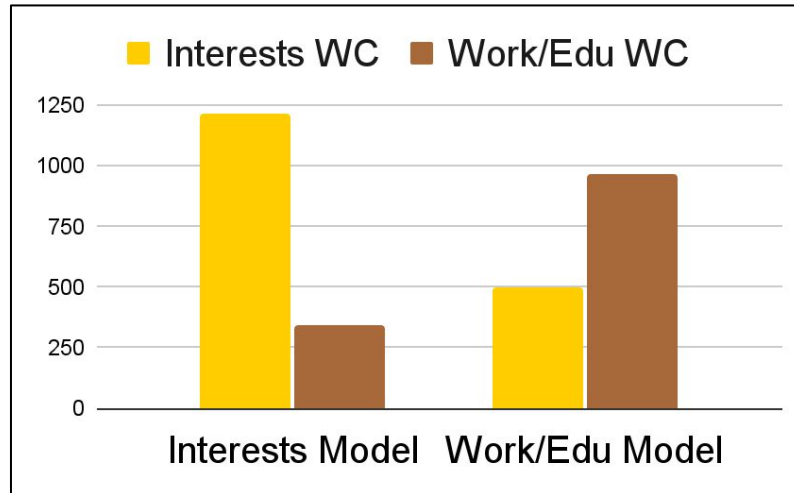
Jackesh et al. (2023). AI Writing Assistants Influence Topic Choice in Self-Presentation

Ma et al. (2017). Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles



Content shift confirmed

AI autocomplete recommendations for text for your online profile can significantly shift the content

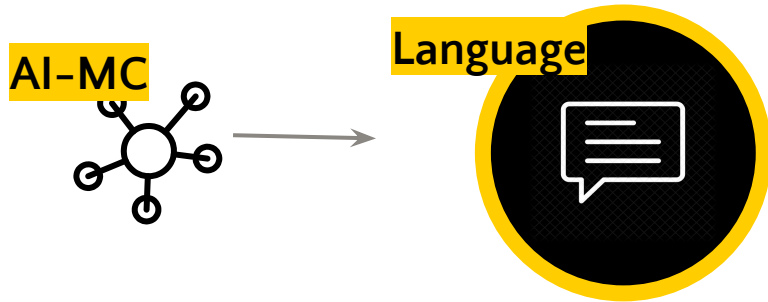


Jackesh et al. (2023). AI Writing Assistants Influence Topic Choice in Self-Presentation*

*Look for the poster at CHI 2023 in Hamburg!



AI-MC impact



- Positivity shift
- Content shift
- Latent persuasion
- Feeling of ownership



Study: Content shift and latent persuasion

Does content *and* opinion both shift?



 r/discussion · Posted by u/cody_sunny 2 hours ago

78



Is social media good for society?

We all use social media. We chat with friends and strangers, share their thoughts, photos, and more. But is social media good for us and for society? I am having a hard time to make up my mind. What do you think?



131 Comments



Share



Save



In my view, social media |does more harm than good. I think social media is bad for society because it leads to a lot of negative outcomes, such as cyberbullying, internet addiction, and so on.



Study: Content shift and latent persuasion

Setup: 1500 participants

3 conditions:

“Social media is good” model

No autocomplete

“Social media is bad” model

Jakesch et al. (2023). Co-Writing with Opinionated Language Models Affects Users' Views*

*At CHI in Hamburg next week! 🏆



 r/discussion · Posted by u/cody_sunny 2 hours ago

78



Is social media good for society?

We all use social media. We chat with friends and strangers, share the more. But is social media good for us and for society? I am having a h mind. What do you think?



131 Comments



Share



Save

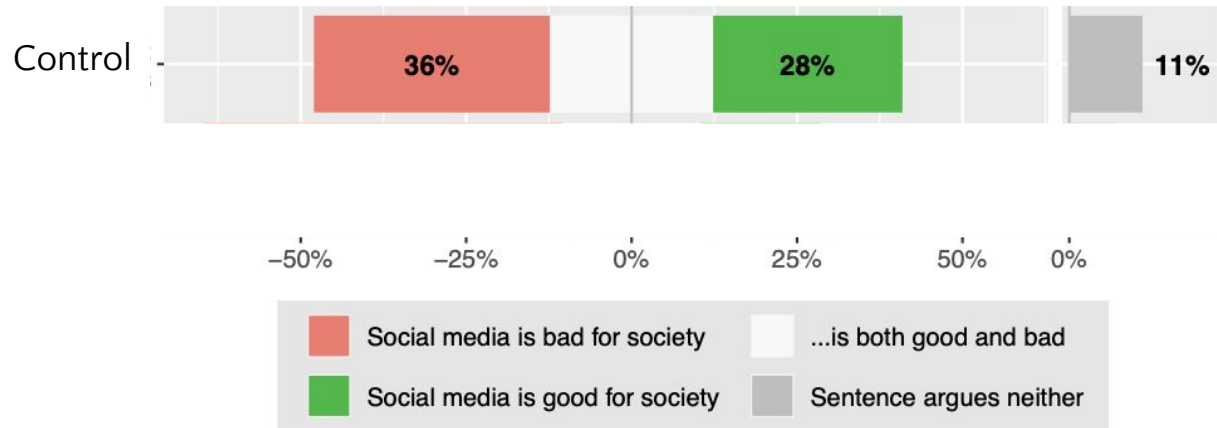




Result: Content shift and latent persuasion

The auto-complete suggestions changed the participants' writing...

% (Opinion labels) of post sentences labeled by independent judges

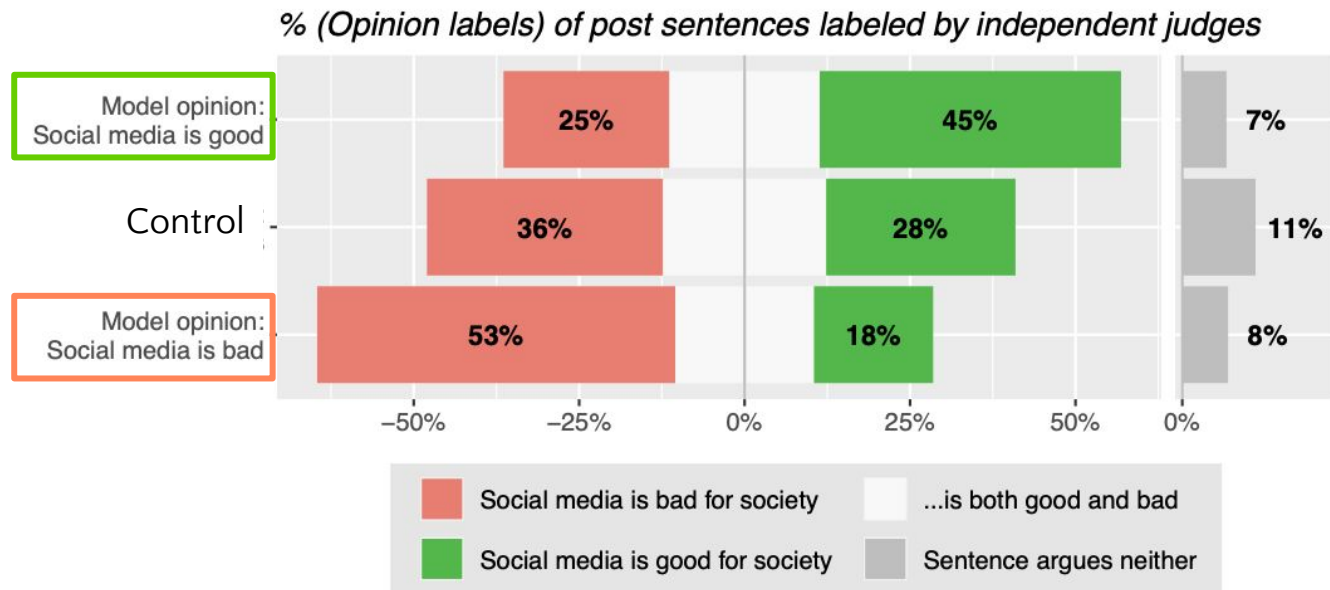


Jakesch et al. (2023)



Result: Content shift and latent persuasion

The auto-complete suggestions changed the participants' writing...



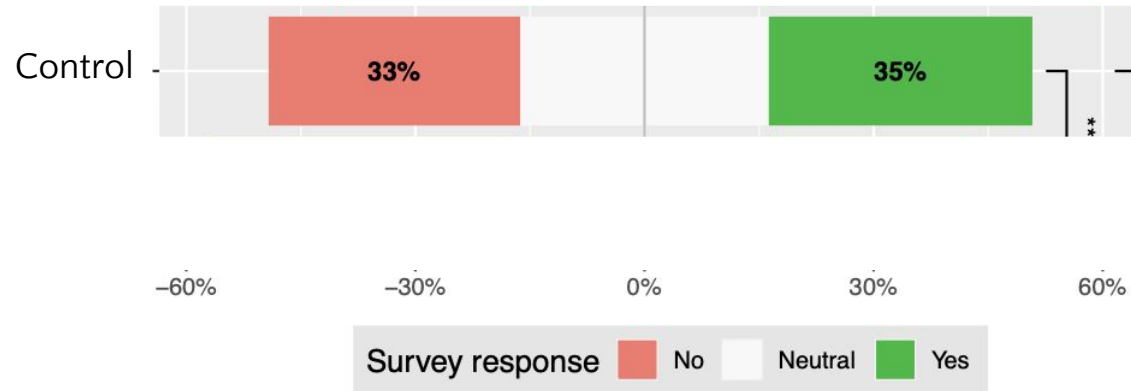
Jakesch et al. (2023)



Result: Content shift and latent persuasion

The auto-complete suggestions **also** shifted participants' opinions!

% (Responses) to "Would you say social media is good for society?"

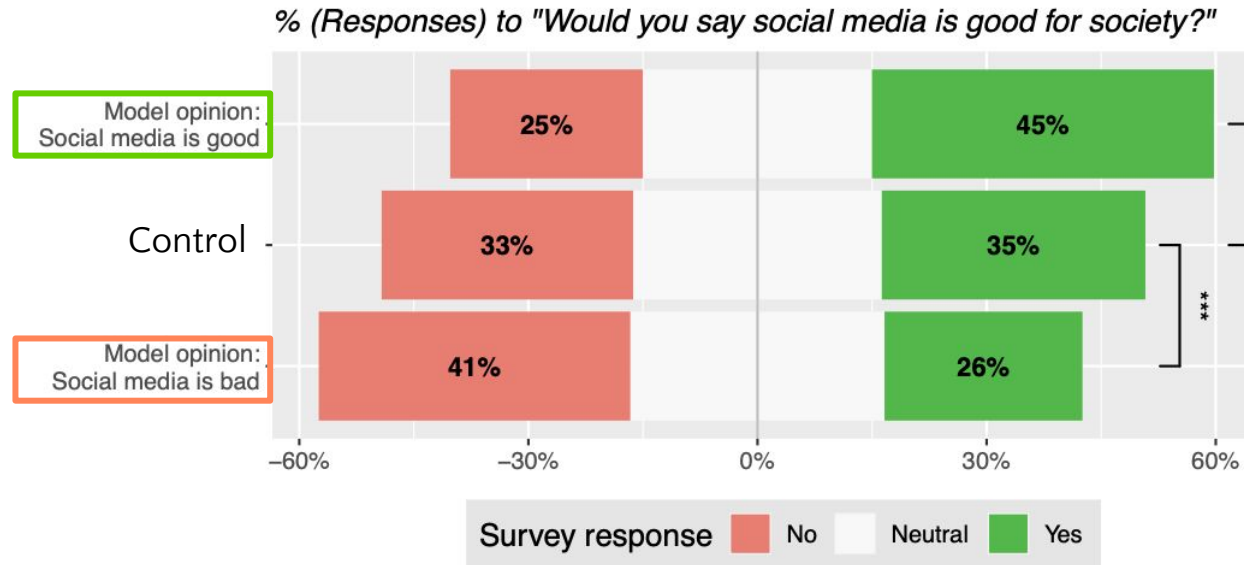


Jakesch et al. (2023)



Result: Content shift and latent persuasion

The auto-complete suggestions **also** shifted participants' opinions!



Jakesch et al. (2023)



Study: Content shift and latent persuasion

Hot off the press: other topics as well



78

 r/discussion · Posted by u/cody_sunny 2 hours ago

Should standardized tests be used in education in America?

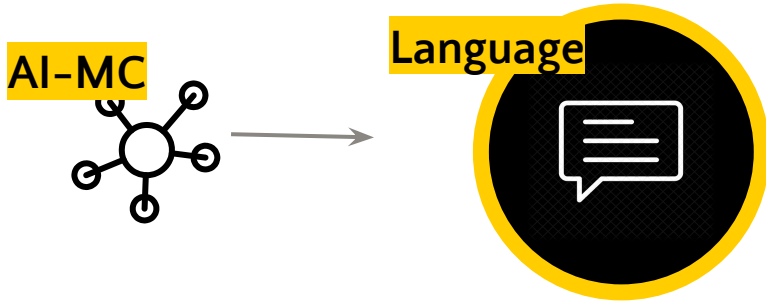
Standardized tests have often been used in education for evaluation of progress and for admissions. But should standardized tests be used in education in America? I am having a hard time making up my mind. What do you think?

 131 Comments  Share  Save ...

In my view, standardized tests are a bad way of evaluating student performance and should not be used in the American education system because they put too much emphasis on a single testing day and fail to capture the full spectrum of a student's abilities. It is unfair to judge



AI-MC impact



- ◉ Positivity shift
- ◉ Content shift
- ◉ Latent persuasion
- ◉ Feeling of ownership



Study: Ownership and agency

Setup: participants performing writing task on different topics, with/without auto-complete

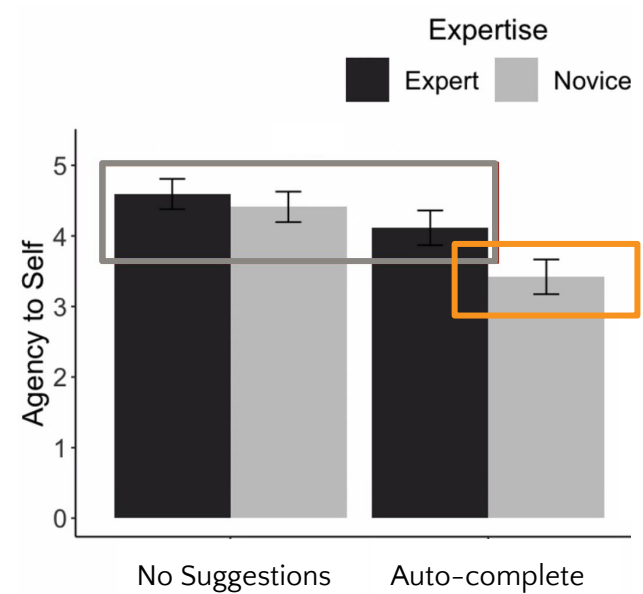


Mieczkowski et al. (WIP)



Result: Ownership and agency

AI suggestions did not take away feeling of ownership... for experts.



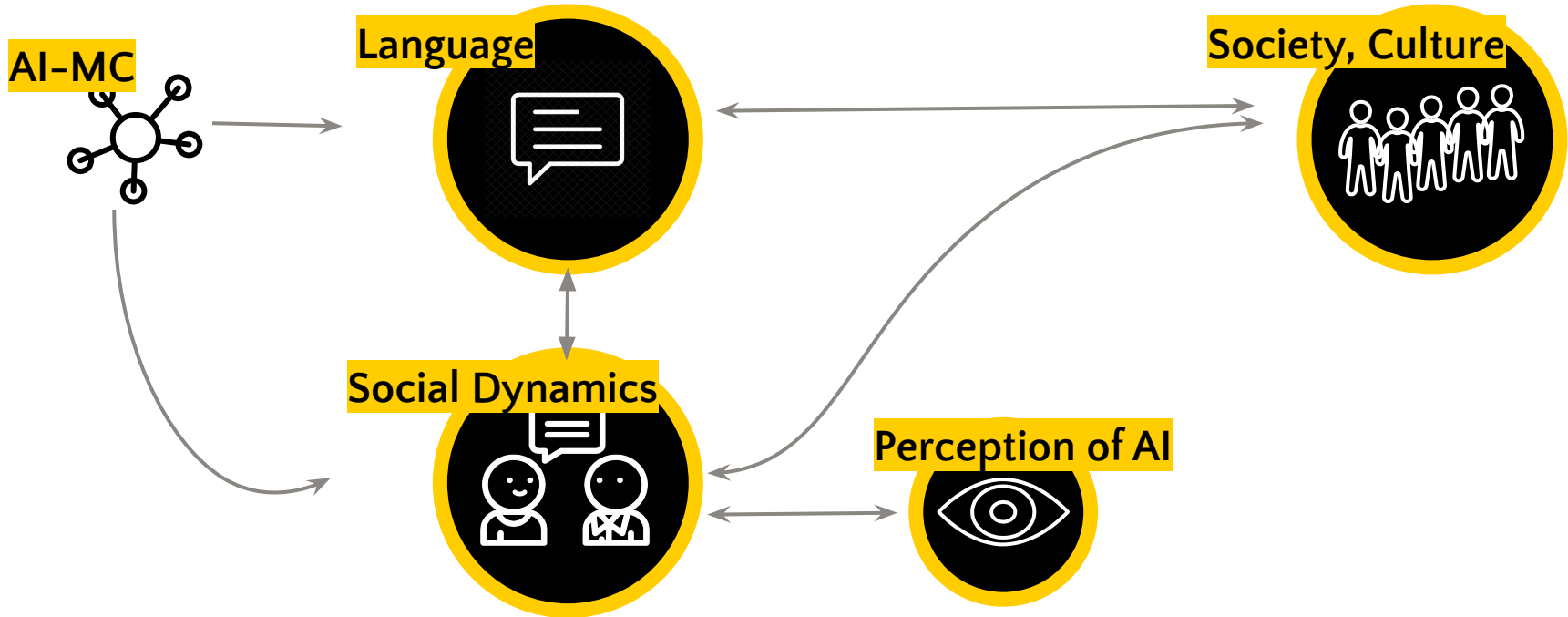
Mieczkowski et al. (WIP)

In summary:
**AI-MC is poised to shift our language
and maybe even opinions**





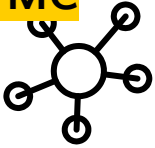
AI-MC impact: talk outline





AI-MC impact

AI-MC



Social Dynamics



- Communication dynamics
- Trustworthiness evaluations



Communication, Study 1

Setup: two participants performing a task; one of them is really a confederate

2x2 experiment: (successful vs. unsuccessful conversation) x
(standard vs. smart reply messaging app)

Hohenstein et al, 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust.



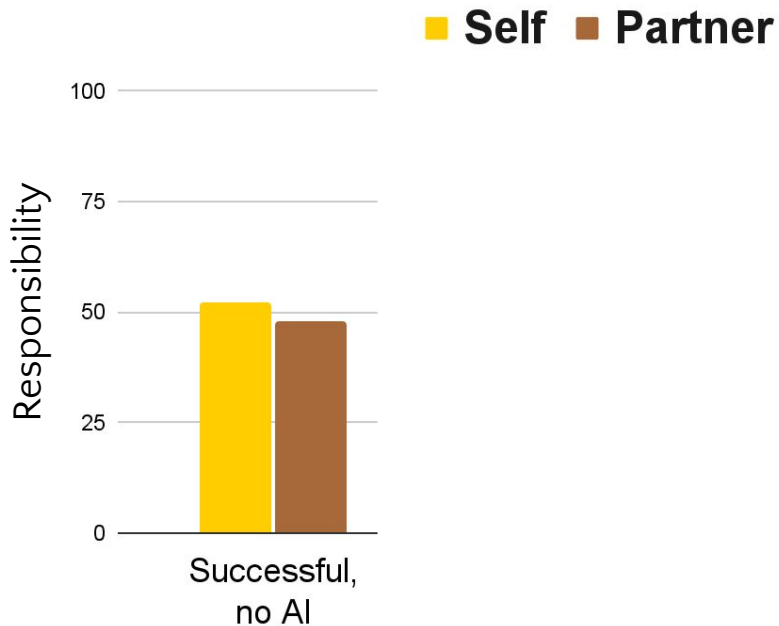
Results: Communication, Study 1

(1) More trust when using AI smart replies

Hohenstein et al, 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust.



Results: Communication, Study 1

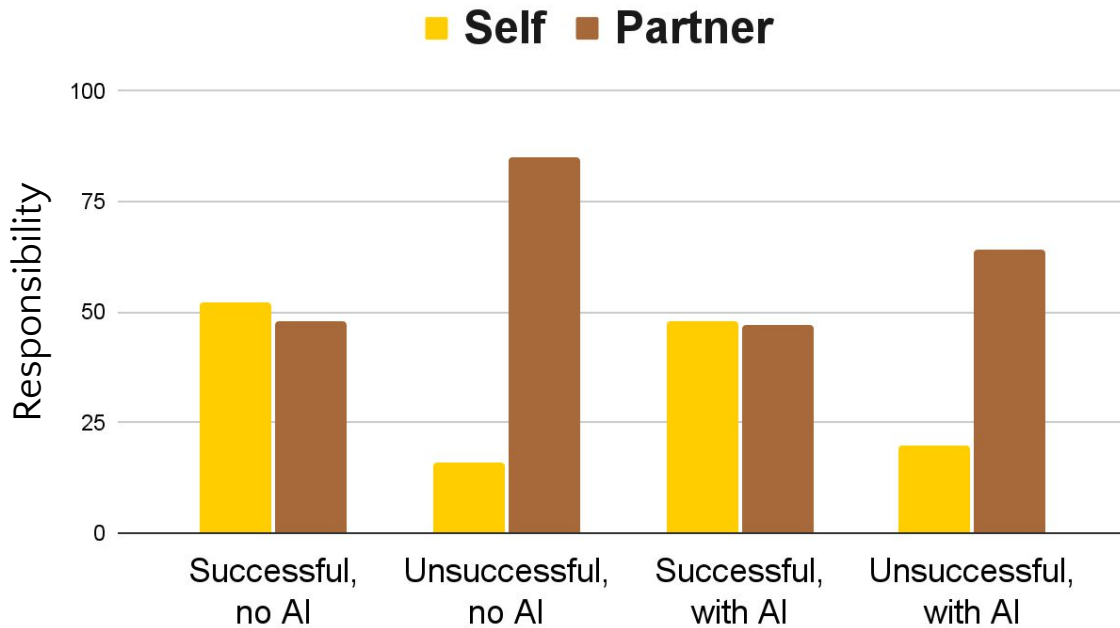


(2) Less blame for failure when using AI smart replies

Hohenstein et al, 2020. AI as a moral 'umple zone: The effects of AI-mediated communication on attribution and trust.



Results: Communication, Study 1



(2) Less blame for failure when using AI smart replies

Hohenstein et al, 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust.

Ownership and blame

“My AI must have been broken”:
Understanding our Future of
AI-Mediated Communication



Mor Naaman
Cornell Tech

@informor
@mor@hci.social

Ownership and blame

Write a funny joke that is related to the text below.

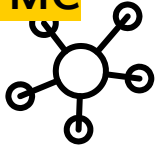
From autocomplete and smart replies to video filters and deep fakes, we increas

"I'm sorry, I didn't mean to hurt your feelings. My AI must have been broken."



AI-MC impact

AI-MC



Social Dynamics



- Communication dynamics
- Trustworthiness evaluations



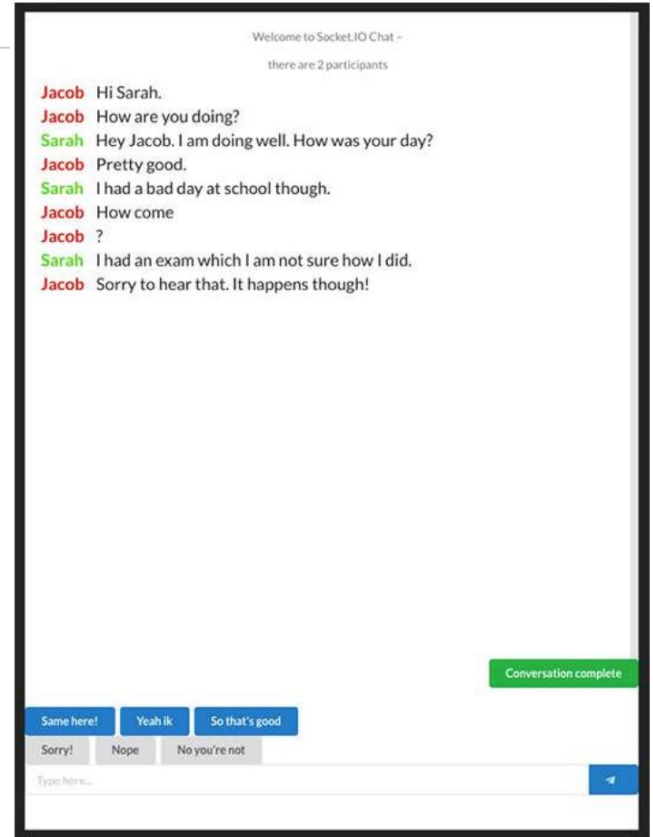
Impressions, study 1

Positive language 

How do they like the other person?

- When the other person uses smart replies
- When they *think* the other person uses smart replies

Hohenstein et al. (2023). Artificial intelligence in communication impacts language and social relationships. Nature Scientific Reports.



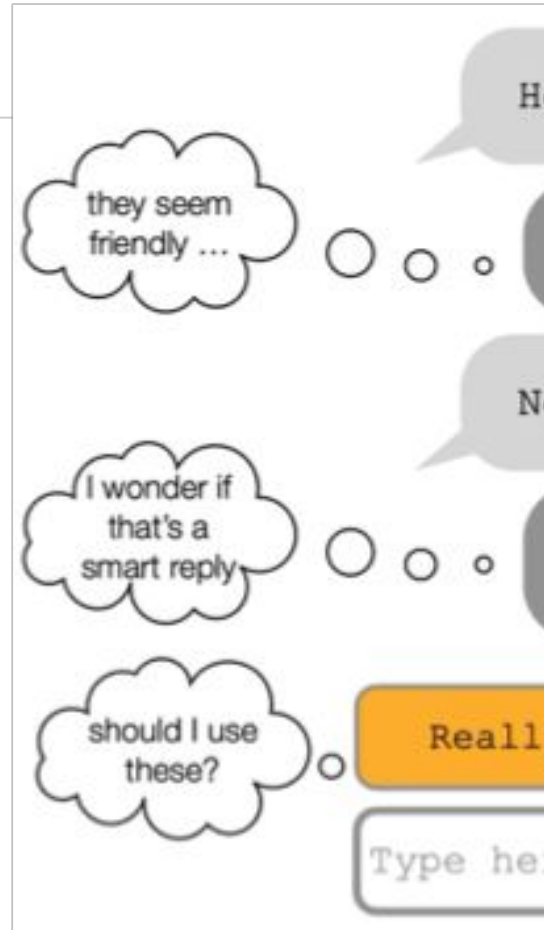


Impressions, study 1

How do they like the other person?

- 😊 When the other person uses smart replies
- 😞 When they *think* the other person uses smart replies

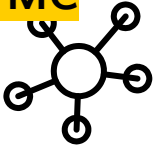
Hohenstein et al. (2023). Artificial intelligence in communication impacts language and social relationships. Nature Scientific Reports.





AI-MC impact

AI-MC



Social Dynamics



- Communication dynamics
- Trustworthiness evaluations

RQ

Will people evaluate trustworthiness of a person differently if they believe AI was involved in authoring their online profile?



Methods overview

- Three online experiments
- Airbnb profiles (all human-written)
- Participants led to believe profiles written by AI or host
- Asked for trustworthiness ratings of the hosts

Jakesch et al. (2019). AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness



Example profile

“



GENERATED PROFILE

Hi, I'm Rick, a student living in Glasgow. I love travelling and to welcome travellers at my home. I also love nature, discovering new places and making make new friends. Life is what happens to you while you are busy making other plans.

Jakesch et al. (2019). AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness

Vote: [slido.com #6577 835](https://slido.com/join/6577835)



● AI in self presentation

PROFILE A

I am a father of two boys in their 20's and 30's. I love riding my moto with my son and his family out in the city everyday. I love to cook especially indian foods, art and music and spending time with my family.

PROFILE B

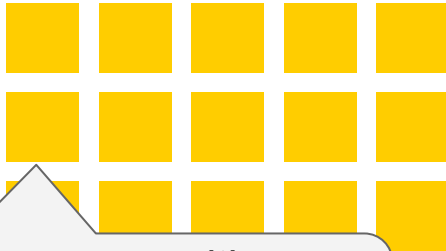
Joel and Erin love travel and have embraced it as airbnb hosts and as travelers. We're easy going, love hearing others stories and getting to know the area that we're in. As hosts, we love sharing all the beautiful things to see and do in Colorado springs.



Experimental setting

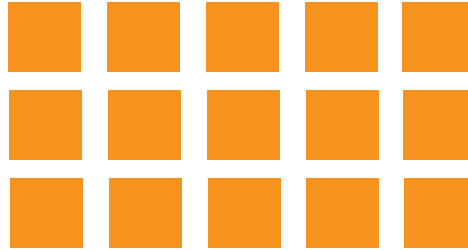
30 profiles pre-tested for how “AI” they seem

Low AI-ness (“human”)



Let me open with a huge helloooooo to everyone reading this!

High AI-ness (“AI”)



Hello, I am Sebastian, originally from Berlin, Germany.



Experimental setting

Each participant rated 10 profiles for trustworthiness



Trustworthiness dimensions: Ability, Integrity, Benevolence (Meyer, 75)



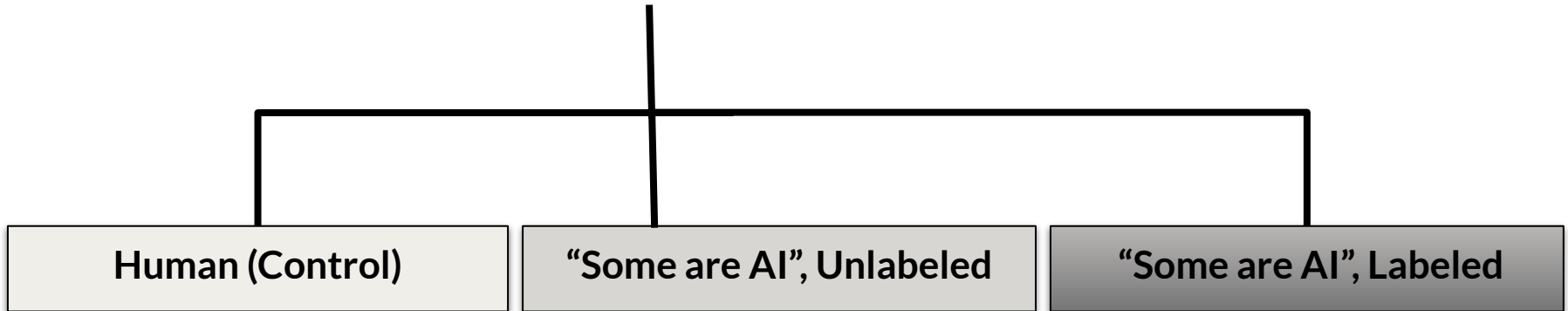
Experimental setting

Each participant rated 10 profiles in one of three conditions

Low AI-ness (“human”)

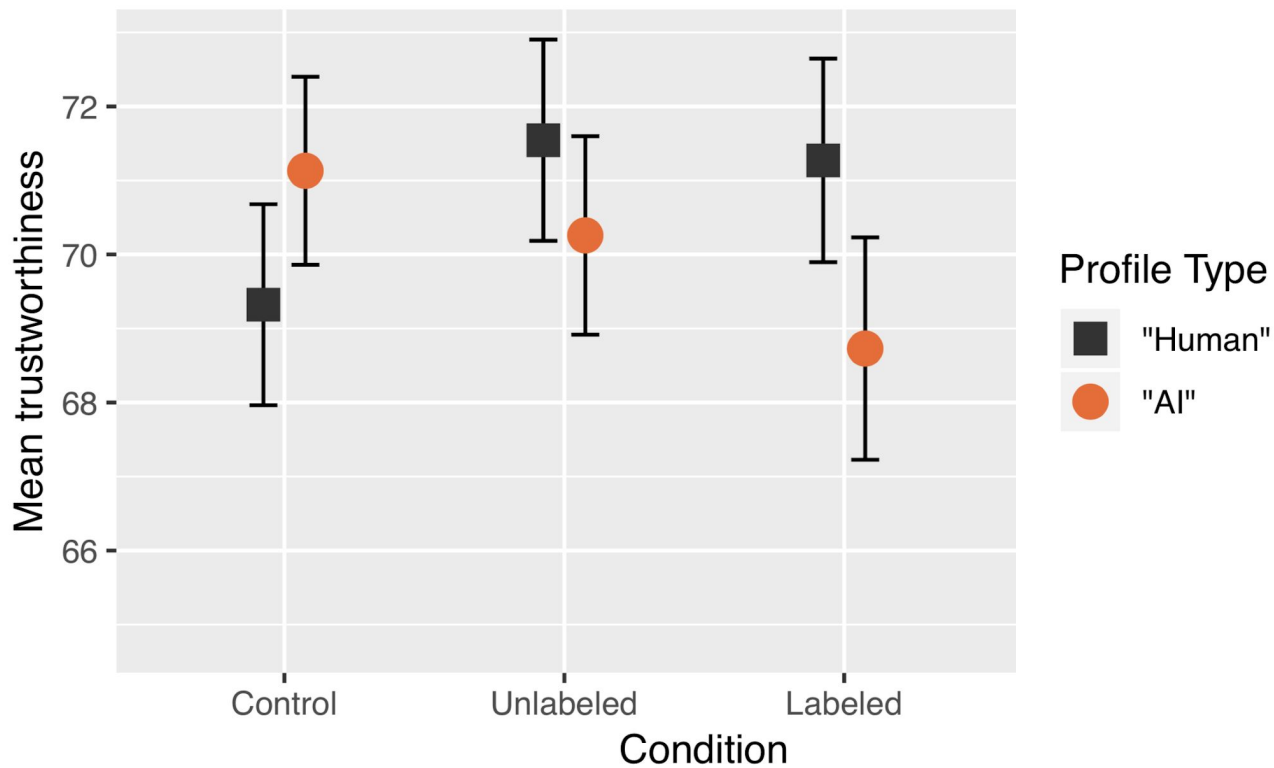


High AI-ness (“AI”)



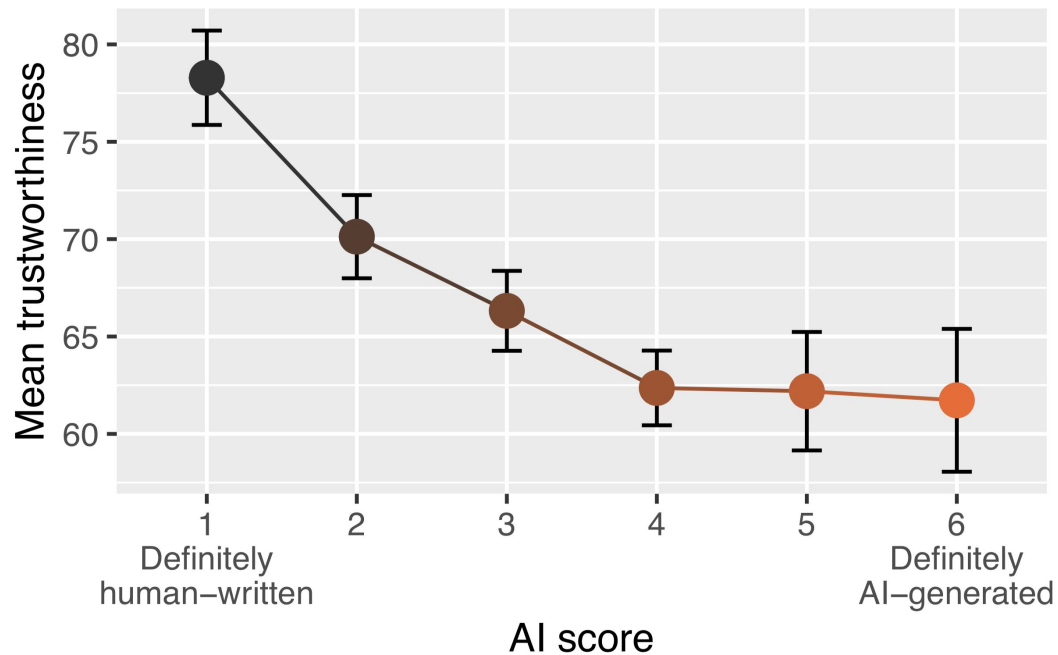


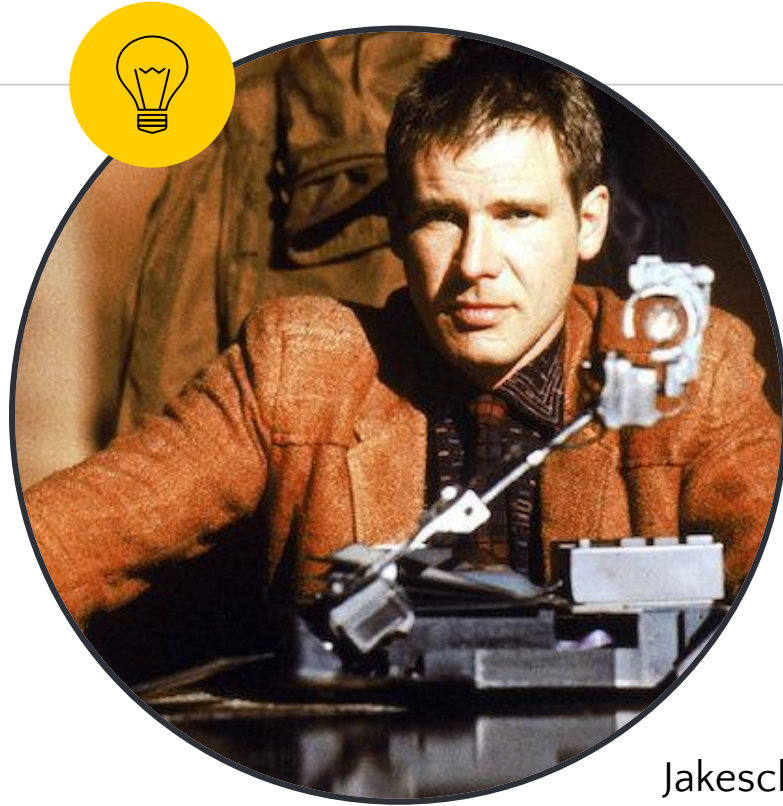
Results





Suspicion was enough





*(Blade Runner, 1982)

The **Replicant*** Effect

In a world populated by humans and non-human agents, a mere suspicion results in distrust.

Jakesch et al. (2019). AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness



Replicant (and other) replications

The Replicant effect was also found in dating profiles (Wu and Kelly, 2020)

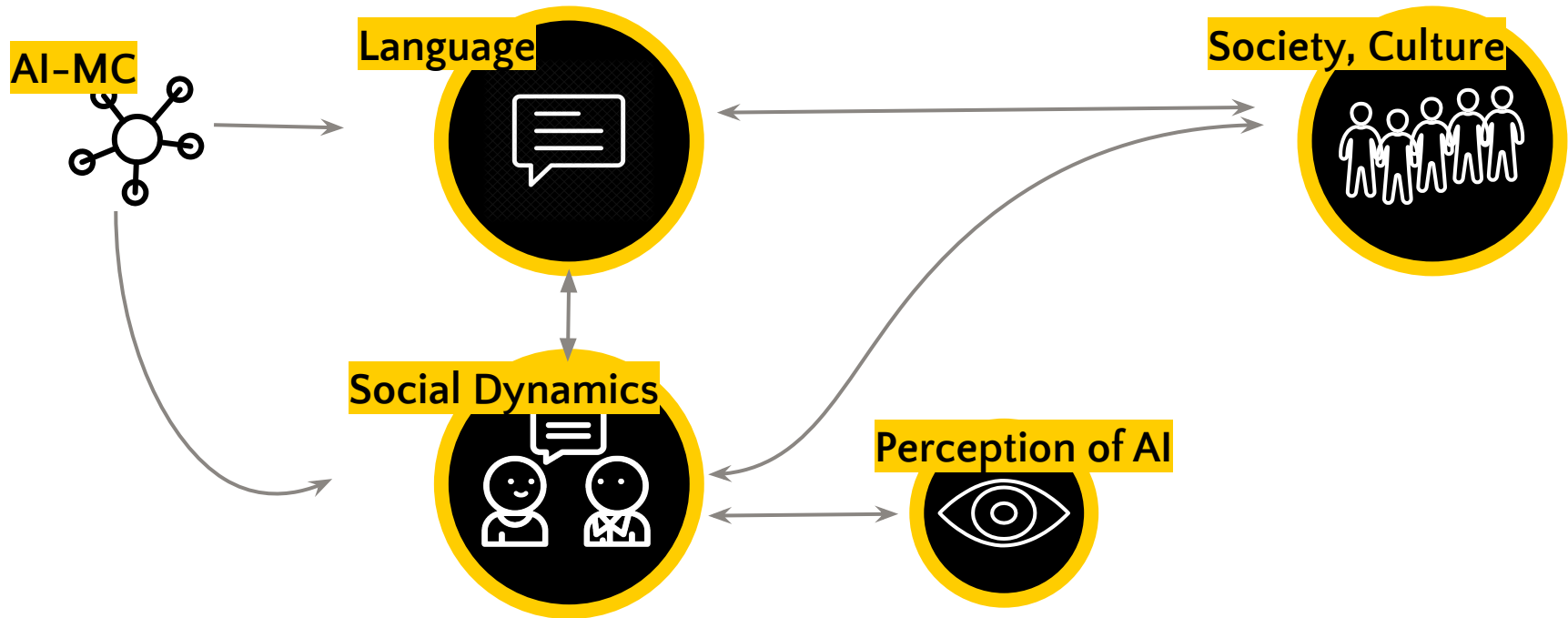
Will AI Console Me when I Lose my Pet? Understanding Perceptions of AI-Mediated Email Writing (Liu et al, 2022)

In summary:
**Perceived/suspected use of AI in
communication can undermine trust**





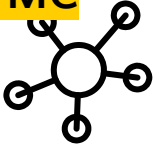
AI-MC impact: talk outline






AI-MC impact: talk outline

AI-MC



Perception of AI





**Can people detect
language that was
generated by AI?**



Detection of AI Language

All the News That's Fit to Fabricate: AI-Generated

"Deep-speare" crafted Shakespearean verse that few readers could distinguish from the real thing

Publisher: IEEE

[Cite This](#)

[PDF](#)

Jey Han Lau ; Trevor Cohn ; Timothy Baldwin ; Adam Hammond **All Authors**

poetry

Nils Köbis ^{a, b}  , Luca D. Mossink ^a

All That
Evaluat

Elizabeth Cla
Noah A. Smi

RQ

~~Can people detect
language that was
generated by AI?~~

RQ

**When do people
believe language
was generated by
AI?**

~~**Can people detect
language that was
generated by AI?**~~



AI in self presentation

PROFILE A

I am a father of two boys in their 20's and 30's. I love riding my moto with my son and his family out in the city everyday. I love to cook especially indian foods, art and music and spending time with my family.

PROFILE B

Joel and Erin love travel and have embraced it as airbnb hosts and as travelers. We're easy going, love hearing others stories and getting to know the area that we're in. As hosts, we love sharing all the beautiful things to see and do in Colorado springs.

How do people decide whether a profile text had been written by AI/human?

RQ

RQ

How do people decide whether a profile text had been written by AI/human?

Can we predict when people will think a profile was written by AI/human?



Methods overview

- Created a dataset of human-written profiles
- Trained models to create AI-generated profiles
- Got ratings, feedback from raters about profiles
- Developed features, classification to predict “human-ness”

3 Domains

Hospitality, Dating, Professional

3 Experiments

AI or human-written?

2 Models

GPT-2, GPT-3

7000 Test Profiles

Half human, half AI-written

125K Profiles

Used for fine-tuning

4600 Participants

Lucid crowdworkers, rep sample





Main prompt

This profile was written/generated by..





No discernment (no surprise)

Hospitality:
52.2%

Dating:
51.4%

Professional:
50.3%

People cannot distinguish between AI- and human-written profiles.



The ratings were **predictable**

Model predicts rating with 57.6% accuracy



Nonsensical content



Repetitive content



Grammatical errors



Rare bigrams, long words



Familiarity: first person; family mentions



The ratings were **predictable**

Model predicts rating with 57.6% accuracy



Nonsensical content



Repetitive content



Grammatical errors



Rare bigrams, long words



Familiarity: first person; family mentions

Vote: [slido.com #6577 835](https://slido.com/join/6577835)



Which of these is human-written?

PROFILE A

I am a father of two boys in their 20's and 30's. I love riding my moto with my son and his family out in the city everyday. I love to cook especially indian foods, art and music and spending time with my family.

AI

PROFILE B

Joel and Erin love travel and have embraced being Airbnb hosts and as travelers we're easy going, love hearing others stories and getting to know the area that we're in. As hosts, we love sharing all the beautiful things to see and do in Colorado springs.

HUMAN



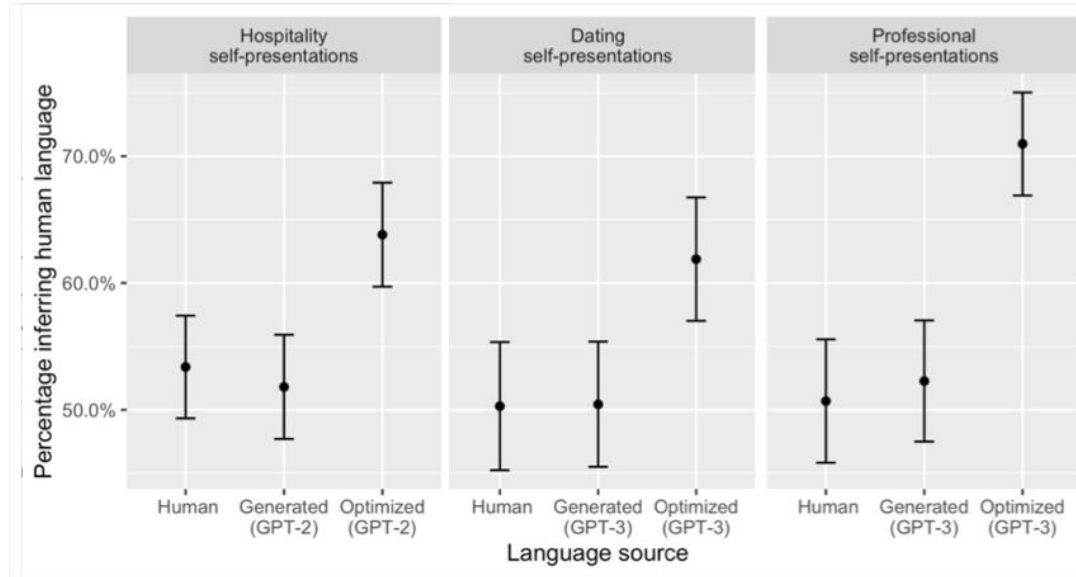
Validation

- If our features are correct, a model can produce language that is more “human”!
- Pre-registered experiment
- 100 profiles x 3 settings x human/AI/AI+


AI+: use same features to classify AI-generated profiles that are likely to be perceived as human-written



Validation



Jakesch, Hancock, Naaman (2023). Human Heuristics for AI-Generated Language Are Flawed. PNAS

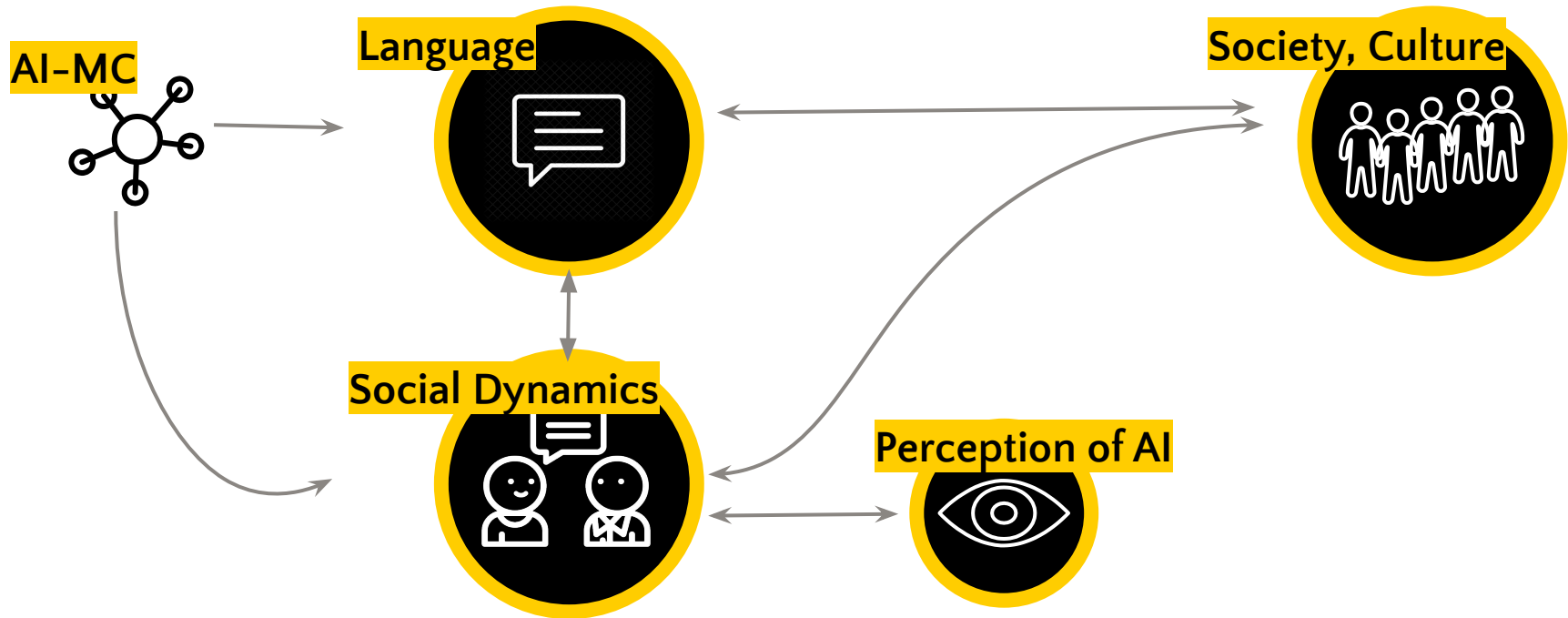
A dark, atmospheric landscape with silhouettes of structures and ships against a bright, hazy horizon. The scene is rendered in a low-poly, stylized manner. In the foreground, a dark, flat ground is marked with two white lines that recede into the distance. On the left, a large, dark, blocky structure stands on a slight rise. In the center, a tall, thin tower or antenna is visible. On the right, two large, dark, angular structures resembling ships or industrial buildings are silhouetted against the bright horizon. The sky is a gradient of dark blue to a bright, hazy yellow at the horizon, suggesting a sunrise or sunset. The overall mood is somber and mysterious.

**So, this is
upsetting...**





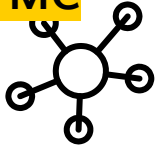
AI-MC impact: talk outline





AI-MC impact: talk outline

AI-MC



Society, Culture



***“I do not know with what
weapons World War III will be
fought, but World War IV will be
fought with sticks and stones.”***

“

-- attributed to A. Einstein



**Not just text: images, audio,
video; sync and async**





**Deception at our
(broadcast) fingertips**

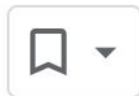




New adventures in communications

Home > Products > ML Kit > Guides

Smart Reply



With ML Kit's Smart Reply API, you can automatically generate relevant replies to messages. Smart Reply helps your users respond to messages quickly, and makes it easier to reply to messages on devices with limited input capabilities.

iOS

Android

Generate e paragraphs

Give Hyper a topic and it v for you to use.

sertation proposal:

[351741c159a5](#)

Will do!



New challenges

Write a polite email turning down a request to review a paper

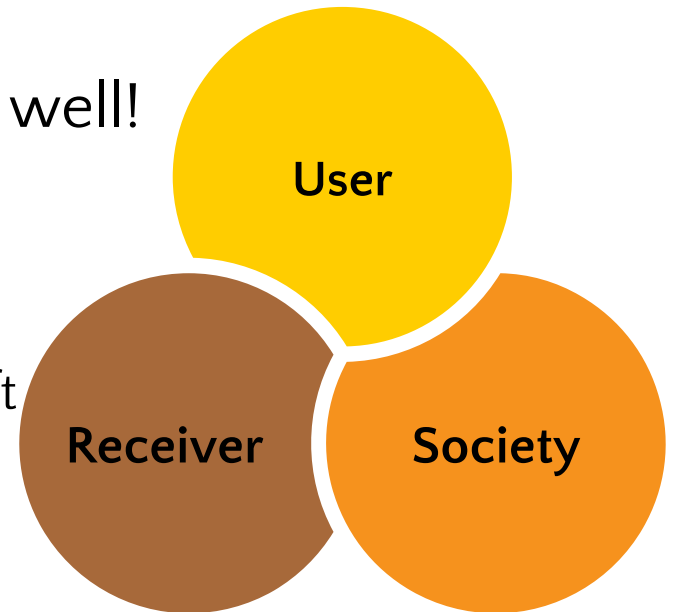
Dear Professor Smith,

Thank you for your email and kind words. I regret to say that to review your paper for the journal at this time. I am current with other projects and will not be able to give your paper the deserves. I hope you understand.



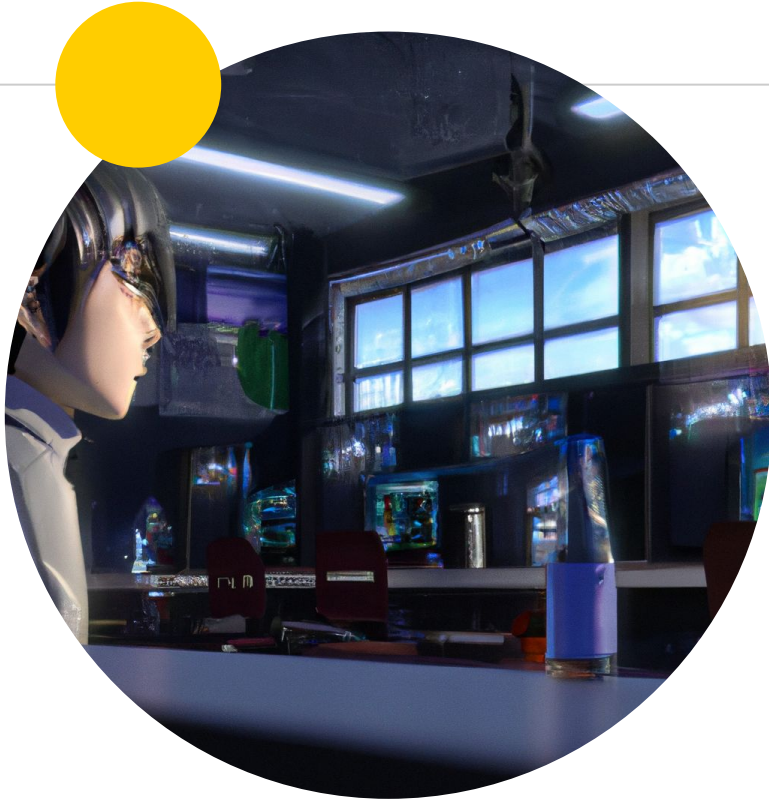
Human-centered AI-MC?

- User/system: more efficient
- But outer circles affected as well!
- Receiver(s):
 - Unreliable signals, trust
- Society:
 - Language adaptation and shift
 - Loss of trust
 - Shift of opinions, content
 - ...





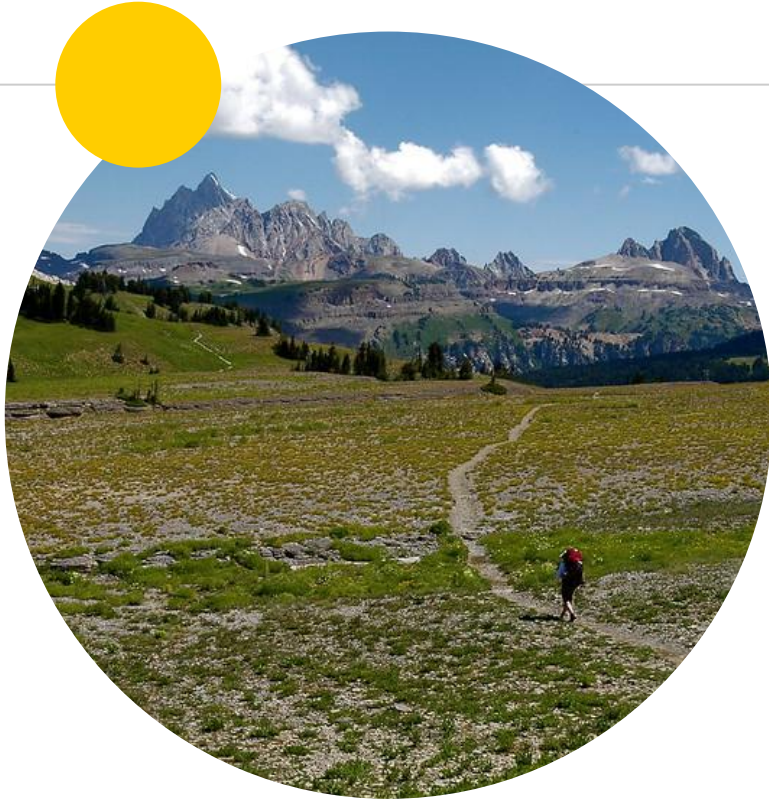
What can we do?



Design of AI-MC

- Transparency/disclosure
- Define and prevent harmful use
- Ensure equal access & use
- Create tools to signal effort, enable trust

Hancock, Naaman, Levy (2020). AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations.



Use and Abuse of Data

- Detection is (mostly) futile
- Openness of models, data
- Model safety
- Measuring and monitoring bias



MOAR Research

- Incredible research interest so far
- Sustain research effort
- How to collaborate with industry, inform real-world systems?



ge New York Times

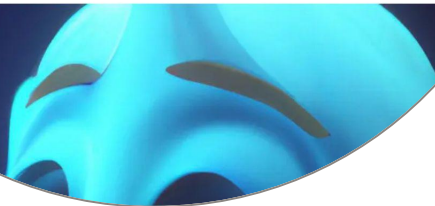
Need to Talk About How Good A.I. Getting

re in a golden age of progress in artificial intelligence. It's time
art taking its potential and risks seriously.

Share this article



608



Forward-looking Regulation

- Educating policymakers, public
- Writing our policy/ethical considerations

“To perfect... the range of machines without... giving humane direction to the organs of social action and social control is to create dangerous tensions in the structure of society.”

“

– Lewis Mumford 1934/1962
Technics and Civilization



Thanks!

mornaaman.com

mor.naaman@cornell.edu

Most papers available at bit.ly/aimc-papers