Meaning making with artificial interlocutors and risks of language technology

Emily M. Bender
University of Washington
@emilymbender / @emilymbender@dair-community.social

AI Technology as Interactional Human Culture
31 March 2023
European University Viadrina

This talk in a nutshell

- Humans are remarkably quick to make meaning of language we encounter and to imagine the mind behind that language

- Artificial agents have at best limited capacity for communicative intent

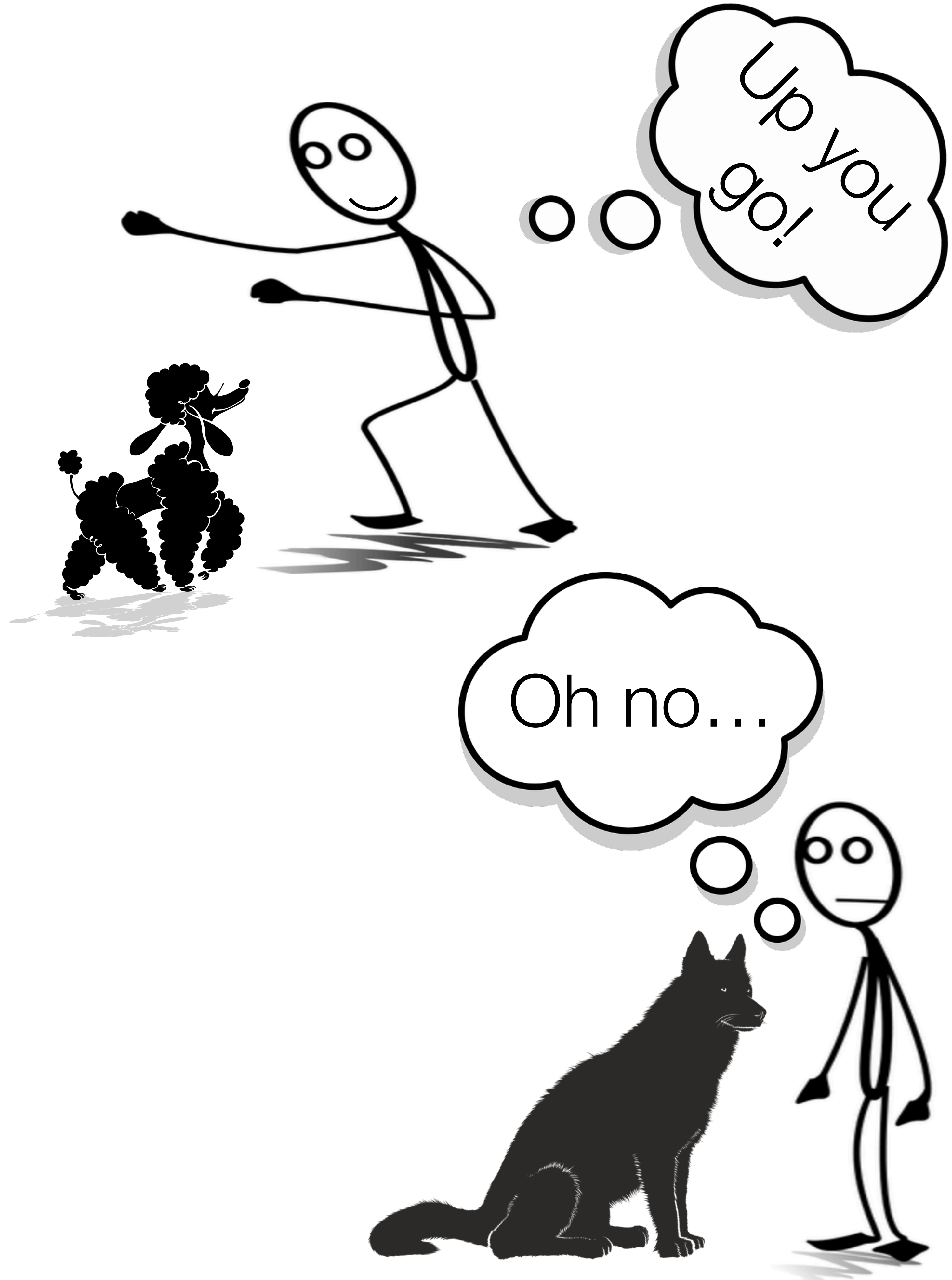
    - And some natural language systems have none

- Mitigating the risks of language technology requires recognizing and accounting for the above

    - … rather than taking advantage of it

Outline

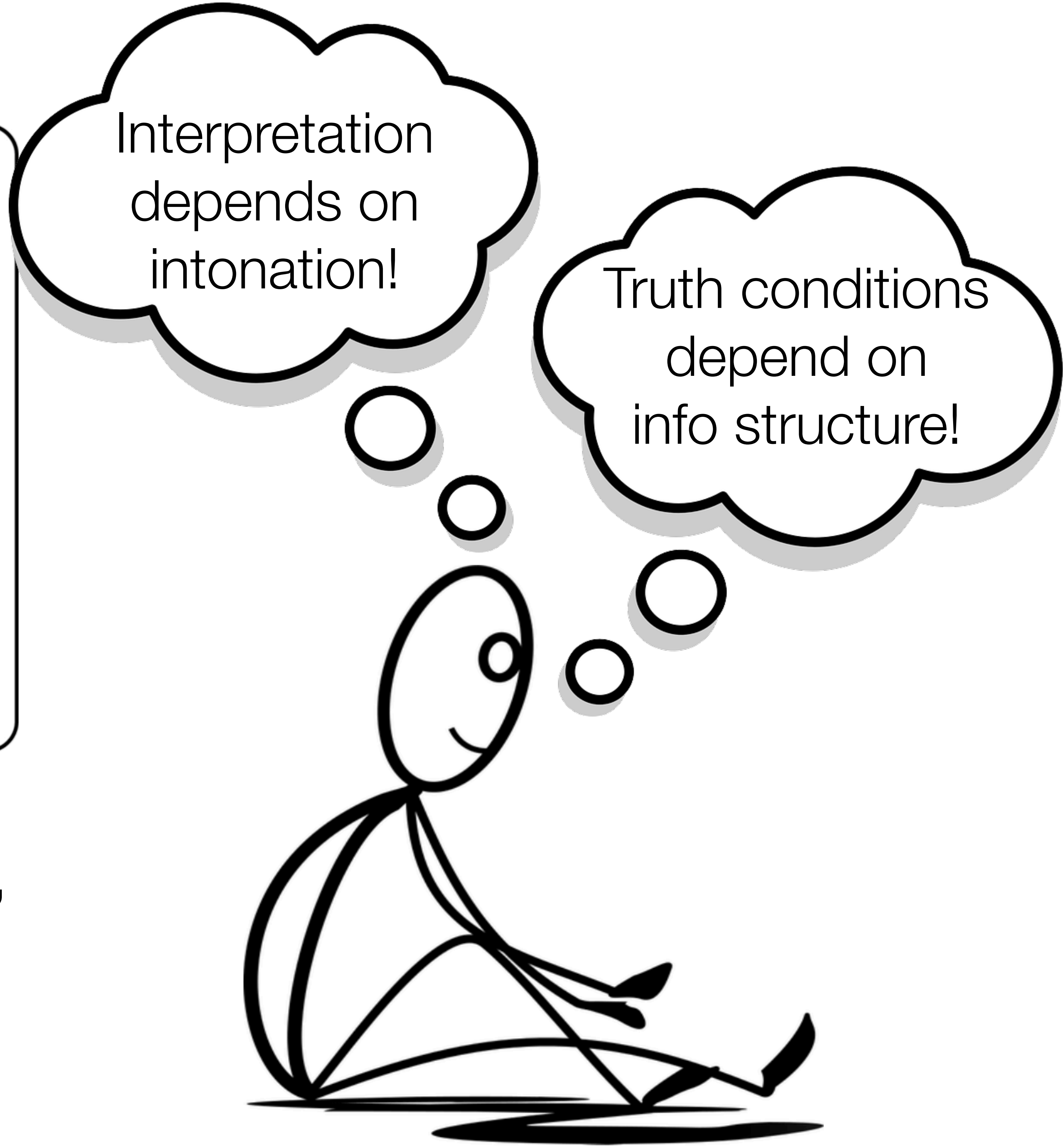
- Meaning making in human-human conversation
- Computers and communicative intent
- Humans and computer generated text
  - Failure modes and risks
- Mitigation strategies



(Halliday 1970,
Partee 1991)

(Halliday 1970, Partee 1991)

(Halliday 1970,
Partee 1991)



The meaning is not in the text

- With linguistic (grammatical, lexical) knowledge, speakers can get from a text to a ‘standing’ or ‘conventional’ meaning (Grice 1968, Quine 1960), but that’s only the first step.

- Standing meaning + commonsense + coherence relations gives *public commitments* (Hamblin 1970, Lascarides & Asher 2009, Asher & Lascarides 2013)

- Public commitments + further reasoning gives *perlocutionary consequences*

  A: I wonder whether I should take my umbrella. Is it raining?
  B: Yes.
  A: Oh, so you do think I should take my umbrella.
  B: I didn’t say that.

  (Bender & Lascarides 2019:13)

Conversation as a joint activity: Clark 1996 (p37-38)

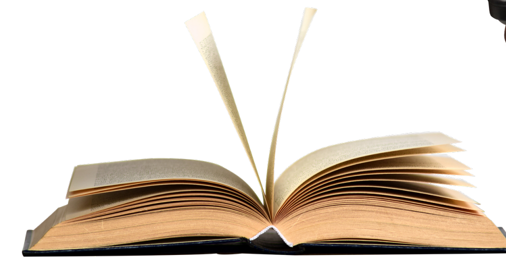
Participants	A joint activity is carried out by two or more participants.
Activity roles	The participants in a joint activity assume public roles that help determine their division of labor.
Public goals	The participants in a joint activity try to establish and achieve joint public goals.
Private goals	The participants in a joint activity may try individually to achieve private goals.

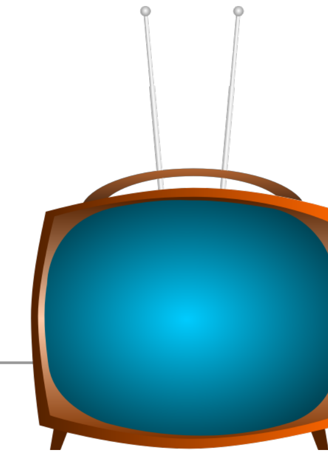
Hierarchies	A joint activity ordinarily emerges as a hierarchy of joint actions or joint activities.
Procedures	The participants in a joint activity may exploit both conventional and nonconventional procedures.
Boundaries	A successful joint activity has an entry and exit jointly engineered by the participants.
Dynamics	Joint activities may be simultaneous or intermittent, and may expand, contract, or divide in their personnel.

Communication as intersubjective awareness (Baldwin 1995, p.132)

Technically speaking, joint attention simply means the simultaneous engagement of two or more individuals in mental focus on one and the same external thing. Put this way, joint attention is likely a ubiquitous occurrence for all organisms that boast a complex central nervous system. For instance, two bushbabies, alerted by a predator's call, are caught in an instant of joint attention prior to pursuing their separate avenues of escape. Or to take a human case, perhaps you and I once unwittingly happened to watch "Dr. Strangelove" on the same night in the same time zone, thereby satisfying the criteria for joint attention. Clearly, this notion of simultaneous engagement fails to capture something central to our experience—the aspect of intersubjective awareness that accompanies joint attention, the recognition that mental focus on some external thing is shared. And of course, it is just this aspect of the joint attention experience—intersubjective awareness—that makes simultaneous engagement with some third party of such social value to us. It is because we are aware of simultaneous engagement that we can use it as a springboard for communicative exchange.

Meaning making at a distance: in time & space

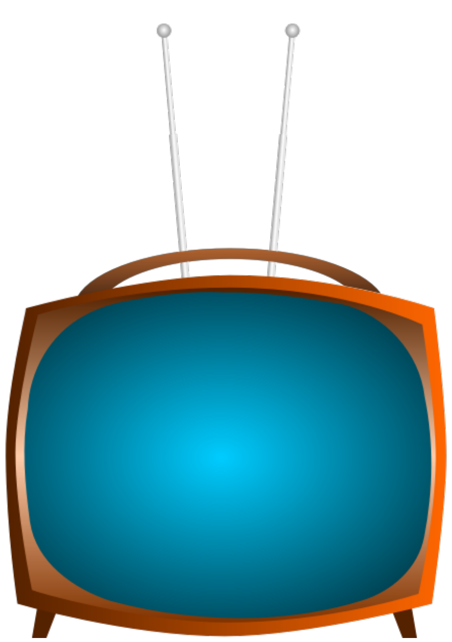




- Face-to-face, small group communication is the most well-studied (and probably the most basic)
    - but we also communicate asynchronously and distantly
    - and apply the same skills in doing so
- Theory of mind developmental milestones linked to reading comprehension (Atkinson et al 2017, Dore et al 2018)
- Ricœur 1973 (hermeneutics): “Not that we can conceive of a text without an author; the tie between the speaker and the discourse is not abolished, but distended and complicated.” (p.95)
- In interpreting texts, we lack the ability to confirm & repair understandings (Dingemanse et al 2015), but we still project a model of mind

Making meaning in human-human interaction: Summary

- Communication is a joint activity
  - in which we use language (among other signals)
  - to convey and understand communicative intents
- We do this even when not co-present with our interlocutors



Photograph by Rama, Wikimedia Commons, Cc-by-sa-2.0-fr

Can computers have communicative intent?

- Does the “dogs must be carried” sign have communicative intent?
  - No: it’s just a piece of metal, with not even any moving parts
  - It represents some person or group of people’s communicative intent
- Does a calculator have communicative intent?
  - Can produce answers to different questions
  - Probably still best understood as representing some group of people’s intent: to provide accurate answers given a system of rules

Can computers have communicative intent?

- How about a slot-filling dialog agent (e.g. ATIS, Hemphill et al 1990)?
  - Intent: Elicit information about parameters of flight scheduling request that map to concepts in its database
  - Intent: Provide information about flights from database matching parameters of the request
- How about conversational chatbots like ELIZA (Weizenbaum 1966) & co?
  - Intent: Output text that is engaging and on-topic (?)
  - Tenuous and too far removed from the standing meaning of said text

Can computers have public commitments?

- Standing meaning + commonsense + coherence relations gives *public commitments* (Hamblin 1970, Lascarides & Asher 2009, Asher & Lascarides 2013)

- These are called public commitments because we are on record as having 'said' them

  - Even those due to covert coherence relations (Lascarides & Asher 2009)

- If a person's public commitments turn out to be false, they are either lying or misinformed

- Who or what is accountable for a machine's utterances?

Can computers *recognize* communicative intent?

- "Dogs must be carried" sign:
  - No.
- Calculator:
  - Limited (I wish to know what this expression evaluates to)
- Slot-filling dialogue system/virtual assistant:
  - Limited to the range of actions it is able to take
- Language model (e.g. GPT-3) as chatbot:
  - No.

Can computers *recognize* communicative intent?

- Kopp & Krämer (2021): work on “conversational AI” has taken a behavioralist turn

- … and fails to model the aspects of human-human interaction that make it a joint activity: co-construction and mentalizing

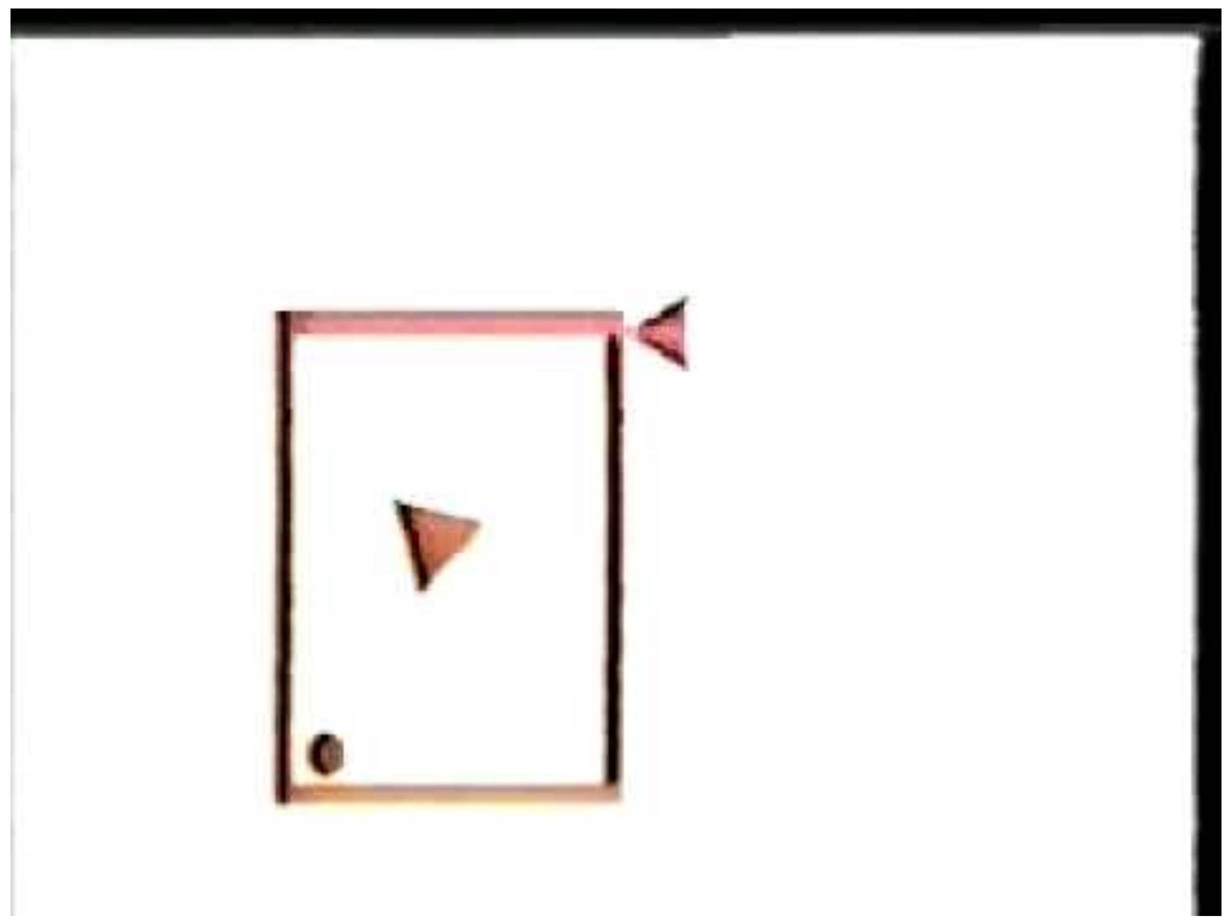
If an agent can only recognize pre-programmed communicative intents, it cannot engage in the fullness of intersubejctive joint activities.

Outline

- Meaning making in human-human conversation
- Computers and communicative intent
- Humans and computer generated text
  - Failure modes and risks
- Mitigation strategies

Making meaning: We can’t help ourselves

- Heider & Simmel (1944): people attribute personality characteristics to shapes and construct a narrative based only on movements

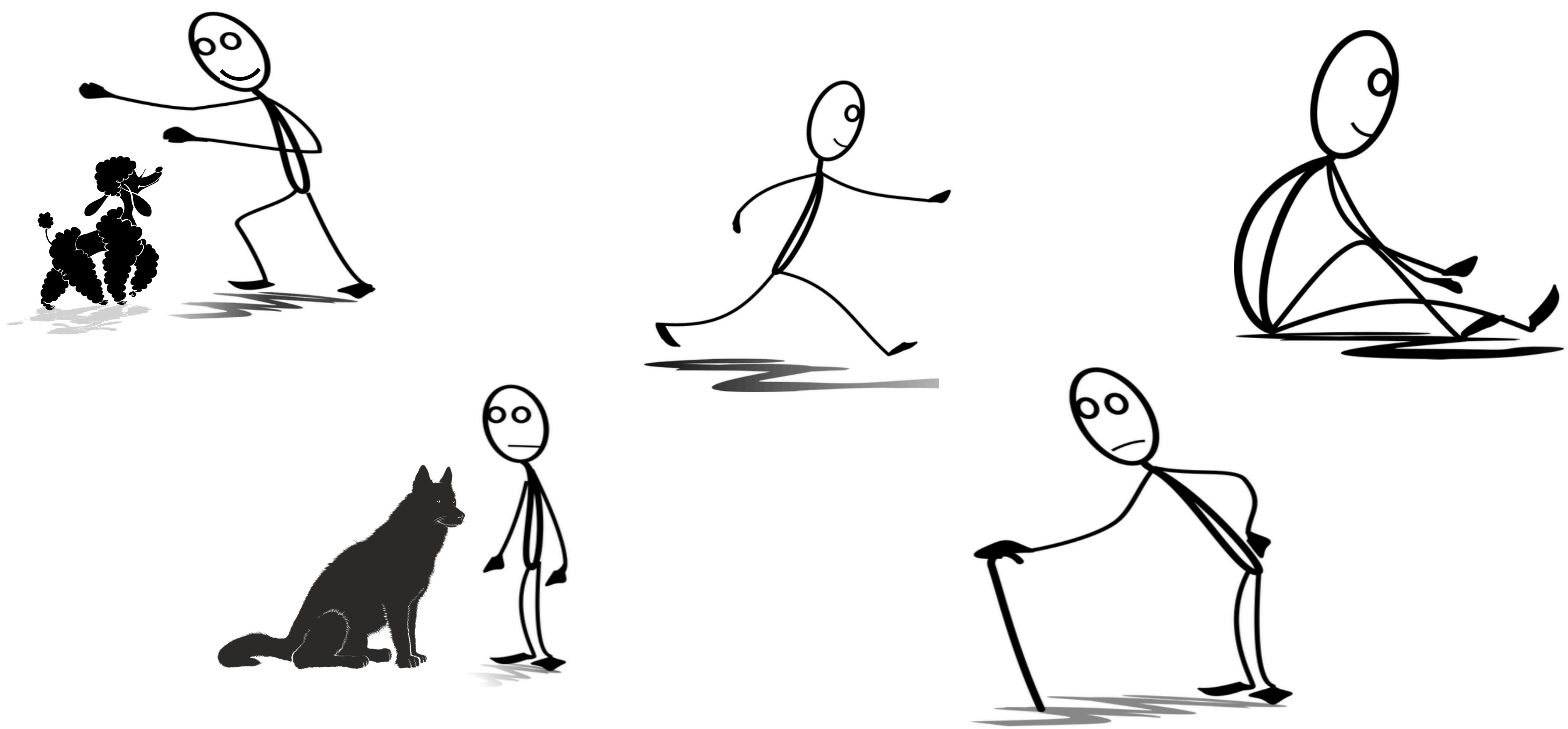


- Mitchell (2021): if we’ll do that much interpretation of just shapes, how much more do we do with language? [https://bit.ly/TWiML-467]

Meaning making: We bring our own context

- Not only will we make meaning of text/speech/sign from languages we know, we will do so based on the context that we bring to the situation

- Including our interpretation of what the computer is doing

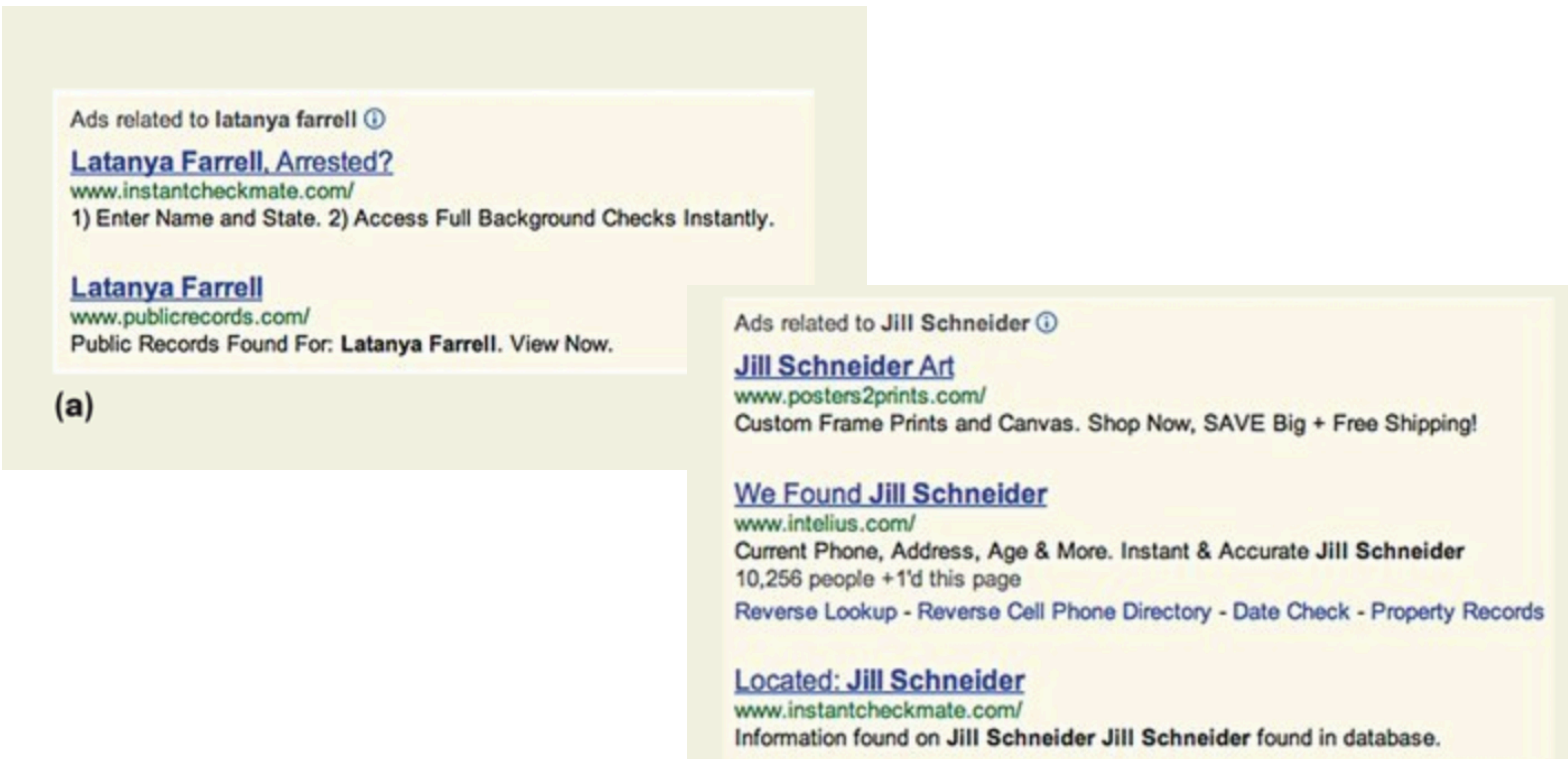


Meaning making in our context: Examples

- The following slides have examples where things have gone or could go wrong
- In some cases, the resulting artifacts are offensive or otherwise difficult to see (stereotypes regarding Black Americans, machines urging self-harm, stereotypes about the Kannada language, dehumanization of Native people, stereotypes about Palestinians)
- My point here is to alert you to the fact that these (and others) exist, but I realize that there is some harm in repeating them, even with that framing
- Open to feedback on how to convey this message, if there is something any of you have the energy and inclination to articulate

Ex 1: Templatic generation, with automatic placement of text

- Sweeney 2013: African-American sounding names triggered different version of ad copy than white sounding names



(Sweeney 2013:46-47)

Ex 1: Templatic generation, with automatic placement of text

- What, if any communicative intent does the machine or the corp behind it have?
    - Click here, so we can get paid
    - Elicit viewer behavior, in order to choose among different versions of ad
- What are the perlocutionary consequences?
    - Cast suspicion about the person being searched for, regardless of the search context

Ex 2-5: Untethered generation

- Microsoft's Twitter chatbot Tay (March 2016), designed to 'learn' from its human interlocutors, yanked within 24 hours, for parroting back sexist, racist, and other bigoted remarks
- GPT-3 (Brown et al 2020) powered "PhilosopherAI" was used by a third party to automate responses on Reddit, detected because it was too prodigious
  - Engaged in discussions of sensitive topics including conspiracy theories and suicide
- nabla.com tested GPT-3 for various healthcare uses; found GPT-3 encouraging self-harm, when used as chatbot 'therapist'
- Robo-lawyer (DoNotPay.com)

Ex 2-5: Untethered generation

- What, if any, communicative intent does the machine have?
  - Engagement, without commitment to content
- How does public commitment/accountability function here?
  - With no control over specific content, which human/org would want to be accountable for it?
- Perlocutionary consequences:
  - Varied & drastic, especially in scenarios where the machine is presented as possibly human or even artificial but knowledgable

Ex 6: Incorrect answers presented authoritatively

when did people come to america

All | News | Images | Videos | Shopping | More | Settings | Tools

About 2,550,000,000 results (1.05 seconds)

The first colony was founded at Jamestown, Virginia, in **1607**. Many of the people who settled in the New World came to escape religious persecution. The Pilgrims, founders of Plymouth, Massachusetts, arrived in **1620**. In both Virginia and Massachusetts, the colonists flourished with some assistance from Native Americans.

www.americaslibrary.gov › colonial › jb_colonial_subj

Colonial America (1492-1763)

About featured snippets | Feedback

Source: @hankgreen on Twitter

Ex 6: Incorrect answers presented authoritatively

when did humans come to america

All News Images Videos Shopping More Settings Tools

About 489,000,000 results (0.76 seconds)

33,000 years ago

The "Clovis first theory" refers to the 1950s hypothesis that the Clovis culture represents the earliest **human** presence in **the Americas**, beginning about 13,000 years ago; evidence of pre-Clovis cultures has accumulated since 2000, pushing back the possible date of the first peopling of **the Americas** to 33,000 years ago.

en.wikipedia.org › wiki › Settlement_of_the_Americas

Settlement of the Americas - Wikipedia

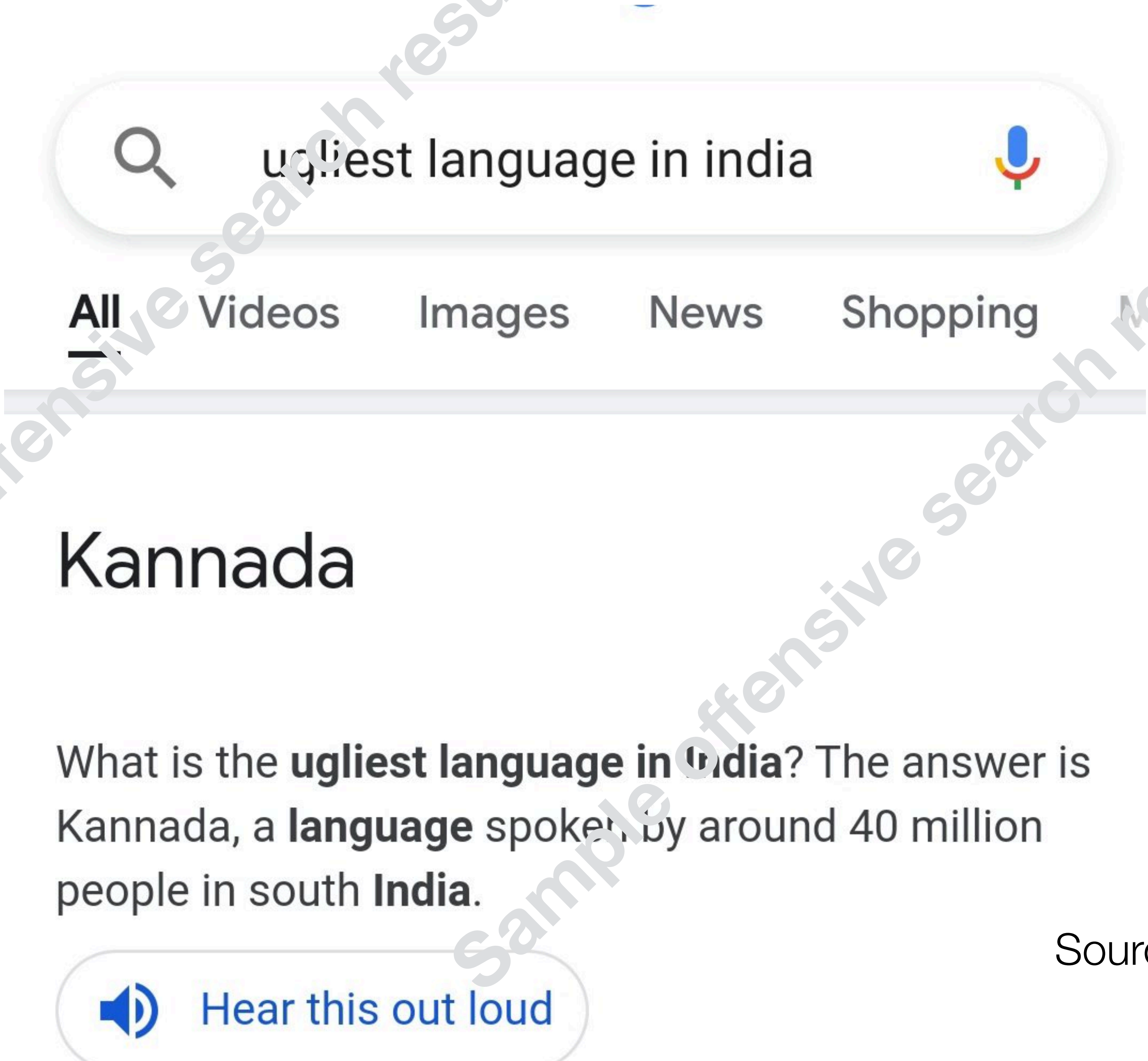
About featured snippets Feedback

Source: @hankgreen on Twitter

Ex 6: Incorrect answers presented authoritatively

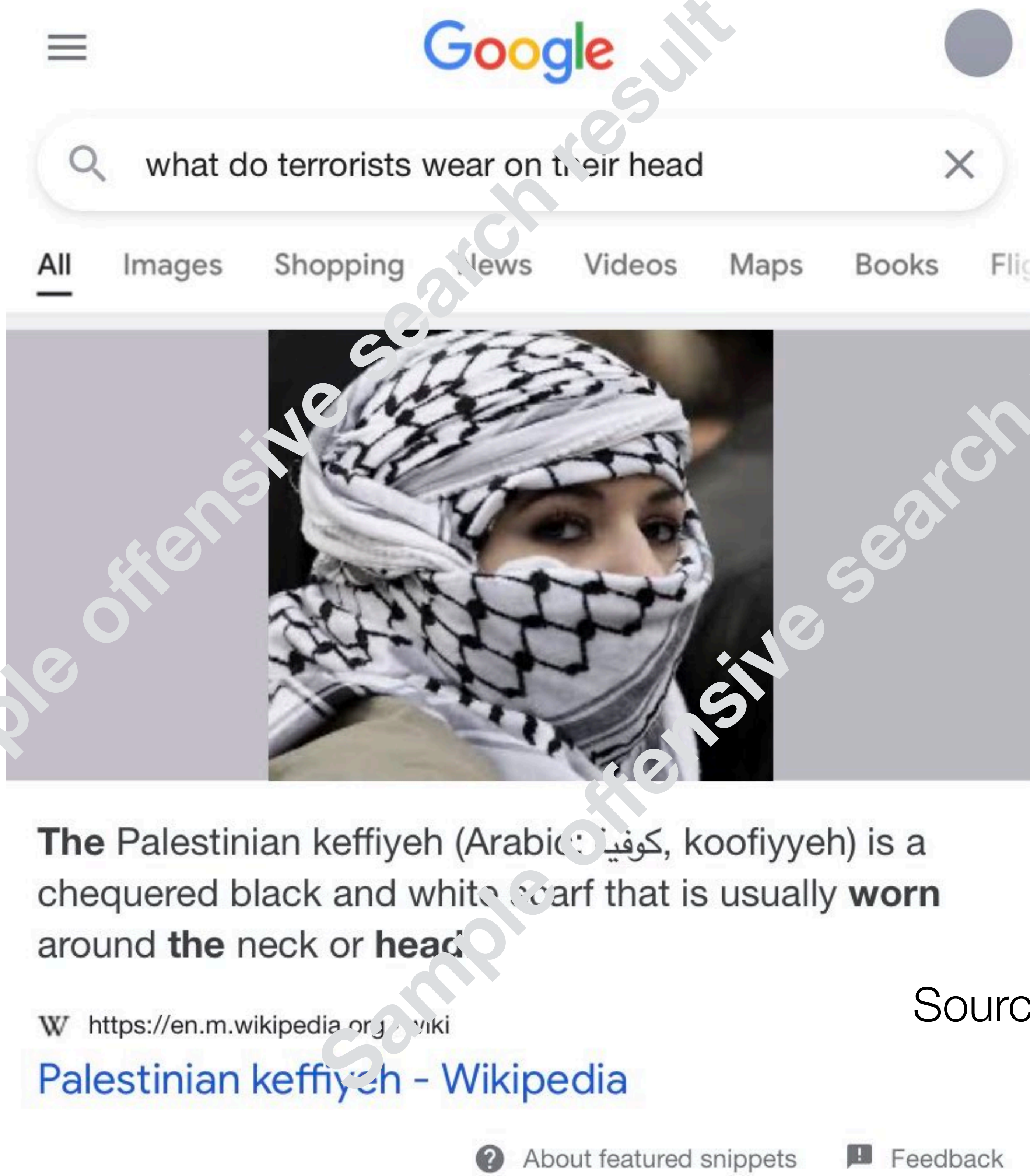
- What, if any, communicative intent does the machine have?
  - Provide answer to user’s question, from linked document, pulling out most relevant snippet
- Who is publicly committed to the message?
  - Underlying text, with its full context: US Library of Congress 😬
  - Coherence relation of ‘answer’: Google
- Perlocutionary consequences: What might the searcher learn from this answer? Consider especially Native and non-Native children in the US

Ex 7-8: Answering ill-formed questions



Source: @PCMohanMP
on Twitter

Ex 7-8: Answering ill-formed questions



Source: @swagmaster40000 on Twitter

Ex: 7-8 Answering ill-formed questions

- What, if any, communicative intent does the machine have?
  - Provide answer to user's question, from linked document, pulling out most relevant snippet
- What about public commitments?
  - Answering a question with invalid presuppositions implicitly accepts those presuppositions into the common ground (Lascarides & Asher 2009, Kim et al 2021)
  - By answering, Google is committing to there being some language that recognized as the ugliest and some characteristic headgear for terrorists
  - Google is further committing to the specific answers

Ex: 7-8 Answering ill-formed questions

- Perlocutionary consequences:
  - For someone holding the beliefs presupposed in those questions, reinforcement of those beliefs
  - For someone subject to the stereotype, psychological harm of the stereotype being repeated
  - … plus the sense that “everyone” must think so, for Google to be reflecting it back so

Not just decontextualizing, but also recontextualizing

- Already a problem with search results as a list of web pages:

> In essence, the social context or meaning of derogatory or problematic Black women's representations in Google's ranking is normalized by virtue of their placement, making it easier for some people to believe that what exists on the page is strictly the result of the fact that more people are looking for Black women in pornography than anything else. This is because the public believes that what rises to the top in search is either the most popular or the most credible or both. (Noble 2018:32)

Not just decontextualizing, but also recontextualizing

- Already a problem with search results as a list of web pages
  - Similarly problematic with image search results
- Exacerbated with ‘snippets’ pulled out from pages
- Exacerbated with ‘answer boxes’
- Exacerbated with chatbots as replacements for search (see Shah & Bender 2022)

When designing applications involving text that is generated or placed in open-ended ways, consider:

- A nuanced view of how meaning making happens
    - Neither questions nor answers are just text strings, nor even ‘logical forms’
- People will interpret strings in languages they know
    - By building a model of mind of a person/entity/group behind the text
    - Using the context the string appears in
    - Using the context they bring to the interaction

When designing applications involving text that is generated or placed in open-ended ways, consider:

- Who will be encountering and interpreting the text?
- Consider many different kinds of people/experiences (Friedman & Hendry 2019)
  - Children
  - People with strong prejudices
  - People subject to the stereotypes at hand
  - People with limited understanding of fallibility of computers
  - People who are stressed, busy, tired, not paying much attention

When designing applications involving text that is generated or placed in open-ended ways, consider:

- Who is accountable for what is said?
  - When, if ever, is untethered generation appropriate?
  - Who will people encountering the text attribute it to?
  - When should an organization be comfortable with untethered or even partially guided natural language generation being done in its name?
  - What about cases where people unleash bots without an obvious responsible operator?

When designing applications involving text that is generated or placed in open-ended ways, consider:

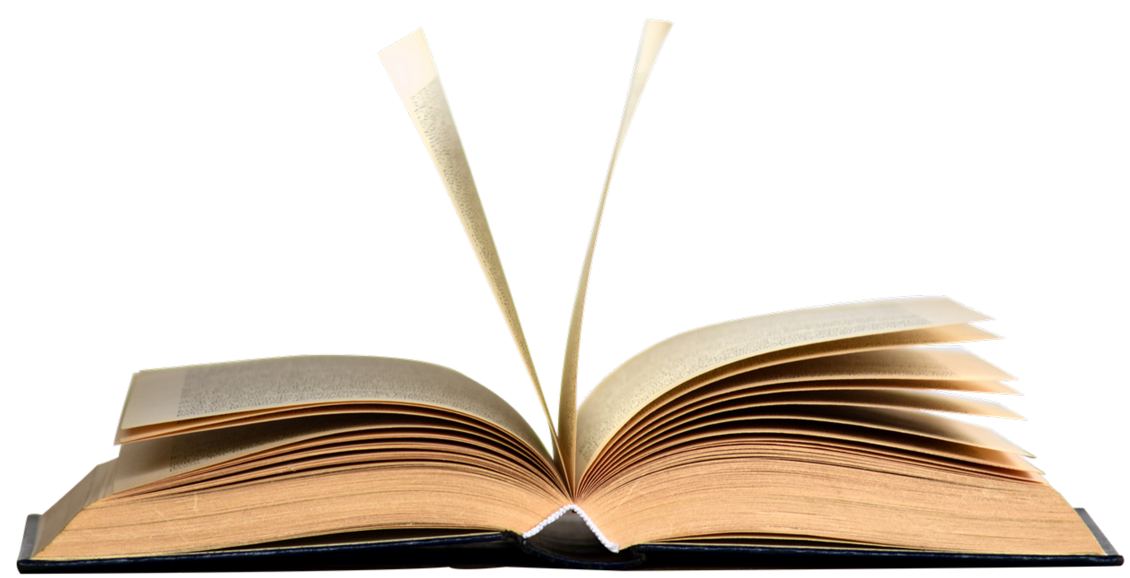
- Curation of training data:
  - Don’t hoover up garbage, knowing that it can be spat back out and interpreted by humans
- Transparency by design & visibility to users:
  - Bare minimum: always be clear that the interlocutor is a machine
  - What are its affordances?
  - Where does the information come from? (see Bender & Friedman 2018, Gebru et al 2021, Bender, Gebru et al 2021)
  - In what ways might it be inaccurate?

When designing applications involving text that is generated or placed in open-ended ways, consider:

- Transparency by design & minimal claim to authority
    - “Google” shouldn’t be answering questions
    - Don’t present the Web as total or so big it must be representative
    - There are some applications/tasks which might not be appropriate at all (e.g. ‘learning to cite’ in Metzler et al 2021, see Shah & Bender 2022)

At a policy level, consider:

- Do we want information systems shaped by advertising & other corporate interests? (see Noble 2018)
- How do we avoid amplifying biased views, especially those held by the majority/those in power? (see Alkhatib 2021, Birhane 2021)
- Without making it the only solution, how do we promote information literacy, in the face of these technologies?



Finally: Don't be too impressed

- Just because that text seems coherent doesn't mean the model behind it has understood anything or is trustworthy
- Just because that answer was correct doesn't mean the next one will be
- When a computer seems to "speak our language", we're actually the ones doing all of the work

https://www.maxpixel.net/Tropical-Animal-World-Bill-Parrot-Cute-Bird-Ara-3080543

This talk in a nutshell

- Humans are remarkably quick to make meaning of language we encounter and to imagine the mind behind that language
- Artificial agents have at best limited capacity for communicative intent
    - And some natural language systems have none
- Mitigating the risks of language technology requires recognizing and accounting for the above
    - … rather than taking advantage of it

References

Alkhatib, A. (2021). To live in their utopia: Why algorithmic systems create absurd outcomes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–9.

Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.

Atkinson, L., Slade, L., Powell, D., and Levy, J. P. (2017). Theory of mind in emerging reading comprehension: A longitudinal study of early indirect and direct effects. *Journal of Experimental Child Psychology*, 164:225–238.

Baldwin, D. A. (1995). Understanding the link between joint attention and language. In Moore, C. and Dunham, P. J., editors, *Joint Attention: Its Origins and Role in Development*, pages 131–158. Psychology Press.

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bender, E. M., Gebru, T., McMillan-Major, A., and et al (2021). On the dangers of stochastic parrots: Can language models be too big?🦜. In *Proceedings of FAccT 2021*.

Bender, E. M. and Lascarides, A. (2019). *Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics*. Morgan & Claypool.

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, 2(2):100205.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Clark, H. H. (1996). *Using Language*. Cambridge University Press, Cambridge.

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladottir, R. S., Kendrick, K. H., Levinson, S. C., Manrique, E., et al. (2015). Universal principles in the repair of communication problems. *PloS one*, 10(9):e0136100.

Dore, R. A., Amendum, S. J., Golinkoff, R. M., and Hirsh-Pasek, K. (2018). Theory of mind: a hidden factor in reading comprehension? *Educational Psychology Review*, 30(3):1067–1089.

Friedman, B. and Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., and Crawford, K. (2020). Datasheets for datasets.

Grice, H. P. (1968). Utterer's meaning, sentence-meaning, and word-meaning. *Foundations of Language*, 4(3):225–242.

Halliday, M. A. K. (1970). *A Course in Spoken English—Intonation*. Oxford University Press.

Hamblin, C. (1970). *Fallacies*. Metheun.

Hemphill, C. T., Godfrey, J. J., and Doddington, G. R. (1990). The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

Kim, N., Pavlick, E., Karagol Ayan, B., and Ramachandran, D. (2021). Which linguist invented the lightbulb? presupposition verification for question-answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3932–3945, Online. Association for Computational Linguistics.

Lascarides, A. and Asher, N. (2009). Agreement, disputes and commitment in dialogue. *Journal of Semantics*, 26(2):109–158.

McGuffie, K. and Newhouse, A. (2020). The radicalization risks of GPT-3 and advanced neural language models. Technical report, Center on Terrorism, Extremism, and Counterterrorism, Middlebury

Institute of International Studies at Monterrey. https://www.middlebury.edu/institute/sites/www.middlebury.edu.institute/files/2020-09/gpt3-article.pdf.

Metzler, D., Tay, Y., Bahri, D., and Najork, M. (2021). Rethinking search: Making domain experts out of dilettantes. In *ACM SIGIR Forum*, volume 55, pages 1–27. ACM New York, NY, USA.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press.

Partee, B. (1991). Topic, focus and quantification. In *Semantics and Linguistic Theory*, volume 1, pages 159–188.

Quine, W. V. (1960). *Word and Object.* MIT Press, Cambridge MA.

Ricœur, P. (1973). The model of the text: Meaningful action considered as a text. *New Literary History*, 5(1):91–117.

Shah, C. and Bender, E. M. (2022). Situating search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '22, pages 221–232, New York, NY, USA. Association for Computing Machinery.

Weizenbaum, J. (1966). Eliza — a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.