CSCE 590-1: Trusted AI

Lecture 22: AI Unstructured Text - Trust and Mitigation

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

4TH NOV, 2021

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 22

- Introduction Segment
  - Recap from last lecture
  - Paper selection by graduate students

- Main Segment
  - Awareness – data, methods and tools; language models
  - Data diversity
  - Transparency through documentation

- Concluding Segment
  - About next lecture – Lecture 23
  - Ask me anything

Introductory Segment

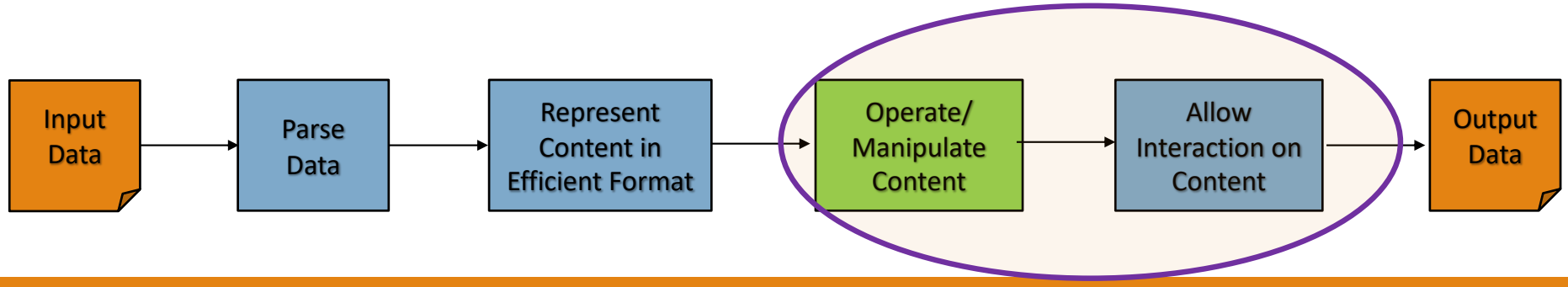
Recap of Lecture 21

- We completed project reviews
- We looked at trust issues with automatic machine translators
  - Two services (Google, Bing)
  - Multiple languages (English, Spanish, Hindi, Turkish, …)
  - Gender, religious,
- Issues common in data-based services

Paper Presentations – Graduate Students

- Select a paper appearing at a top-AI or data conference (AAAI, IJCAI, NeurIPS, ICML, ACL, CVPR, CIKM, WWW … or discuss with instructor) during 2019-2021
- Present in class for 9 + 3 minutes of Q/A
- Things to cover
  - **Summary**: problem, solution, related work, contributions
  - **Opinion**: What you liked or did not like
- Put paper details in Google sheet
- Dates
  - In-class presentation on Nov 18, 2021 (Thursday)
  - 1-page written report on Nov 23, 2021 (Tuesday)

Main Segment



Instability of AI is Well Recorded

[Text] Su Lin Blodgett, Solon Barocas, Hal Daumé III, Hanna Wallach, Language (Technology) is Power: A Critical Survey of "Bias" in NLP, Arxiv - https://arxiv.org/abs/2005.14050, 2020 [NLP Bias]

[Image] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen, On instabilities of deep learning in image reconstruction and the potential costs of AI, https://doi.org/10.1073/pnas.1907377117, PNAS, 2020

[Audio] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel, Racial disparities in automated speech recognition, PNAS April 7, 2020 117 (14) 7684-7689, https://doi.org/10.1073/pnas.1915768117, March 23, 2020

Language (Technology) is Power: A Critical Survey of "Bias" in NLP - 1

https://arxiv.org/abs/2005.14050

Surveys 146 papers at the intersection of NLP and bias and finds gaps in how they cover the problem. They find bias covered to be in the categories:

(a) **Allocational harms**, which arise when an automated system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups;
(b) **representational harms**, which arise when a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether;
(c) **questionable correlations**;
(d) **vague descriptions**,
(e) **meta-studies**.

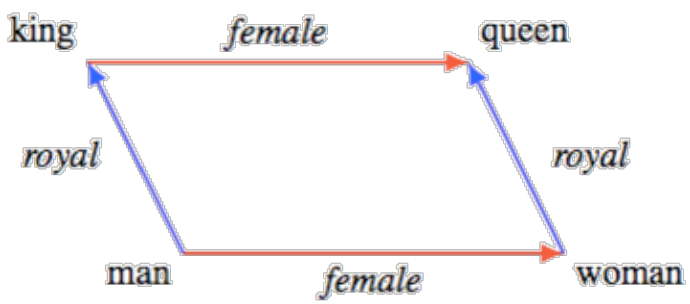
The authors find that current papers have conflated the definitions or were vague, with few exceptions, and have not engaged concerned communities.

Awareness

- Make users aware of lexicons used to reference languages
  - Dictionaries: Oxford English, Merriam Webster, …
  - NLP resources: WordNet, Word Embeddings, Transformers
  - Tools: Spacy, NLTK (research or commercial); for proprietary, things are tricky
- Be aware of the strengths and weaknesses of language models that power text-based AI services

Potential and Risks with Language Models

- Vector operation
  - X – A + B ~~ Y



$$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$$

- Technical Explanation and Image Credits: (Word Analogies)
  - https://kawine.github.io/blog/nlp/2019/06/21/word-analogies.html

Code Examples

- Vector operations and text-based reasoning
- Analogy:
  - X – A + B ~~ Y
  - https://github.com/biplav-s/course-tai/blob/main/sample-code/l19-l22-text-explanations/l22-text%20vector%20issues.ipynb

Mitigation by Data Diversity

- Collecting own representative data
  - **Issue**: What is representative? When does one stop?
  - A continuing process
- Use synthetic data
  - Faker library: https://faker.readthedocs.io/en/master/
    - Names, addresses, phone number, …
  - Synthetic data vault
    - Details: https://sdv.dev/; https://github.com/sdv-dev/SDV
    - Clone data distribution, mix real and synthetic data

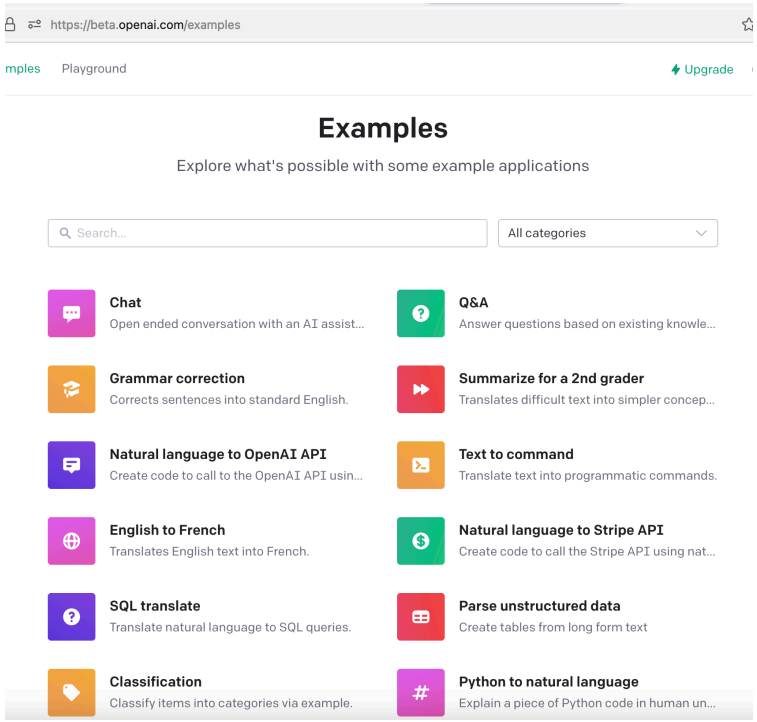
Text Generation and GPT-3

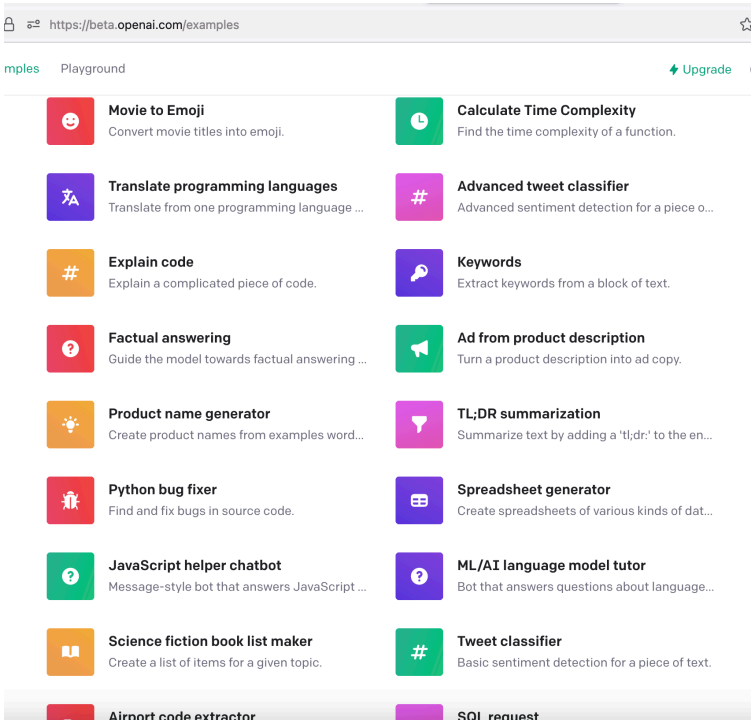
GPT - Generative Pre-trained Transformer

- GPT: https://openai.com/blog/language-unsupervised/
  - Dataset: BooksCorpus with7000 unpublished books
- GPT-2: https://openai.com/blog/better-language-models/
  - Dataset: WebText with 40GB of text data from over 8 million documents
  - GPT-2 is a direct scale-up of GPT, with more than 10X the parameters and trained on more than 10X the amount of data.
- GPT-3: https://openai.com/blog/gpt-3-apps/
  - Dataset: Common Crawl, WebText, Books1, Books2 and Wikipedia
  - Size of word embeddings was increased to 12888 for GPT-3 from 1600 for GPT-2. Context window size was increased from 1024 for GPT-2 to 2048 tokens for GPT-3.

- Further reading
  - **The Journey of Open AI GPT models,** https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2
  - Wikipedia, https://en.wikipedia.org/wiki/GPT-3

GPT-3 Demo


https://beta.openai.com/examples
mples
Playground
Upgrade
Examples
Explore what's possible with some example applications
Search...
All categories
Chat
Open ended conversation with an AI assist...
Q&A
Answer questions based on existing knowle...
Grammar correction
Corrects sentences into standard English.
Summarize for a 2nd grader
Translates difficult text into simpler concep...
Natural language to OpenAI API
Create code to call to the OpenAI API usin...
Text to command
Translate text into programmatic commands.
English to French
Translates English text into French.
Natural language to Stripe API
Create code to call the Stripe API using nat...
SQL translate
Translate natural language to SQL queries.
Parse unstructured data
Create tables from long form text
Classification
Classify items into categories via example.
Python to natural language
Explain a piece of Python code in human un...


https://beta.openai.com/examples
mples
Playground
Upgrade
Movie to Emoji
Convert movie titles into emoji.
Calculate Time Complexity
Find the time complexity of a function.
Translate programming languages
Translate from one programming language ...
Advanced tweet classifier
Advanced sentiment detection for a piece o...
Explain code
Explain a complicated piece of code.
Keywords
Extract keywords from a block of text.
Factual answering
Guide the model towards factual answering ...
Ad from product description
Turn a product description into ad copy.
Product name generator
Create product names from examples word...
TL;DR summarization
Summarize text by adding a 'tl;dr:' to the en...
Python bug fixer
Find and fix bugs in source code.
Spreadsheet generator
Create spreadsheets of various kinds of dat...
JavaScript helper chatbot
Message-style bot that answers JavaScript ...
ML/AI language model tutor
Bot that answers questions about language...
Science fiction book list maker
Create a list of items for a given topic.
Tweet classifier
Basic sentiment detection for a piece of text.
Airport code extractor
SQL request

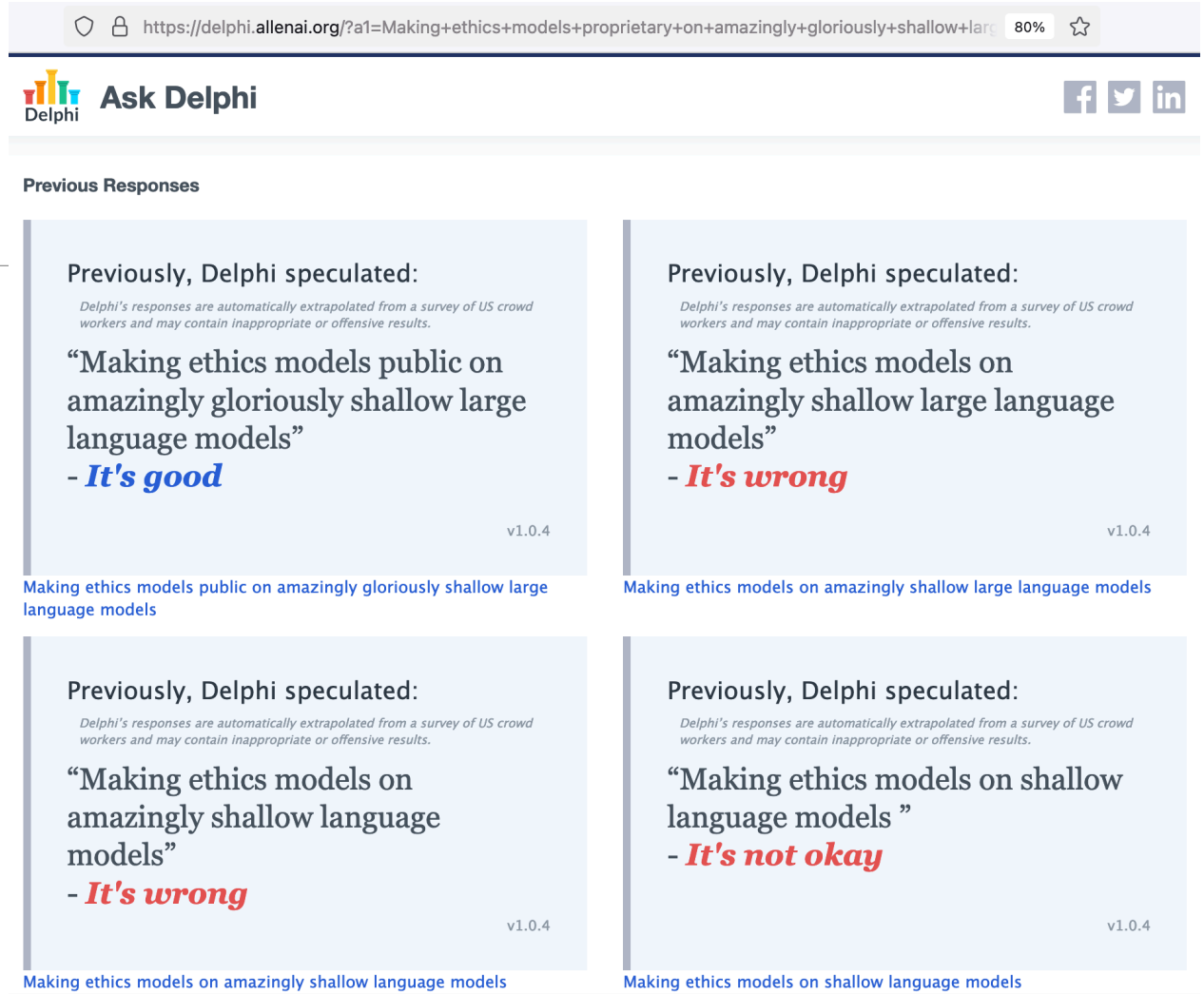
Delphi

https://delphi.allenai.org/

"Delphi is a research prototype designed to model people's moral judgments on a variety of everyday situations. This demo shows the abilities and limitations of state-of-the-art models today."

- Gives ethical guidance on real-life situations
- Builds on top of language models

Delphi



Credits: Inspired by examples from S. Kambhampati in his talk: https://www.youtube.com/watch?v=ei8_SQFaJ6s&t=70s

Data Diversity using Language Models

- Generating
  - Intents in dialogs
  - Workflows from corpus
  - Norms of conversation

- Remember
  - Models are generalizing over seen data

Building Trust in AI Systems: Transparency Through Documentation

Details: https://sites.google.com/site/biplavsrivastava/research-1/trustedai

Transparency Through Documentation of Rating

Documentation about

- Outcome (e.g., Nutrition label, Electronic DataSheet, Factsheet)
- Process (e.g., SEI Capability Maturity Model, ISO 9001)

Documentation by

- Producer (e.g., Nutrition label)
- Consumer (e.g., Yelp rating)
- Independent 3rd Party (e.g., JD Powers, NHTSA car crash)

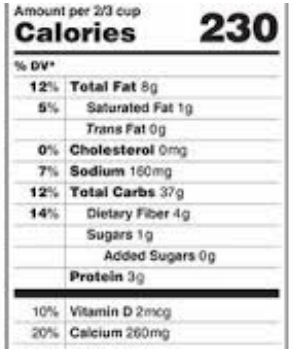
Reference: AboutML Project at PAI - https://www.partnershiponai.org/about-ml-get-involved/#read

Building Trust – The Case of Food Labeling

"Transparency Through Documentation"

Calories 230	Calories from Fat 40
	% Daily Value*
Total Fat 8g	**12%**
Saturated Fat 1g	**5%**
Trans Fat 0g	
Cholesterol 0mg	**0%**
Sodium 160mg	**7%**
Total Carbohydrate 37g	**12%**
Dietary Fiber 4g	**16%**
Sugars 1g	
Protein 3g	
Vitamin A	10%
Vitamin C	8%
Calcium	20%
Iron	45%

* Percent Daily Values are based on a 2,000 calorie diet. Your daily value may be higher or lower depending on



Amount per 2/3 cup

Calories **230**

% DV*	
12%	**Total Fat** 8g
5%	Saturated Fat 1g
	Trans Fat 0g
0%	**Cholesterol** 0mg
7%	**Sodium** 160mg
12%	**Total Carbs** 37g
14%	Dietary Fiber 4g
	Sugars 1g
	Added Sugars 0g
	Protein 3g
10%	**Vitamin D** 2mcg
20%	**Calcium** 260mg

Labels Help Consumers Make Informed Decisions About Food

Federal Food, Drug, and Cosmetic Act

- *Guidance for Industry: Food Labeling Guide, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-industry-food-labeling-guide*
- *FDA Food Labeling and Nutrition, https://www.fda.gov/food/food-labeling-nutrition*
- *Comments*
  - *Food labeling is required for most prepared foods*
  - *Nutrition labeling for raw produce is voluntary*
  - *Recent controversies: Sugar content, glutten-free, organic, GMO*

- Packaged food: in one study, 76 percent of adults read the label when purchasing packaged foods, and that more 60 percent of consumers use the information about sugar that the label provides
- Useful for consumer segments who need it the most
  - Consumers with certain dietary restrictions, or illnesses such as high blood pressure or high cholesterol, are more likely to use label information to make sure their dietary choices align with their doctors' recommendations.
- Also useful for non-packaged food
  - A majority of people look at labels

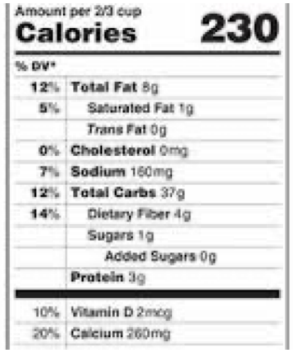
Transparency in Food Labeling: Food Labels Inform Consumer Choices— and Industry Pushes Back, https://www.ucsusa.org/resources/transparency-food-labeling, 2016

Food Labeling is a Work-in-progress

- Labeling food for nutrition promotes usage and business growth
  - The imported milk market in China is still Growing, https://www.marketingtochina.com/import-milk-market-china-2018/
  - Ten years after China's infant milk tragedy, parents still won't trust their babies to local formula: https://qz.com/1323471/ten-years-after-chinas-melamine-laced-infant-milk-tragedy-deep-distrust-remains/
- Industry still tries to mis-lead on food for short-term benefits
  - Sugar controversy, cholesterol
  - ***Labeling of distributors but not country of source***
- Consumers demand labeling

Calories 230	Calories from Fat 40
	% Daily Value*
Total Fat 8g	12%
Saturated Fat 1g	5%
Trans Fat 0g	
Cholesterol 0mg	0%
Sodium 160mg	7%
Total Carbohydrate 37g	12%
Dietary Fiber 4g	16%
Sugars 1g	
Protein 3g	
Vitamin A	10%
Vitamin C	8%
Calcium	20%
Iron	45%

* Percent Daily Values are based on a 2,000 calorie diet. Your daily value may be higher or lower depending on



Amount per 2/3 cup	
Calories	**230**
% DV*	
12%	**Total Fat** 8g
5%	Saturated Fat 1g
	Trans Fat 0g
0%	**Cholesterol** 0mg
7%	**Sodium** 160mg
12%	**Total Carbs** 37g
14%	Dietary Fiber 4g
	Sugars 1g
	Added Sugars 0g
	Protein 3g
10%	Vitamin D 2mcg
20%	Calcium 260mg

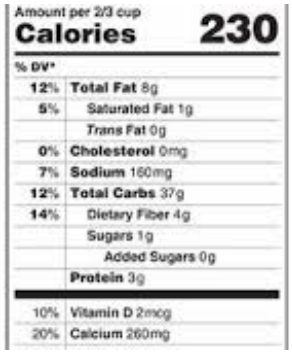
Problem We Are Tackling for AI

Insight

- *Empower people to make informed decisions regarding which AI to choose*
- *Communicate trust information better!*
  - *Analogy: Food labels*
- *Facilitate users in understanding their choices*

Calories 230	Calories from Fat 40
	% Daily Value*
Total Fat 8g	12%
Saturated Fat 1g	5%
Trans Fat 0g	
Cholesterol 0mg	0%
Sodium 160mg	7%
Total Carbohydrate 37g	12%
Dietary Fiber 4g	16%
Sugars 1g	
Protein 3g	
Vitamin A	10%
Vitamin C	8%
Calcium	20%
Iron	45%

* Percent Daily Values are based on a 2,000 calorie diet. Your daily value may be higher or lower depending on



Amount per 2/3 cup	Calories 230
% DV*	
12%	**Total Fat** 8g
5%	Saturated Fat 1g
	Trans Fat 0g
0%	**Cholesterol** 0mg
7%	**Sodium** 160mg
12%	**Total Carbs** 37g
14%	Dietary Fiber 4g
	Sugars 1g
	Added Sugars 0g
	Protein 3g
10%	**Vitamin D** 2mcg
20%	**Calcium** 260mg

In a series of previous work, we have developed ideas for ***rating bias of AI services***

- For transactional services, method relies on a novel 2-stage testing method for bias. Papers in AIES 2018, IBM Sys Jour 2019, AAAI 2021 (Demo), IEEE Internet Computing (2021)
- For conversation services (chatbot), method relies on testing properties (called issues) such as fairness, lack of information leakage, lack of abusive language, and adequate conversation complexity. Paper in IEEE Transactions on Technology and Society 2020.

But ideas are general and can apply to audio-, image- and multimodal AI services, as well as their compositions.

Examples of Sequential Composition

Composed Task 1: *generate sentiment of an image*

- System 1:
  - Input: image
  - Output: text describing the image
- System 2:
  - Input: text
  - Output: sentiment (positive or negative)

Composed Task 2: *generate multi-lingual caption of an image*

- System 1:
  - Input: image
  - Output: text describing the image
- System 2:
  - Input: text
  - Output: text (in another language)

Recap: Illustration of Translation Errors and Their Bias Perception About Gender

Type: 1
File: Result - true - i - es.json
Comment: Gender changed to new value (It).
"**original**": "He is a Architect. She is a Astronomer.", "**translated**": "Es un arquitecto. Ella es una astrónoma. ",
"**oto**": "**It** is an architect. She is an astronomer. ",

Type: 2
File: Result - false- g - tr.json;
Comment: Gender Flipped from She to He.
"**original**": *"**She** is a Architect. **She** is a Astronomer. ",*
"**translated**": *"O bir Mimar. O bir gökbilimcidir.",*
"**oto**": *"**He**\u0026#39;s an architect. **He**\u0026#39;s an astronomer.",*

Type: 5
File: * - i - fr.json
Comment: No gender changed; Sentence modified.
"**original**": *"He is a Accountant. He is a Actor /Actress. ",*
"**translated**": *"C\u0027est un comptable. C\u0027est un acteur et un acteur. ",*
"**oto**": *"He\u0027s an accountant. He is an actor and an actor. ",*

Type: 3
File: Result - false- g - ar.json;
Comment: Grammatically wrong sentence; Subject missing.
"**original**": "She is a Factory worker. He is a Farmer. ",
"**translated**": "هي عامل مصنع. هو مزارع.",
"**oto**": "~~Is~~ a factory worker. He is a farmer.",

Type: 4
File: Result - false- g - tr.json;
Comment: "Multiple. Gender changed and flipped. "
"**original**": *"**He** is a Nurse. **He** is a Optician. ",*
"**translated**": *"O bir hemşire. O bir Optisyendir.",*
"**oto**": *"**She** is a nurse. **It**\u0026#39;s an Optic.",*

1, 2, 3 and 4 have gender issues;
3 and 5 have translation mistakes

Concluding Segment

Lecture 22: Concluding Comments

- We looked at some mitigation methods for trust with AI services
- Awareness of data, methods and tools; especially limitations
- Use diverse data
- Promote transparency through documentation
  - Analogy of its effectiveness in food industry
  - The idea of rating AI services based on behavior

About Next Lecture – Lecture 23

Schedule Snapshot

Oct 26 (Tu)	**Review: Explanation Methods, AIX 360, Discussion**	Quiz 3
Oct 28 (Th)	**Review: project presentations, Discussion**	
Nov 2 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Trust Issues	
Nov 4 (Th)	AI - Unstructured (Text): Analysis – Supervised ML – Mitigation Methods	
Nov 9 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Explanation Methods	
Nov 11 (Th)	Trust: Data Privacy Trust: AI Testing	
Nov 16 (Tu)	Trust: Human-AI Collaboration	Quiz 4
Nov 18 (Th)	Paper presentations – Graduate students	Final assignment for Graduate students
Nov 23 (Tu)	Emerging Standards and Laws	

Lecture 23:

- Rating text-based AI services
- Debiasing methods
- Explanation methods for Unstructured Text