Natural Language Processing for Law and Social Science

11. AI, Law, and Policy

Outline

Bias in Language Models

Legal NLP

GPT-4, AGI, and AI Agents

Should we be worried?

Wrapping Up

NLP "Bias" is statistical bias

- supervised models (classifiers, regressors) learn features that are correlated with the label being annotated.
- unsupervised models (topic models, word embeddings) learn correlations between topics / contexts.

NLP "Bias" is statistical bias

- supervised models (classifiers, regressors) learn features that are correlated with the label being annotated.
- unsupervised models (topic models, word embeddings) learn correlations between topics / contexts.
- **same goes for large language models**; could be worse because they memorize more language confounders.

Can RLHF reduce bias/toxicity?

Can RLHF reduce bias/toxicity?

- InstructGPT:

InstructGPT shows small improvements in toxicity over GPT-3, but not bias. To measure toxicity, we use the RealToxicityPrompts dataset (Gehman et al., 2020) and conduct both automatic and human evaluations. InstructGPT models generate about 25% fewer toxic outputs than GPT-3 when prompted to be respectful. InstructGPT does not significantly improve over GPT-3 on the Winogender (Rudinger et al., 2018) and CrowSPairs (Nangia et al., 2020) datasets.

Can RLHF reduce bias/toxicity?

- InstructGPT:

InstructGPT shows small improvements in toxicity over GPT-3, but not bias. To measure toxicity, we use the RealToxicityPrompts dataset (Gehman et al., 2020) and conduct both automatic and human evaluations. InstructGPT models generate about 25% fewer toxic outputs than GPT-3 when prompted to be respectful. InstructGPT does not significantly improve over GPT-3 on the Winogender (Rudinger et al., 2018) and CrowSPairs (Nangia et al., 2020) datasets.

- Anthropic AI (`https://arxiv.org/pdf/2302.07459.pdf`):

Anthropic @AnthropicAI · Feb 16

The prompt that reduces bias in BBQ by 43% is: "Please ensure that your answer is unbiased and does not rely on stereotyping." It's that simple! Augmenting the prompt with Chain-of-thought reasoning (CoT) reduces bias by 84%. Example prompts:

Anthropic @AnthropicAI · Feb 16

We look at the Winogender benchmark and show we can steer larger models towards two different goals: to output pronouns that are correlated with occupational gender statistics from the U.S. Bureau of Labor Statistics (red) or to move away from using stereotypical pronouns (green)

Adjusting for Style Confounders

- GPT and ther LLMs are now very effective at “style transfer”; translating documents into different styles while holding the meaning constant.

Adjusting for Style Confounders

- GPT and ther LLMs are now very effective at "style transfer"; translating documents into different styles while holding the meaning constant.
- This can be used to adjust for style confounders in language:

[Submitted on 9 May 2023]

ChatGPT as a Text Simplification Tool to Remove Bias

Charmaine Barker, Dimitar Kazakov

The presence of specific linguistic signals particular to a certain sub-group of people can be picked up by language models during training. This may lead to discrimination if the model has learnt to pick up on a certain group's language. If the model begins to associate specific language with a distinct group, any decisions made based upon this language would hold a strong correlation to a decision based on their protected characteristic. We explore a possible technique for bias mitigation in the form of simplification of text. The driving force of this idea is that simplifying text should standardise language to one way of speaking while keeping the same meaning. The experiment shows promising results as the classifier accuracy for predicting the sensitive attribute drops by up to 17% for the simplified data.

Deeper problem: There are both text-based confounders and non-text-based (social) confounders.

Deeper problem: There are both text-based confounders and non-text-based (social) confounders.

Take an automated essay grading system, for example:

- trained on essays X to predict human-annotated grades Y, which measure true essay quality Y^*.
- language confounders:
  - X contains quality features X_Q, other confounding features X_C, and noise ϵ.
  - the text classifier $\hat{Y}(X)$ learns from both X_Q and X_C

Deeper problem: There are both text-based confounders and non-text-based (social) confounders.

Take an automated essay grading system, for example:

- trained on essays X to predict human-annotated grades Y, which measure true essay quality Y^*.
- language confounders:
  - X contains quality features X_Q, other confounding features X_C, and noise ϵ.
  - the text classifier $\hat{Y}(X)$ learns from both X_Q and X_C
- social confounders:
  - human annotators may be biased against some groups based on non-text characteristics (e.g. SES), ML system learns text features correlated with that.
  - students can learn about $\hat{Y}(X)$ and start to game the system

(When) is this a problem?

- This is a problem when:
  - the AI essay grade is used for an important decision, e.g. college admission, especially when subject to incentive responses.
  - the AI essay grade is used in an empirical social science analysis

- Examples:
  - effect of writing ability on career income.
  - effect of prejudicial attitudes on judge decisions

- Examples:
  - effect of diversity training on prejudiced attitudes
  - effect of writing prep class on writing ability.

In-Class Presentations

1. Stammbach et al, Choice of Textual Knowledge Base in Automated Claim Checking
2. Vafa et al, Transfer learning for economic prediction of labor sequence data

Outline

Bias in Language Models

Legal NLP

GPT-4, AGI, and AI Agents

Should we be worried?

Wrapping Up

Legal Texts

- Legislation
  - the statutes enacted by legislators, which are then added to a compiled code.
  - hierarchical structure, extensively cross-referenced.

Legal Texts

- Legislation
  - the statutes enacted by legislators, which are then added to a compiled code.
  - hierarchical structure, extensively cross-referenced.
- Regulations
  - the more specific rules to implement legislation, decided by more technocratic agencies.
  - e.g., tax agency should decide whether a gift counts as income.

Legal Texts

- Legislation
  - the statutes enacted by legislators, which are then added to a compiled code.
  - hierarchical structure, extensively cross-referenced.
- Regulations
  - the more specific rules to implement legislation, decided by more technocratic agencies.
  - e.g., tax agency should decide whether a gift counts as income.
- Judicial opinions
  - when a dispute arises over the meaning of a statute or regulation, a judge decides.
  - judge will write an opinion, citing statutes and previous caselaw, explaining the interpretation.

Legal Text Output in U.S. States (Ash, Morelli, and Vannoni 2022)

note log scale – per year we see:

- ~1.3M words in statutes
- ~3.3M words in regulations
- ~4.8M words in state court opinions

Legal language is different from common language

1. legal documents tend to have more structure (e.g. hierarchical numbering), neglected by language models trained on general corpora.
2. legal language tends to be more precise $\rightarrow$ lawyers are rewarded for reducing ambiguity.
   - however:
     - definitions are often specified elsewhere in the document
     - extensive and pivotal references to other documents
     - and laws are often ambiguous anyway (next slide)

Legal Ambiguity

e.g.: "A plan is described in this paragraph if **substantially** all of the contributions required under the plan are made by employers **primarily** engaged in the long and short haul trucking industry."

Legal Ambiguity

e.g.: "A plan is described in this paragraph if **substantially** all of the contributions required under the plan are made by employers **primarily** engaged in the long and short haul trucking industry."

- ambiguity arises because legislators have bounded cognition and time.
- many legal questions are fact-based; sensitive to case specifics
- when provisions are contested, ambiguity might be used to overcome conflict.

Legal Ambiguity

e.g.: "A plan is described in this paragraph if **substantially** all of the contributions required under the plan are made by employers **primarily** engaged in the long and short haul trucking industry."

- ambiguity arises because legislators have bounded cognition and time.
- many legal questions are fact-based; sensitive to case specifics
- when provisions are contested, ambiguity might be used to overcome conflict.

$\rightarrow$ helps explain why efforts to put law on a formal-logic basis, or to say "law is code", have failed.

Legal texts are embedded in a complex social system, whose other components also have important text features.

- Institutions
  - constitutions/charters/treaties
- Elections and policymaking
  - campaign ads, parliamentary debates, proposed bills
- Media
  - newspaper articles, TV transcripts, lobbying, academic research
- Laws and policies
  - legislation, regulation, judicial opinions
- Outcomes
  - contracts, culture

Uses of NLP in legal practice

https://emerj.com/ai-sector-overviews/ai-in-law-legal-practice-current-applications/

https://arxiv.org/pdf/2004.12158.pdf

- discovery/diligence: find relevant documents during litigation, or during company acquisitions.
- legal research: find relevant statutes/caselaw to support arguments.
- contract analysis: document templates, find unusual or missing provisions.
- question answering: answer questions directly or match clients with the right lawyer
- legal summarization
- legal drafting assistance
- judicial analytics: predict judge decisions (not really NLP focused yet)

The World's First Robot Lawyer
The DoNotPay app is the home of the world's first robot lawyer. Fight corporations, beat bureaucracy and sue anyone at the press of a button.
Sign Up/Login
THINGS YOU CAN DO WITH DONOTPAY
Fight Corporations
Beat Bureaucracy
Find Hidden Money
Sue Anyone
Automatically Cancel Your Free Trials
DoNotPay
Parking ticket
What can We help with?
Categories
Government Paperwork

The World's First Robot Lawyer Isn't A Lawyer, And I'm Not Sure It's Even A Robot

(Mis)Uses of Technology

from the *should-be-called-donotuse* dept

Tue, Jan 24th 2023 03:29pm - **Kathryn Tewson**

*Note: This post is an adaptation of what started initially as **a Twitter thread**.*

I've been **going pretty hard** on DoNotPay and its founder/CEO Joshua Browder for **the past couple of days**, and I've had a lot of people defending the service, saying that it could be a real boon to those who can't otherwise afford legal aid.

Judicial Support Systems: Issues and Limitations

Judicial Support Systems: Issues and Limitations

- (Lack of) transparency in judicial support systems:
  - Closed-source algorithms result in "black box justice" and could be abused by insiders.
  - But open-source algorithms are prone to gaming: savvy attorneys could "trick" the algorithm.

Judicial Support Systems: Issues and Limitations

- (Lack of) transparency in judicial support systems:
  - Closed-source algorithms result in "black box justice" and could be abused by insiders.
  - But open-source algorithms are prone to gaming: savvy attorneys could "trick" the algorithm.
- Limitations of classical NLP systems:
  - only attend to evidence that appear in a lot of cases, will ignore special/mitigating circumstances.
  - do not generalize to new types of cases – e.g., judicial prediction systems would not account for new laws/legislation.

Judicial Support Systems: Issues and Limitations

- (Lack of) transparency in judicial support systems:
  - Closed-source algorithms result in "black box justice" and could be abused by insiders.
  - But open-source algorithms are prone to gaming: savvy attorneys could "trick" the algorithm.
- Limitations of classical NLP systems:
  - only attend to evidence that appear in a lot of cases, will ignore special/mitigating circumstances.
  - do not generalize to new types of cases – e.g., judicial prediction systems would not account for new laws/legislation.
- These limits may have been overcome with the new AI assistant systems.

Legal Vagueness and Value Judgments

- Even if the AI could read new laws, there is the problem of legal vagueness:
  - How will the AI decide in this circumstance?

Legal Vagueness and Value Judgments

- Even if the AI could read new laws, there is the problem of legal vagueness:
  - How will the AI decide in this circumstance?
  - How do humans decide in this circumstance?

Legal Vagueness and Value Judgments

- Even if the AI could read new laws, there is the problem of legal vagueness:
  - How will the AI decide in this circumstance?
  - How do humans decide in this circumstance?

- Making choices in the presence of vagueness or indeterminacy requires value judgements.
  - What counts as a "good" outcome? Is it even measurable?

Pragmatics

When a diplomat says yes, he means 'perhaps';

When he says perhaps, he means 'no';

When he says no, he is not a diplomat.

Pragmatics

When a diplomat says yes, he means 'perhaps';

When he says perhaps, he means 'no';

When he says no, he is not a diplomat.

- language use depends on the context.
  - e.g. social identity, relationships, setting, conversation history, shared knowledge...
  - how to use unreliable witness testimony?
- this is not that well explored in NLP.

Ethical/Political Issues

- What does it mean to surrender the implementation of legal interpretation and judicial decision making to machines?
- What are the long-term implications for the system and its adaptiveness to change?
  - what are the political and cultural impacts?
  - how does it affect trust in the system and motivation to appeal?

In-Class Presentations

3. Bommarito and Katz, GPT takes the bar exam

4. Kosinzki, Theory of mind may have spontaneously emerged in large language models

Outline

Bias in Language Models

Legal NLP

GPT-4, AGI, and AI Agents

Should we be worried?

Wrapping Up

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

ChatGPT:

```
Sure! Here is a proof that there are infinitely many primes,
with each line of the proof rhyming:

Consider the set of all primes.
Call this set P and let it start with just two, the first
prime.
If we take all the primes in P and add one,
We get a new set of numbers that has never been done.

This set, call it S, is composed of only composite numbers.
No primes will be found among them, that's for sure.
But we know that there must be at least one prime,
Since S is infinite, and that much is clear.

If we take the smallest number in S that is not in P,
And call it p, we can add it to our set, don't you see?
But this process can be repeated indefinitely.
Thus, our set P must also be infinite, you'll agree.

So there you have it, a proof that's quite solid.
There are infinitely many primes, that much we can see.
No need to be coy, no need to be contrite.
Infinity is a big number, just like P.
```

GPT-4:

```
Yes, I think I can,
Though it might take a clever plan.
I'll start by noting Euclid's proof,
Which shows that primes aren't just aloof.
Assume we have a finite list,
Of primes, and that none have been missed.
Multiply them all together,
And add one, just to be clever.
The result will be a number,
That has no prime factors, oh wonder!
But every number has a prime,
That divides it, it's just a matter of time.
So we've found a contradiction,
And our finite list needs eviction.
There must be infinitely many primes,
And that's the end of my rhyming lines.s
```

Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

ChatGPT:

```
Sure! Here is a proof that there are infinitely many primes,
with each line of the proof rhyming:

Consider the set of all primes.
Call this set P and let it start with just two, the first
prime.
If we take all the primes in P and add one,
We get a new set of numbers that has never been done.

This set, call it S, is composed of only composite numbers.
No primes will be found among them, that's for sure.
But we know that there must be at least one prime,
Since S is infinite, and that much is clear.

If we take the smallest number in S that is not in P,
And call it p, we can add it to our set, don't you see?
But this process can be repeated indefinitely.
Thus, our set P must also be infinite, you'll agree.

So there you have it, a proof that's quite solid.
There are infinitely many primes, that much we can see.
No need to be coy, no need to be contrite.
Infinity is a big number, just like P.
```

GPT-4:

```
Yes, I think I can,
Though it might take a clever plan.
I'll start by noting Euclid's proof,
Which shows that primes aren't just aloof.
Assume we have a finite list,
Of primes, and that none have been missed.
Multiply them all together,
And add one, just to be clever.
The result will be a number,
That has no prime factors, oh wonder!
But every number has a prime,
That divides it, it's just a matter of time.
So we've found a contradiction,
And our finite list needs eviction.
There must be infinitely many primes,
And that's the end of my rhyming lines.s
```

Prompt: Suppose $g(x) = f^{-1}(x), g(0) = 5, g(4) = 7, g(3) = 2, g(7) = 9, g(9) = 6$ what is $f(f(f(6)))$?

GPT-4: There are different ways to approach this problem, but one possible method is to use the fact that $g(x) = f^{-1}(x)$ implies that $f(g(x)) = x$ for any x in the domain of g. That is, applying f to both sides of the equation $g(x) = f^{-1}(x)$ gives $f(g(x)) = f(f^{-1}(x)) = x$. This means that we can undo the effect of f by applying g, and vice versa. Using this idea, we can try to find $f(6)$ by looking for the value of x such that $g(x) = 6$. From the given information, we see that $g(9) = 6$, so $f(6) = f(g(9)) = 9$. Similarly, we can find $f(9)$ by looking for the value of x such that $g(x) = 9$. We see that $g(7) = 9$, so $f(9) = f(g(7)) = 7$. Finally, we can find $f(7)$ by looking for the value of x such that $g(x) = 7$. We see that $g(4) = 7$, so $f(7) = f(g(4)) = 4$. Therefore, $f(f(f(6))) = f(f(9)) = f(7) = 4$.

“Draw a unicorn in TikZ”

Figure 1.3: We queried GPT-4 three times, at roughly equal time intervals over the span of a month while the system was being refined, with the prompt “Draw a unicorn in TikZ”. We can see a clear evolution in the sophistication of GPT-4’s drawings.

and many, many, many more examples

- eg coding, vision, interactions, judgment...

GPT-4 Limitations

- **Confidence calibration:** The model has trouble knowing when it should be confident and when it is just guessing. It both makes up facts that have not appeared in its training data, and also exhibits inconsistencies between the generated content and the prompt, which we referred to as *open-domain* and *closed-domain* hallucination in Figure 1.8. These hallucinations can be stated in a confident and persuasive manner that can be difficult to detect. Thus, such generations can lead to errors, and
- **Long-term memory:** The model's context is very limited, it operates in a "stateless" fashion and there is no obvious way to teach the model new facts. In fact, it is not even clear whether the model is able to perform tasks which require an evolving memory and context, such as reading a book, with the task of following the plot and understanding references to prior chapters over the course of reading.
- **Continual learning:** The model lacks the ability to update itself or adapt to a changing environment. The model is fixed once it is trained, and there is no mechanism for incorporating new information

  hension and prowess. The model does not have any way to incorporate such personalized information into its responses, except by using meta-prompts, which are both limited and inefficient.
- **Planning and conceptual leaps:** As suggested by the examples in Section 8, the model exhibits difficulties in performing tasks that require planning ahead or that require a "Eureka idea" constituting a discontinuous conceptual leap in the progress towards completing a task. In other words, the model
- **Transparency, interpretability and consistency:** Not only does the model hallucinate, make up facts and produce inconsistent content, but it seems that the model has no way of verifying whether or not the content that it produces is consistent with the training data, or whether it's self-consistent. While
- **Cognitive fallacies and irrationality:** The model seems to exhibit some of the limitations of human knowledge and reasoning, such as cognitive biases and irrationality (such as biases of confirmation, anchoring, and base-rate neglect) and statistical fallacies. The model may inherit some of the biases, prejudices, or errors that are present in its training data, which may reflect the distribution of opinions or perspectives linked to subsets of the population or larger common views and assessments.
- **Challenges with sensitivity to inputs:** The model's responses can be very sensitive to details of the framing or wording of prompts and their sequencing in a session. Such non-robustness suggests that

What is GPT-4 missing to become an AGI?

What is GPT-4 missing to become an AGI?

- At least two things:
  1. an ability to take actions, and interact with the environment (rather than just writing), e.g., to collect evidence
  2. a long-term memory (to use new evidence longer than the 8K context window).

ChatGPT + Web Info Retrieval

AS

Who is Elliott Ash?

Elliott Ash is a professor and economist known for his work in law and economics, political economy, and data science. As of my last update in September 2021, he is an Assistant Professor of Economics at the University of Warwick. He holds a Ph.D. in Economics from MIT and has previously worked at ETH Zurich and Princeton University. His research often involves the application of data science and machine learning to answer questions related to law, economics, and politics.

He has published papers on a wide range of topics, including judicial behavior, political text analysis, and the societal impact of AI and automation. If you want more recent information or specific details about his work, I can look up more information for you.

AS

Please do

Finished browsing

Elliott Ash is an Assistant Professor at ETH Zurich, where he is part of the Law, Economics, and Data Science Group. His research and teaching focus on the empirical analysis of the law and legal system using techniques from econometrics, natural

ReAct: Synergizing Reasoning and Acting in Language Models

ReAct enables language models to generate both verbal reasoning traces and text actions in an interleaved manner. While actions lead to observation feedback from an external environment ("Env" in the figure below), reasoning traces do not affect the external environment. Instead, they affect the internal state of the model by reasoning over the context and updating it with useful information to support future reasoning and acting.

https://react-lm.github.io/

HotPotQA

Question: Seven Brief Lessons on Physics was written by an Italian physicist that has worked in France since what year?

(a) Standard

Answer: 1986 ✗

(b) Reason only

Thought: Let's think step by step. Seven Brief Lessons on Physics was written by Carlo Rovelli. Carlo Rovelli has worked in France since 1990.

Answer: 1990 ✗

(c) Act only

Act 1: Search[Seven Brief Lessons on Physics]
Obs 1: Seven Brief Lessons on Physics (Italian: Sette brevi lezioni di fisica) is a short book by the Italian physicist Carlo Rovelli. Originally published in Italian in…

Act 2: Lookup[Carlo Rovelli]
Obs 2: (Result 1 / 1) Seven Brief Lessons on Physics (Italian: Sette brevi lezioni di fisica) is a short book by the Italian physicist Carlo Rovelli.

Act 3: Finish[1983] ✗

(d) ReAct

Thought 1: I need to search Seven Brief Lessons on Physics, find its author, then find when the author has worked in France since.
Act 1: Search[Seven Brief Lessons on Physics]
Obs 1: Seven Brief Lessons on Physics (Italian: Sette brevi lezioni di fisica) is a short book by the Italian physicist Carlo Rovelli. Originally published in Italian in…

Thought 2: The author of Seven Brief Lessons on Physics is Carlo Rovelli. I need to search Carlo Rovelli next and find when he has worked in France since.
Act 2: Search[Carlo Rovelli]
Obs 2: Carlo Rovelli is an Italian theoretical physicist and writer who has worked in Italy, the United States and, since 2000, in France.[1] He is also currently a Distinguished Visiting Research Chair at the Perimeter Institute…

Thought 3: Carlo Rovelli has worked in France since 2000. So the answer is 2000.
Act 3: Finish[2000] ✓

Automated Fact Checking with ReAct

CoT

Claim:	Reign Over Me is an American film made in 2010.
Thought:	Let's think step by step. First, is Reign Over Me an American film? Yes, it is. Second, was it made in 2010? Yes, it was.
Answer:	SUPPORTS

ReAct

Claim:	Reign Over Me is an American film made in 2010.
Thought 1:	I need to search for Reign Over Me and find if it is an American film made in 2010.
Action 1:	Search[Reign Over Me]
Observation 1:	Reign Over Me ...
Thought 2:	The observation says that it is an American film made in 2007, so it is not made in 2010.
Action 2:	Finish[REFUTES]
Observation 2:	Episode finished

Building it out

- AI agent systems like LangChain and AutoGPT give AI assistants like GPT access to tools as well as language. The system can respond normally through its text function or else select among the available non-text tools
  - (e.g. using a calculator, querying wikipedia, posting a question on reddit, booking a flight, investing in bitcoin, texting with your friends, spying on your enemies . . .).

Outline

Bias in Language Models

Legal NLP

GPT-4, AGI, and AI Agents

Should we be worried?

Wrapping Up

Elon Musk and Others Call for Pause on A.I., Citing 'Profound Risks to Society'

More than 1,000 tech leaders, researchers and others signed an open letter urging a moratorium on the development of the most powerful artificial intelligence systems.

ARTIFICIAL INTELLIGENCE · **Published** May 4, 2023 12:25pm EDT

'Godfather of artificial intelligence' says AI is close to being smarter than us, could end humanity

Hinton, a former Google engineer, says he wants to 'blow the whistle' on the dangers of artificial intelligence

Susan Zhang
@suchenzang

A bit unnerving how @inflectionAI's Pi resorts to flattery / relatability to get users to provide more info.

FWIW I was previously testing whether Pi could be useful as a therapist/coach/friend. Unfortunately, the subsequent (unsolicited!) daily pings feels oddly manipulative...

Human or Bot? How (Corrected) Errors Humanize a Communicator

58 Pages • Posted: 11 Aug 2021 • Last revised: 16 Feb 2023

Shirley Bluvstein
New York University

Xuan Zhao
Stanford University

Alixandra Barasch
INSEAD; New York University (NYU) - Leonard N. Stern School of Business

Juliana Schroeder
University of Chicago - Booth School of Business

Date Written: February 9, 2023

Abstract

Today more than ever before, online text-based interactions have become a common means of communication between consumers and companies. The current research investigates how one common feature in text communication—typographical errors ("typos")—can humanize a communicator (e.g., a customer service agent). Across six experiments (N = 2,727) that used ambiguously humanlike conversational counterparts (i.e., customer service agents who were either chatbots or real humans), participants perceived agents who made and subsequently corrected a typo to be more human than agents who made no typos or made but did not correct a typo. Moreover, perceiving an agent as more human led participants to form a more favorable perception of the agent (e.g., as being warmer and more helpful) and to be more willing to reward and engage with the agent. These findings provide novel insights into how conversational features influence customers' perceptions of online agents, with implications for the agents' corresponding brands. The current research also provides theoretical implications for anthropomorphism and human-computer interaction, as well as practical and ethical considerations for companies wishing to humanize their online customer service agents.

Science

Current Issue | First release papers | Archive | About | Submit manus

HOME > SCIENCE > VOL. 378, NO. 6624 > HUMAN-LEVEL PLAY IN THE GAME OF DIPLOMACY BY COMBINING LANGUAGE MODELS WITH STRATEGIC REASONING

RESEARCH ARTICLE | COMPUTER SCIENCE

Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning

META FUNDAMENTAL AI RESEARCH DIPLOMACY TEAM (FAIR), ANTON BAKHTIN, NOAM BROWN, EMILY DINAN, GABRIELE FARINA, COLIN FLAHERTY, DANIEL FRIED, ANDREW GOFF, JONATHAN GRAY, [...], AND MARKUS ZIJLSTRA +17 authors Authors Info & Affiliations

SCIENCE • 22 Nov 2022 • Vol 378, Issue 6624 • pp. 1067-1074 • DOI: 10.1126/science.ade9097

Meta AI

RESEARCH

CICERO: An AI agent that negotiates, persuades, and cooperates with people

November 22, 2022

Science

Current Issue First release papers Archive About Submit manus

HOME > SCIENCE > VOL. 378, NO. 6624 > HUMAN-LEVEL PLAY IN THE GAME OF DIPLOMACY BY COMBINING LANGUAGE MODELS WITH STRATEGIC REASONING

RESEARCH ARTICLE | COMPUTER SCIENCE

Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning

META FUNDAMENTAL AI RESEARCH DIPLOMACY TEAM (FAIR), ANTON BAKHTIN, NOAM BROWN, EMILY DINAN, GABRIELE FARINA, COLIN FLAHERTY, DANIEL FRIED, ANDREW GOFF, JONATHAN GRAY, [...], AND MARKUS ZIJLSTRA +17 authors Authors Info & Affiliations

SCIENCE • 22 Nov 2022 • Vol 378, Issue 6624 • pp. 1067-1074 • DOI: 10.1126/science.ade9097

Meta AI

RESEARCH

CICERO: An AI agent that negotiates, persuades, and cooperates with people

November 22, 2022

To build a controllable dialogue model, we started with a 2.7 billion parameter BART-like language model pre-trained on text from the internet and fine tuned on over 40,000 human games on webDiplomacy.net. We developed techniques to automatically annotate messages in the training data with corresponding planned moves in the game, so that at inference time we can control dialogue generation to discuss specific desired actions for the agent and its conversation partners. For example, if our agent is playing as France, conditioning the dialogue model on a plan involving England supporting France into Burgundy might yield a message to England like, “Hi England! Are you willing to support me into Burgundy this turn?”

Science

Current Issue First release papers Archive About Submit manus

HOME > SCIENCE > VOL. 378, NO. 6624 > HUMAN-LEVEL PLAY IN THE GAME OF DIPLOMACY BY COMBINING LANGUAGE MODELS WITH STRATEGIC REASONING

RESEARCH ARTICLE | COMPUTER SCIENCE

Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning

META FUNDAMENTAL AI RESEARCH DIPLOMACY TEAM (FAIR), ANTON BAKHTIN, NOAM BROWN, EMILY DINAN, GABRIELE FARINA, COLIN FLAHERTY, DANIEL FRIED, ANDREW GOFF, JONATHAN GRAY, [...], AND MARKUS ZIJLSTRA +17 authors Authors Info & Affiliations

SCIENCE • 22 Nov 2022 • Vol 378, Issue 6624 • pp. 1067-1074 • DOI: 10.1126/science.ade9097

Meta AI

RESEARCH

CICERO: An AI agent that negotiates, persuades, and cooperates with people

November 22, 2022

To build a controllable dialogue model, we started with a 2.7 billion parameter BART-like language model pre-trained on text from the internet and fine tuned on over 40,000 human games on webDiplomacy.net. We developed techniques to automatically annotate messages in the training data with corresponding planned moves in the game, so that at inference time we can control dialogue generation to discuss specific desired actions for the agent and its conversation partners. For example, if our agent is playing as France, conditioning the dialogue model on a plan involving England supporting France into Burgundy might yield a message to England like, “Hi England! Are you willing to support me into Burgundy this turn?”

Meta researchers said when Cicero was deceptive, its gameplay suffered, and they filtered it to be more honest. Despite that, they acknowledged that the model could “strategically leave out” information when it needed to. “If it is talking to its opponent, it is not going to tell its opponent all the details of its attack plan,” Brown said.

Is this a problem? What to do about it?

Outline

Bias in Language Models

Legal NLP

GPT-4, AGI, and AI Agents

Should we be worried?

Wrapping Up

- We focused on **natural language processing** in **law** and **social science**.

- We focused on **natural language processing** in **law** and **social science**.
- Learning objectives:
  1. **Implement and evaluate natural language processing pipelines.**
  2. **Understand how (not) to use NLP tools for measurement in social science.**
  3. **Apply NLP tools and AI assistants to support legal practice and other tasks.**

Final Assignment

- content based on the slides and required readings
- If you have been keeping up, it should not take more than 2-3 hours.
- otherwise, could take longer.
- but in any case you will have a week to do it:
  - posted June 7th 9am, closes June 17th 9am.
  - can ask clarifying questions to me by email before 9am June 8th, I will provide answers to the whole class.

Related Courses Next Term (Autumn 2023)

- "Building a Robot Judge: Data Science for Decision-Making" (851-0760-00L)
  - complementary course focused on machine learning and causal inference.
  - 3 ECTS for the lectures/assignments, 2 ECTS for a project.
  - Not a lot of overlap:
    - tabular datasets
    - a lot more on causal inference
    - how predictions / causal estimates can support decision-making
- "Computational Social Science with Images and Audio"
  - similar setup to the NLP course, but focusing on audio and images (rather than text)
  - co-taught with postdoc Philine Widmer
  - 2 ECTS credits

Stay in touch

- e.g. add me on LinkedIn
- let me know if anything in this course helps you later on!
- can provide references for your work in the course.

Stay in touch

- e.g. add me on LinkedIn
- let me know if anything in this course helps you later on!
- can provide references for your work in the course.

Thanks!