Challenges in AI/ML for Health: Bias, Generalizability, Privacy

Luca Foschini, PhD

FDA-MCERSI Workshop on Application of Artificial Intelligence and Machine Learning for Precision Medicine

Feb 17th, 2023

luca@sagebionetworks.org

@calimagna
AI Bot ChatGPT Passes US Medical Licensing Exams Without Cramming – Unlike Students

Alicia Ault
January 26, 2023

ChatGPT can pass parts of the US medical licensing exam, researchers have found, raising questions about whether the AI chatbot could one day help write the exam or help students prepare for it.

AI=Trustworthy?

News > Medscape Medical News

Al Bot ChatGPT Passes US Medical Licensing Exams Without Cramming – Unlike Students

Alicia Ault
January 26, 2023

ChatGPT can pass parts of the US medical licensing exam, researchers have found, raising questions about whether the AI chatbot could one day help write the exam or help students prepare for it.


Reproduced from example viewed on Twitter, unable to find original attribution, please email me if you have info: luca@sagebionetworks.org
Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.
Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.
Figure 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.
Valid → Reproducible

Systematic evaluation of 300+ papers in:

- Computer vision
- Natural language processing
- General Machine Learning (ML)
- Machine learning for health (ML4H)

Evaluation Metrics:
A. Technical replicability
   1. Code available
   2. Public dataset
B. Statistical replicability
   1. Variance reported
C. Conceptual replicability
   1. Multiple datasets

Reproducibility in machine learning for health research: Still a ways to go. McDermott et al., SCIENCE TRANSLATIONAL MEDICINE 2021.
https://arxiv.org/abs/1907.01463
Reliable → NotBrittle

Synthesizing robust adversarial examples Athalye, et al., (ICML) 2018
https://arxiv.org/abs/1707.07397

Deep learning models for electrocardiograms are susceptible to adversarial attack
Han et al., NATURE MEDICINE 2020 https://arxiv.org/abs/1707.07397
SEE ALSO: Adversarial attacks on medical machine learning, Finlayson et al., SCIENCE (2019)
Unknown Bias → Lack of Reproducibility

Machine Learning COVID19 Detection from Wearables: The importance of study design. Nestor et al. (Accepted)
Preprint of prior version available: https://www.medrxiv.org/content/10.1101/2021.05.11.21257052v1
SEE ALSO: The performance of wearable sensors in the detection of SARS-CoV-2 infection: a systematic review. Mitratza & Goodale et al. LANCET DIGITAL HEALTH
Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.
Unmitigated Bias $\rightarrow$ Unfairness

Classifier trained on existing data can exhibit unequal error rates across races

Why Is My Classifier Discriminatory? Chen et al., (NeurIPS) 2018

May 31, 2022

Racial and Ethnic Discrepancy in Pulse Oximetry and Delayed Identification of Treatment Eligibility Among Patients With COVID-19

Ashraf Farzwy, MD, MPH; Tianshi David Wu, MD, MHS; Kunbo Wang, MS; et al


Dissecting racial bias in an algorithm used to manage the health of populations

ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN

Bias in medical AI products often runs under FDA’s radar. Hosgor & Akin STAT+

### Lack of Representation → Unmitigable Bias

Percentages of 518 FDA-approved AI products that submitted data covering sources of bias

<table>
<thead>
<tr>
<th></th>
<th>Aggregate Reporting</th>
<th>Stratified Reporting</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient Cohort</strong></td>
<td>less than 2% conducted multi-race/gender validation</td>
<td>less than 1% approval with performance figures across gender and race</td>
</tr>
<tr>
<td><strong>Medical Device</strong></td>
<td>8% conducted multi-manufacturer validation</td>
<td>less than 2% reported performance figures across manufacturers</td>
</tr>
<tr>
<td><strong>Clinical site</strong></td>
<td>less than 2% conducted multiside validation</td>
<td>less than 1% approvals with performance figures across sites</td>
</tr>
<tr>
<td><strong>Annotator</strong></td>
<td>less than 2% reported annotator/reader profile</td>
<td>less than 1% reported annotator/reader profile</td>
</tr>
</tbody>
</table>
New sources of Bias

Source: https://openai.com/blog/chatgpt/
Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.
Private?

Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person’s name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

Extracting Training Data from Large Language Models. Carlini et al.  
Figure 1: Our extraction attack. Given query access to a neural network language model, we extract an individual person’s name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.
Assess risk, then choose tradeoffs

Science

- Hypothesis generation
- Target Identification
- Screening
- Diagnostics
- Triaging
- Treatment decisions

Health care

Trade Offs (example)

- Valid & Reliable
- Private
- Safe
- Secure & Resilient
- Accountable & Transparent
- Explainable & Interpretable
- Safe
Trust → Verifiable (formally & automatically)

Proof of human-centric design

Standardized data description (e.g., Datasheet for Datasets)
Gebru & Krawford et al. 2018

Standardized model descriptions (e.g., Model Cards)
Mitchell & Gebru et al. 2018

Open benchmarks (e.g., DREAM Challenges)
https://dreamchallenges.org/

Standardized reporting metrics
Open interoperable protocols

Key Dimensions
- Application Context
  - Plan and Design
- Data & Input
  - Collect and Process Data
- AI Model
  - Build and Use Model
  - Verify and Validate
- Task & Output
  - Deploy and Use
- Application Context
  - Operate and Monitor
- People & Planet
  - Use or Impacted by
Thank You

Luca Foschini, PhD
luca@sagebionetworks.org
@calimagna