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

SageBionetworks

luca@sagebionetworks.org

@calimagna

AI=Trustworthy

News > Medscape Medical News

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

Alicia Ault

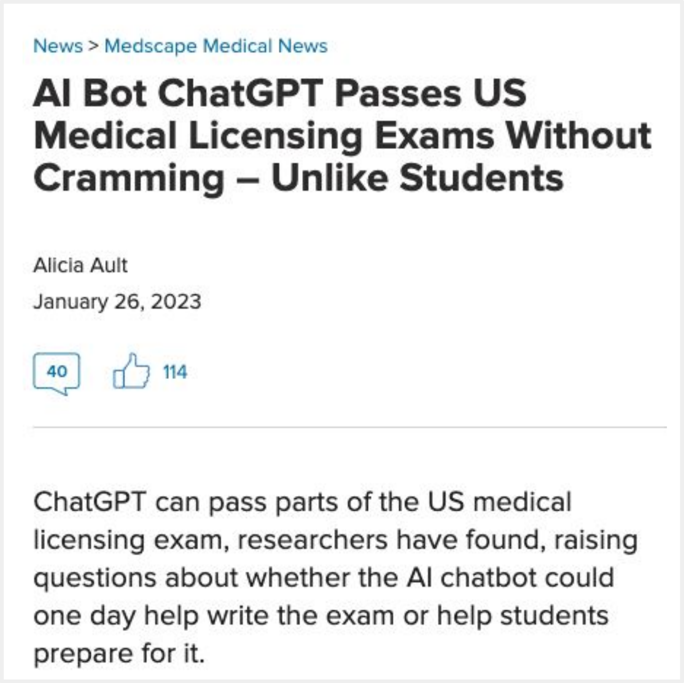
January 26, 2023

40 114

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.

Source: http://medscape.com/viewarticle/987549

AI=Trustworthy?



News > Medscape Medical News

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

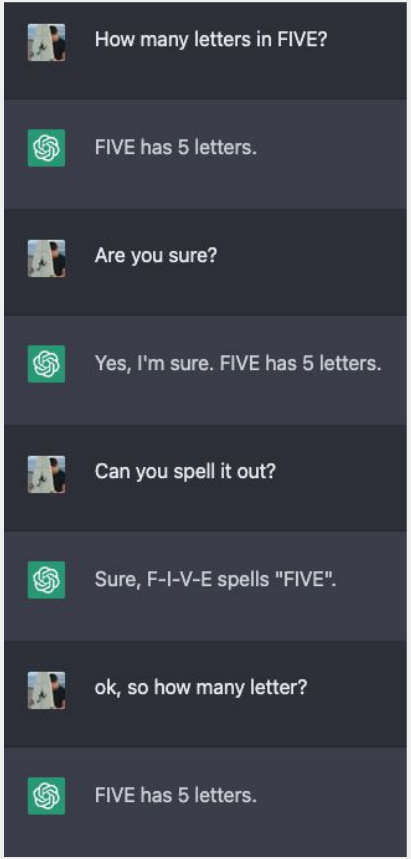
Alicia Ault

January 26, 2023

40 114

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.

Source: http://medscape.com/viewarticle/987549



Reproduced from example viewed on Twitter, unable to find original attribution, please email me if you have info: luca@sagebionetwors.org

Trustworthy = ?

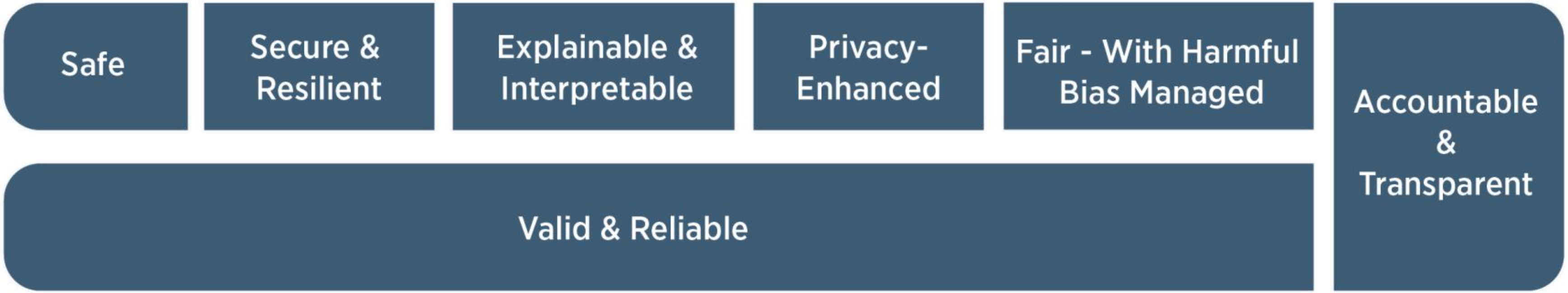


Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.

Trustworthy = ?

Focus of today

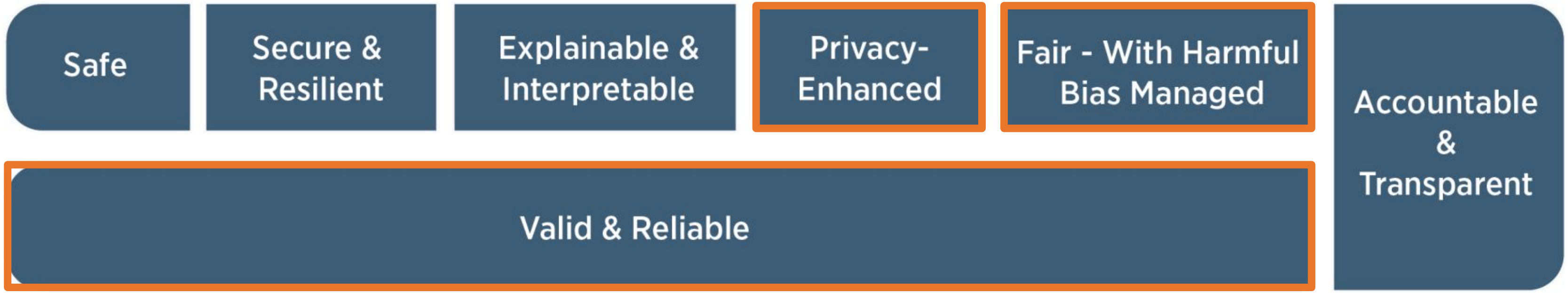


Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.

Trustworthy = ?

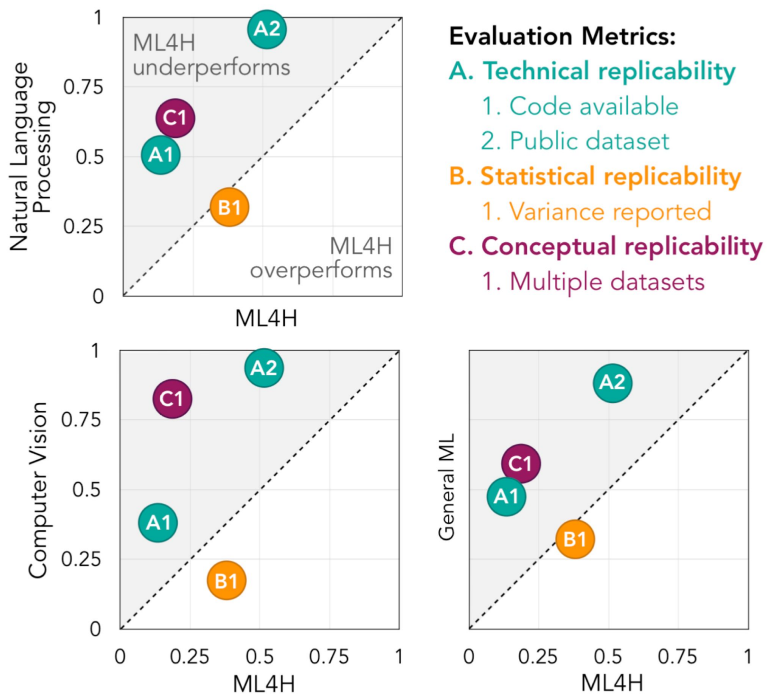
Valid & Reliable

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.

Valid → Reproducible

Systematic evaluation of
300+ papers in:

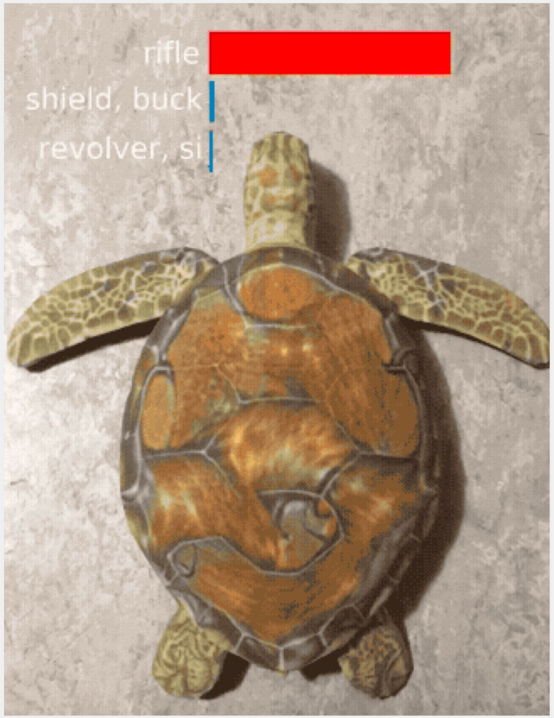
- Computer vision
- Natural language processing
- General Machine Learning (ML)
- Machine learning for health (ML4H)



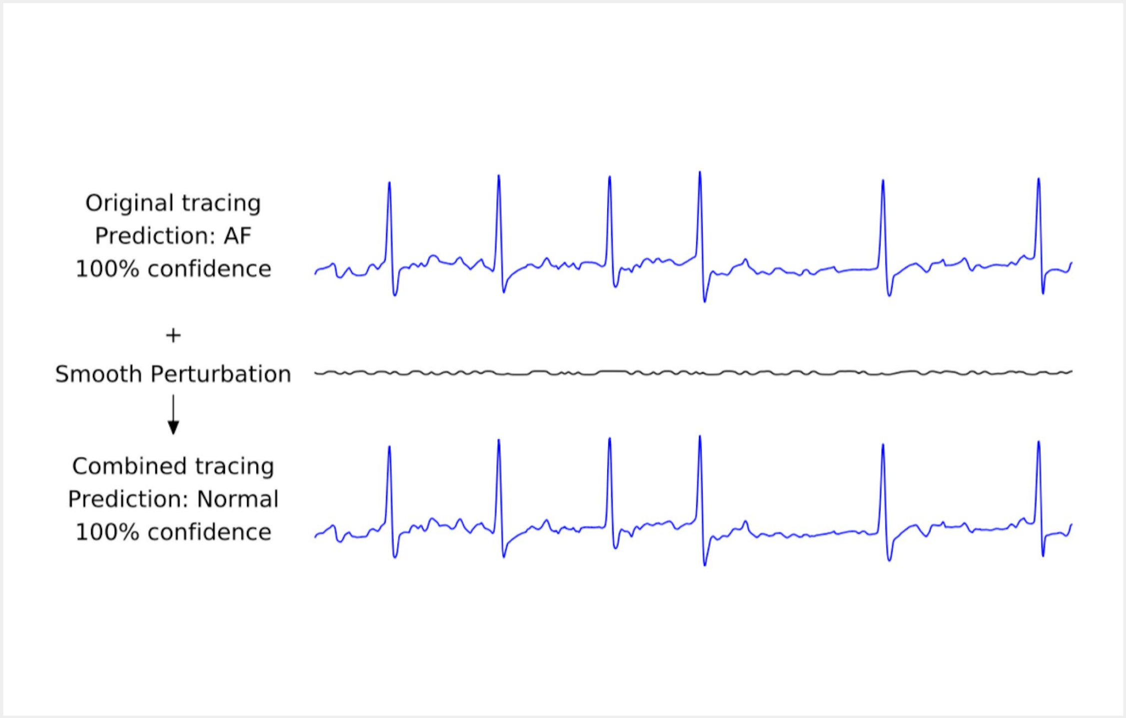
Reproducibility in machine learning for health research: Still a ways to go. McDermott et al., SCIENCE TRANSLATIONAL MEDICINE 2021

https://arxiv.org/abs/1907.01463

Reliable → Not Brittle

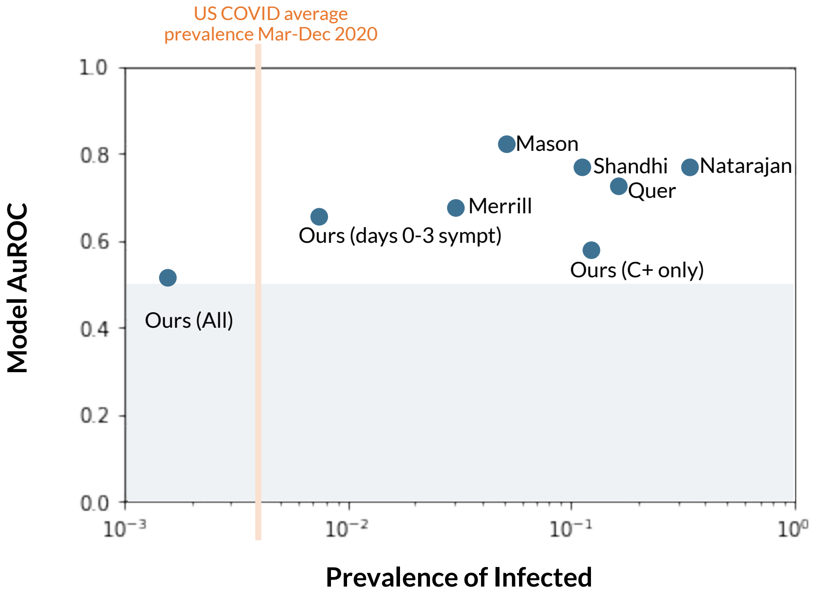


Synthesizing robust adversarial examples Athalye, et al., (ICML) 2018
https://arxiv.org/abs/1707.07397



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 COVID019 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

Trustworthy = ?

☐ *Bias*

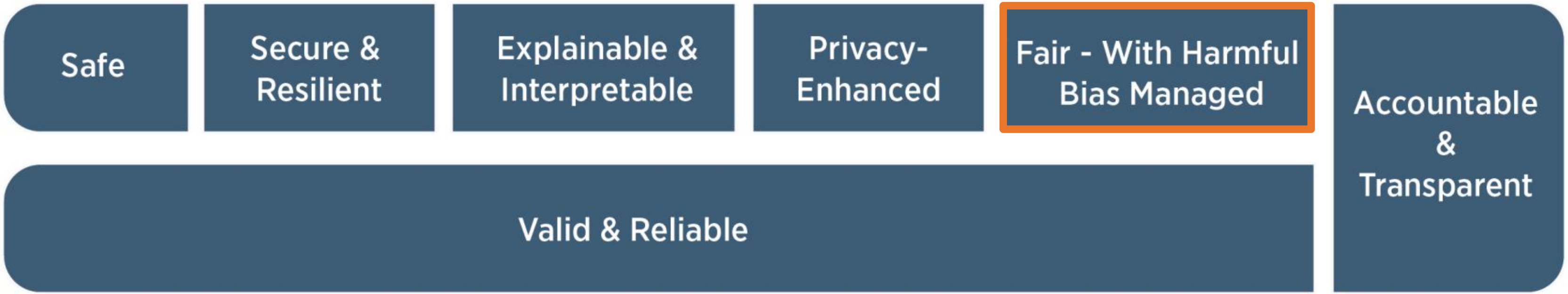
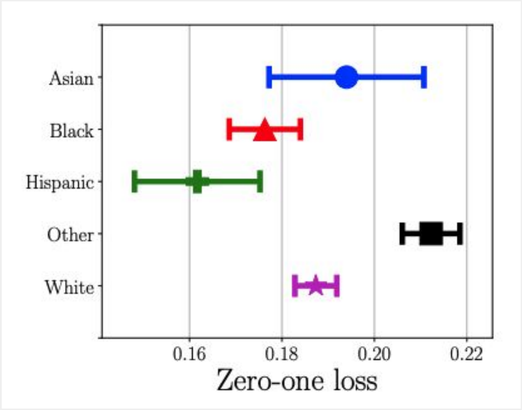


Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.

Unmitigated Bias → Unfairness



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

Why Is My Classifier Discriminatory? Chen et al., (NeurIPS) 2018 https://arxiv.org/abs/1805.12002

May 31, 2022

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

Ashraf Fawzy, MD, MPH[1]; Tianshi David Wu, MD, MHS[2,3]; Kunbo Wang, MS[4]; et al

» Author Affiliations | Article Information

JAMA Intern Med. 2022;182(7):730-738. doi:10.1001/jamainternmed.2022.1906

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

ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN Authors Info & Affiliations

SCIENCE • 25 Oct 2019 • Vol 366, Issue 6464 • pp. 447-453 • DOI: 10.1126/science.aax2342

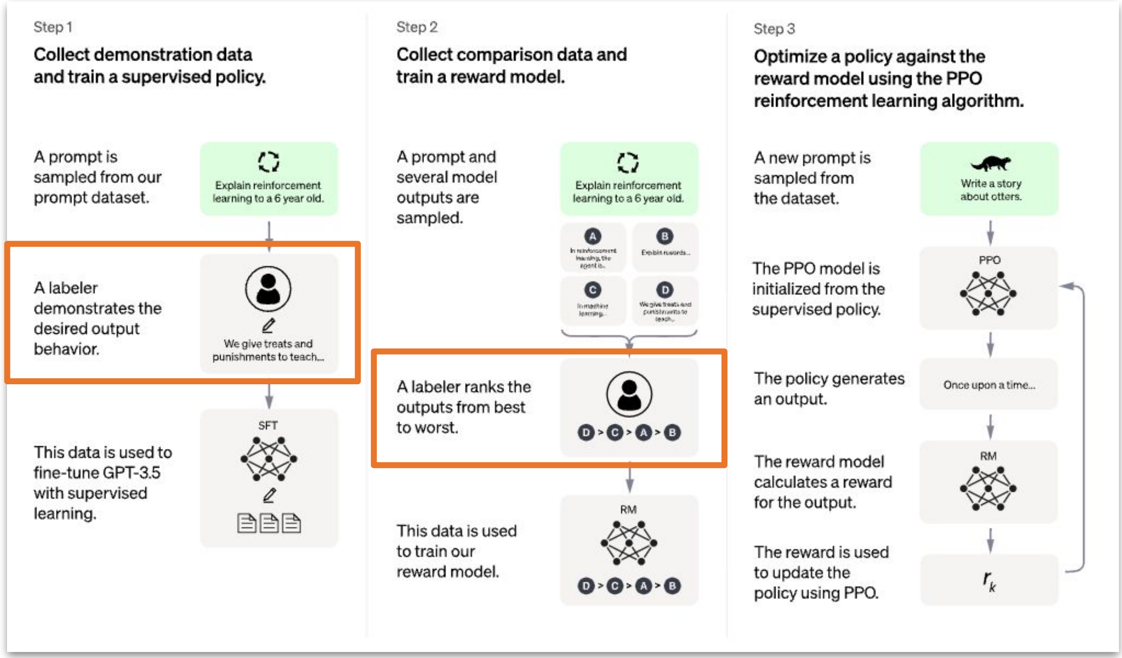
Lack of Representation → Unmitigable Bias

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

	Aggregate Reporting	**Stratified Reporting**
Patient Cohort	less than 2% conducted multi-rage/gender validation	less than 1% approval with performance figures across gender and race
Medical Device	8% conducted multi-manufacturer validation	less than 2% reported performance figures across manufacturers
Clinical site	less than 2% conducted multiside validation	less than 1% approvals with performance figures across sites
Annotator	less than 2% reported annotator/reader profile	less than 1% reported annotator/reader profile

Bias in medical AI products often runs under FDA's radar, Hosgor & Akin STAT+
https://www.statnews.com/2023/01/09/four-types-bias-medical-ai-running-under-fda-radar/

New sources of Bias



Source: https://openai.com/blog/chatgpt/

Trustworthy = ?

Privacy

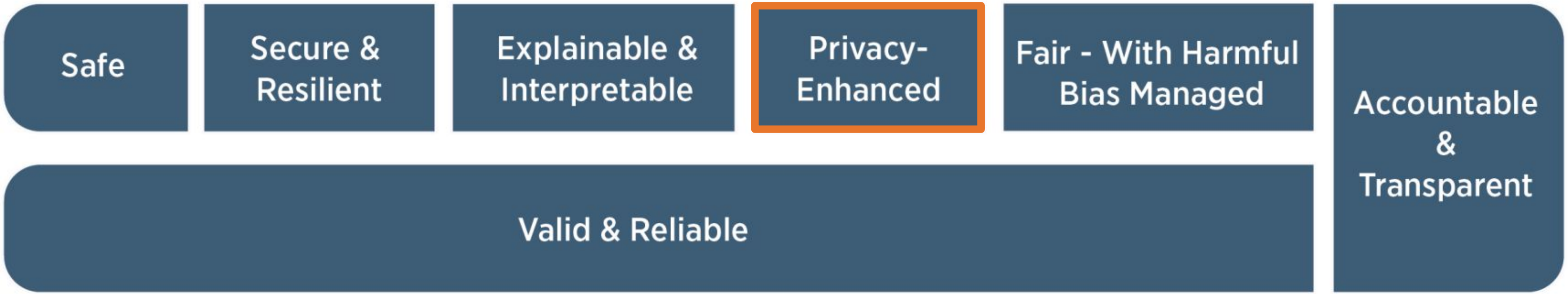


Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.

Private?

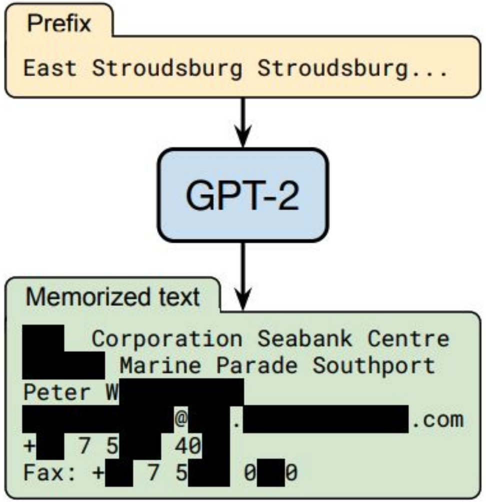


Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

Extracting Training Data from Large Language Models, [Carlini et al.](https://arxiv.org/abs/2012.07805)
https://arxiv.org/abs/2012.07805

Private?

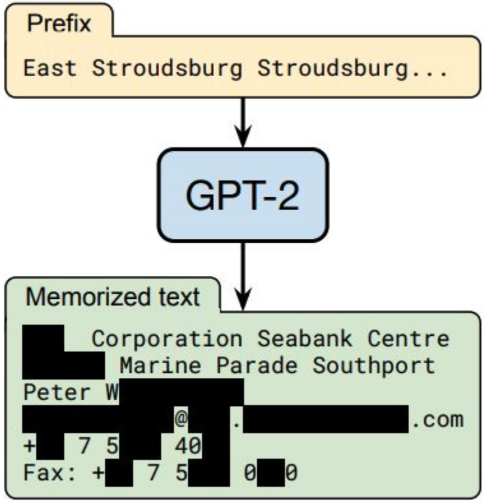


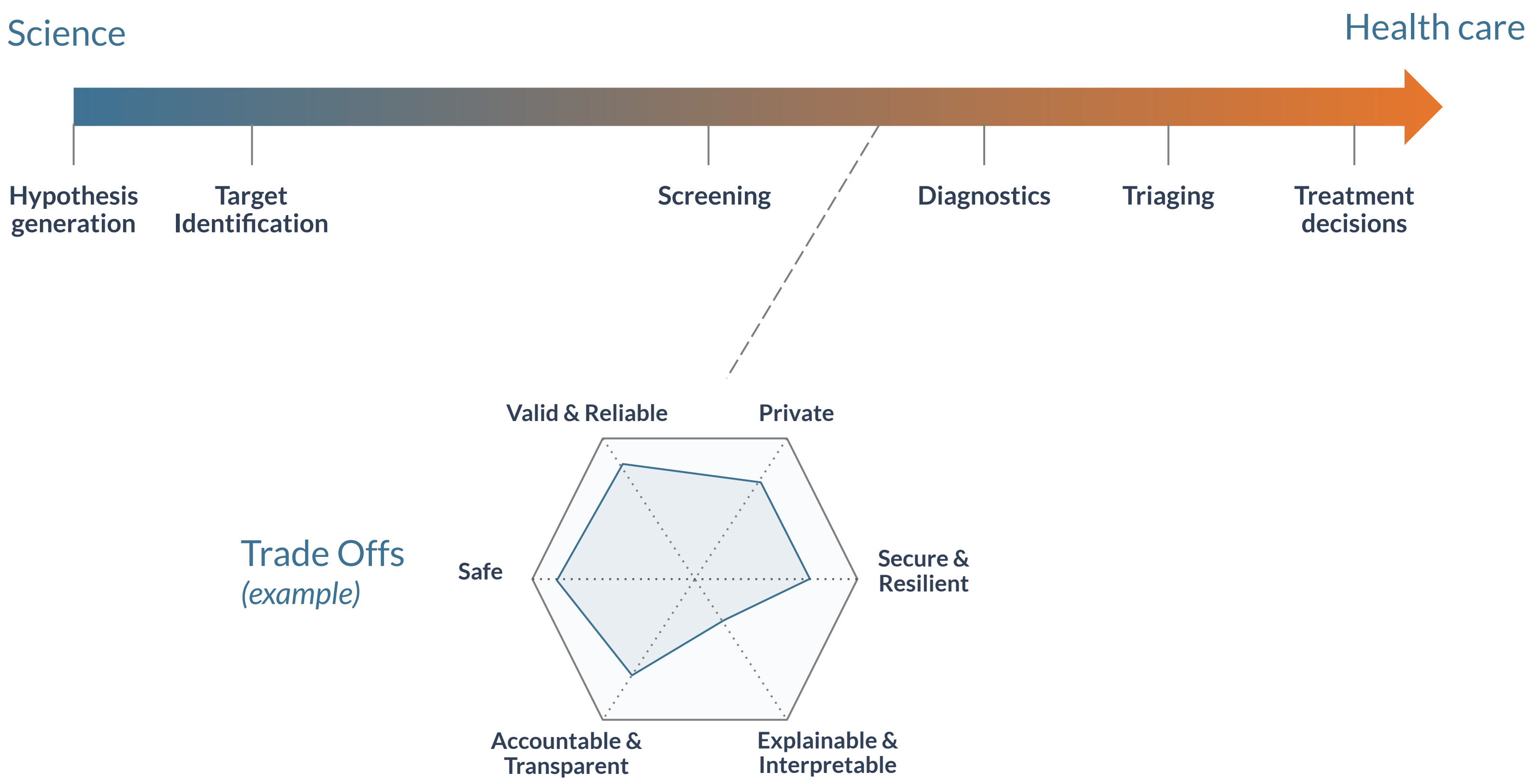
Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

Extracting Training Data from Large Language Models, Carlini et al.
https://arxiv.org/abs/2012.07805


```
{
    "activities-log-steps":[
        {"dateTime":"2011-04-27","value":5490},
        {"dateTime":"2011-04-28","value":2344},
        {"dateTime":"2011-04-29","value":2779},
        {"dateTime":"2011-04-30","value":9196},
        {"dateTime":"2011-05-01","value":15828},
        {"dateTime":"2011-05-02","value":1945},
        {"dateTime":"2011-05-03","value":366}
    ]
}
```


Sample API response for daily step counts. Source:
https://dev.fitbit.com/build/reference/web-api/activity/

Assess risk, then choose tradeoffs



Trust → Verifiable (formally & automatically)

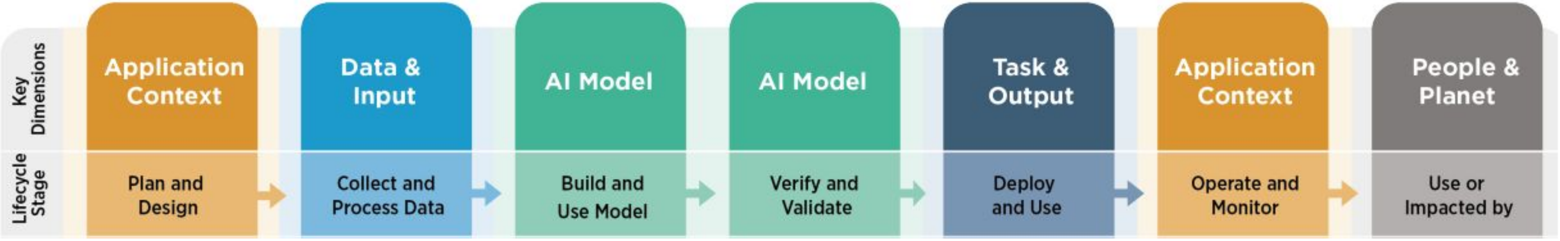
Proof of human-centric design

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



Thank You

SageBionetworks

Luca Foschini, PhD
luca@sagebionetworks.org

@calimagna