National Institute of Standards and Technology

Comment re: Four Principles of Explainable Artificial Intelligence

October 15, 2020

Buoy Health, Inc.
580 Harrison Ave,
Boston, MA 02118

October 15, 2020

National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899

Re: Four Principles of Explainable Artificial Intelligence (NISTIR 8312)

On behalf of Buoy Health, Inc. ("Buoy"), we are pleased to submit a comment in response to the National Institute of Standards and Technology ("NIST") first draft of *Four Principles of Explainable Artificial Intelligence* (NISTIR 8312).

Enclosed is the following:

- Comment re: Four Principles of Explainable Artificial Intelligence

The contact for this comment is Cory Lamz, Esq., Counsel & Data Privacy Officer, legal@buoyhealth.com.

Buoy appreciates the opportunity to submit this comment.

Warmly,

Buoy Health, Inc.

Amy Molten, MD, FAAP
Andrew Dumit
Cory Lamz, Esq., CIPP/US
Darin Baumgartel, PhD
Eddie Reyes
Greg Joondeph-Breidbart
Kun-Hua Tu, PhD
Mackenzie Mayberry

Introduction

Buoy Health, Inc. (“Buoy,” “we,” “our”) leverages artificial intelligence throughout our business. Given the accelerated pace of the advancement of AI solutions for healthcare and the potential value for leveraging data for public health, we are grateful for the consideration and development of explainability principles defined in Four Principles of Explainable Artificial Intelligence (NISTIR 8312) - Explanation, Meaningful, Explanation Accuracy, and Knowledge Limits. We acknowledge the value in developing core principles related to explainable AI, and we hope to continue to participate in this conversation as it develops.

Our comment addresses each of the four principles of explainability in the context of AI systems deployed in the consumer healthcare space. We also discuss the significance of both systematic bias and fairness as they relate to the principle of knowledge limits, and we argue that any organization leveraging AI in healthcare should account for these concepts when designing their systems.

Buoy’s AI Health Assistant product asks end users questions related to their symptoms during a short, conversational exchange. Based on these answers, our tool provides relevant medical information that empowers end users to self-diagnose and take further action, when necessary. The engine that determines what medical information is provided to end users based on their answers, as well as what next steps are relevant, is powered by AI. We therefore consider it crucial to analyze the relationship and application of these four principles of explainability to AI systems deployed in healthcare, particularly as they may relate to products like Buoy’s AI Health Assistant and the context of consumer healthcare.

Explanation

The principle of Explanation is paramount to the user experience in AI-based, consumer-facing healthcare applications. As NISTIR 8312 indicates, explanation is a vital element of trustworthy AI, and we believe this is particularly important in the healthcare industry due to the criticality and personal nature of the interactions between the user and the AI system. In consumer healthcare, trust is essential in the patient-provider relationship.[1] So too should be the case with an AI-based system. Therefore, we would argue that an individual interacting with an AI-based digital health tool is more likely to engage with medical information provided by a system that they trust.

Meaningful

[1] Rosemary Rowe, Michael Calnan, Trust relations in health care—the new agenda, *European Journal of Public Health*, Volume 16, Issue 1, February 2006, Pages 4–6, https://doi.org/10.1093/eurpub/ckl004

The principle of Meaningful plays a pivotal role in explainable AI, especially in the context of consumer healthcare. First, prior knowledge of a health concern is a critical component of meaningfulness, as prior knowledge could impact an end user's interpretation of an output. Designers of AI systems in the consumer healthcare space must be especially aware of this, and the AI system they design should seek to understand how to frame the explanations in pursuit of meaningfulness. Explanations should, therefore, be concise, unambiguous, and address preconceived biases and prior knowledge of the end user. To account for the wide variety of familiarity, designers of these tools should take care to remove medical jargon, ensure all explanations are at a reasonable cognitive level, and promote ethical transparency by explaining for what the system was optimized (e.g., clinical benefit vs. profit). Second, explainable AI systems in the consumer healthcare space must consider the end user's health timeline, and meaningful explanations should contextualize the current predictions to the end user's possible future states of health - an AI system that explains its decisions in a meaningful way at one moment in time must also be aware of potential future states and make its explanation meaningful in those future states as well.

Explanation Accuracy

The principle of Explanation Accuracy is of particular relevance in the consumer healthcare space due to the nature of the user's intent and prior knowledge of medicine. End users of a symptom assessment tool often fall into two distinct categories: (1) those who are researching symptoms with concerns about a particular condition or diagnosis, and (2) those who are in the initial stages of research and lack such pre-existing concerns. Explanation accuracy manifests in different ways for these two categories. For both, it is essential to explain to the end user the underlying motivating factors for the symptom assessment tool's output(s) - for example, communicating to a user that they verified a set of cardinal symptoms, which may be indicative of a certain illness among a certain percentage of individuals. In addition, for end users that may fall into the first category, providing reasoning for the output and a justification for what was *not* included in the output (i.e., counterfactuals) is essential in the furtherance of explanation accuracy - for example, indicating that the AI did not suggest a certain condition about which the user was concerned, because the user did not verify a set of symptoms that are typical of that illness.

Knowledge Limits

The principle of Knowledge Limits must be carefully considered in the consumer healthcare space. There are two components to knowledge limits in the healthcare space to consider when designing explainable AI systems: (1) known unknowns and (2) the unknown unknowns.

The known unknowns in the consumer healthcare space typically take the form of examples provided in NISTIR 8312. On one hand, a medical AI system may not have been trained for the end user currently interacting with the system, so the system should generate explanations in accordance with that fact. This is analogous to the picture of the apple in the bird-classifier example. On the other hand, the system should communicate its confidence about examples like the given one. In other words, the AI system should explain it is X% accurate for users similar to one currently using the system.

The unknown unknowns in the consumer healthcare space for explainable AI stem from three requirements:

First, these systems must make predictions at a particular moment in time along the continuous process of health. Developers of these systems should take care to explain the temporal uncertainty of these predictions. This could be achieved in the healthcare consumer space by explaining how predictions differ as symptoms and signs change over time or if new ones arise.

Second, developers must be aware of the myriad of possible user errors (i.e., erroneous inputs) in this space, including purposeful user errors, such as not wanting to admit an embarrassing symptom or condition, or accidental errors, such as not understanding a medical term or if a symptom or sign is important in this illness context.

Third, explanations must consider the full range of data points used to generate an output and clearly communicate to a user all information relevant to that result. This must include both information that led to that prediction as well as the data points that were considered but irrelevant to the final prediction. This way, individual judgement can validate whether or not the result is produced with all potential known and unknown inputs in mind for any given clinical situation. See, for example, AMA Journal of Ethics "Should Watson Be Consulted for a Second Opinion," which provided a hypothetical example of legal liability: “A particular medication recommendation regimen is recommended (by Watson and) used by physician, ignoring other contraindicating patient data because of the physician's assumption that Watson had evaluated that information. Was that info not included when producing a result? Was present but not in the form that made it relevant or useful? Or was it missing at the time of the result?”[2] If knowledge limits are not specified, it may not be possible for a user of the AI system to know if the result is complete or appropriate.

[2] Luxton, David D, Should Watson be Consulted for a Second Opinion? *AMA Journal of Ethics*. Volume 21, Issue 2, February 2019: https://journalofethics.ama-assn.org/article/should-watson-be-consulted-second-opinion/2019-02

The principle of Knowledge Limits also has profound implications on the topics of systematic bias and fairness. Just as humans are subject to behavior that is directed by unconscious bias, inequality in healthcare research and practice create systematic shortcomings in the form of an imbalanced data foundation that may not be identifiable or recognized by the AI system itself.

Training AI systems requires large amounts of data. Although current sources like electronic health records ("EHRs"), insurance claims data, and other private aggregated data are widely available, they can also be fragmented, inaccurate, incomplete, outdated, or otherwise biased. Even if investments are made in the creation of high-quality datasets, in the context of consumer healthcare, these gaps will likely remain. The long tail of medical data, organizational boundaries, and local shifts in disease trends can shift the dataset - as can insufficient data from underrepresented populations - such that explicit identification of knowledge limits for any given output would be required for the prospective generation and maintenance of unbiased results.

An AI system must be able to identify the knowledge limits of any output - the unknown unknowns of data quality that produce a result. In the consumer healthcare space, not doing so may yield heightened risk and a variety of downstream consequences. Therefore, we suggest that "fairness" be incorporated as an important element of outcomes, directly related to the principle of Knowledge Limits in an AI system. To do this, a proactive approach to ensuring that data sets encompass people of all races, genders, ideologies, and interests is necessary to create diversity of both the development and training pools for AI systems. Otherwise, in healthcare in particular, the known risks associated with bias and inequality may inadvertently train the AI to perpetuate such bias and inequality and, worse, affect the data or shift the resulting model toward unintended, discriminatory outputs on health outcomes for subgroups (e.g., age, ethnicity, sex, socioeconomic status, location, and genetics) who already experience social inequity. Worse still, the system's biases may be hidden by a seemingly reasonable output if the core fundamental approach to detecting bias in the historical data is not established and expected.

A baseline in fairness could guide proactive processes and output analysis according to particular population subgroups. It is especially important to conduct an analysis of an output and its explanation to understand the potential impact of bias as it relates to current clinical practice and to detect both the benefits and potential harm. We recommend these considerations be included in the discussion of the principle of Knowledge Limits.

Thank you for the opportunity to contribute to this important conversation.