Evaluating the Explainability and Interpretability of AI Systems

...

Dr. David Broniatowski, Dr. Valerie Reyna, Dr. Armin Vosoughi, Ms.Alaysha Shearn

Introduction

- Significant attention has been devoted to establishing credible methodologies for evaluating precursors of public trust in AI systems
  - National Artificial Intelligence Initiative Act, 2020 [1]
  - NIST's AI Risk Management Framework [2]

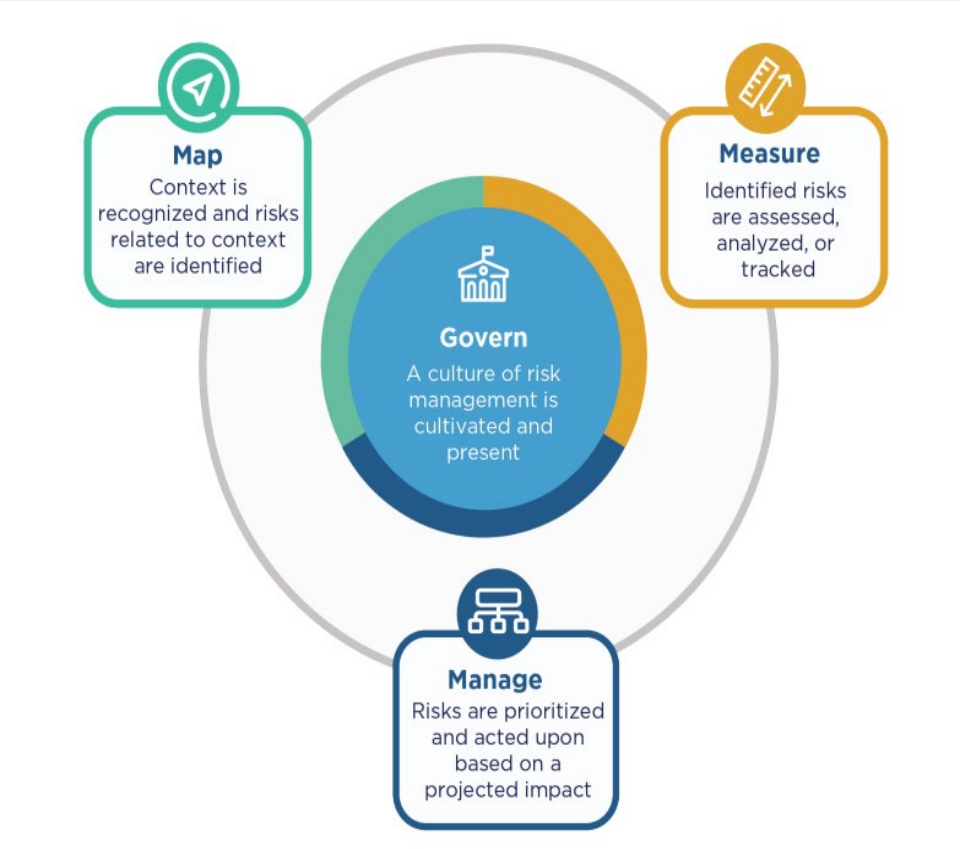

Figure 1. Functions organize AI risk management activities at their highest level to govern, map, measure, and manage AI risks. [2]

Introduction

- Model explainability and interpretability are two major precursors of trust that are often conflated.
- The purpose of this study is to fill the gap by providing psychometric evaluations of explanability and interpretability

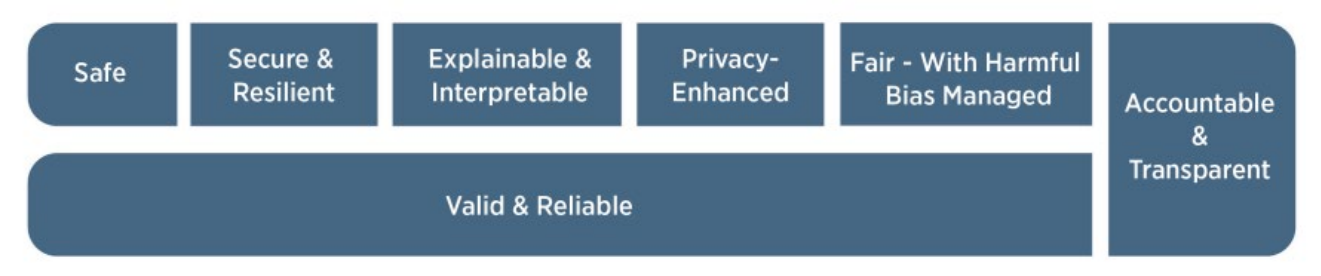

Figure 2. Characteristics of Trustworthy AI Systems [2]

Model explainability and interpretability must be understood to address public trust.

Model Interpretability

- Trust judgments of models stem from how people interpret model outputs in context.
  - The model output must “make sense” to them
- An intuitive judgment that we seek to formalize and capture.
- “If the system can explain its reasoning, we then can verify whether that reasoning is sound with respect to these auxiliary criteria.” [3]

Fuzzy Trace Theory (FTT) is a leading theory that provides guidance on how humans ma
sense of and use information to make decisions. [4]

Fuzzy Trace Theory (FTT)

- Humans encode stimuli (e.g., model output) into multiple mental representations in parallel
- These mental representations vary from one another in their level of precision
  - Gist : intuitive, informed, bottom-line meaning of stimuli in context.
  - Verbatim: literal, detailed representations of stimuli (e.g., raw, uninterpreted system output).
  - Humans prefer to rely on the least precise gist representation that makes meaningful distinctions when making a decision
  - Some people, especially novices who do not yet have well-formed gists, instead rely on verbatim representations.
- Verbatim representations can serve as inputs to algorithmic thinking
  - Computational systems are verbatim information processors.

Humans makes decisions based on the gist. By interpretability we mean the gist that the
extract.

Fuzzy Trace Theory Example

- One of the most basic gists is the difference between some and none [5]
- Assumptions:
  - 2 classes
  - Balanced training set
- Gists
  - K-nearest neighbor - none
    - 50% is like a coin toss
  - Naive Bayes - some
    - Accuracy is equivalent to 73.3% if you flip the class label
  - SVM - too good to be true
- A computer scientist would extract the gist
- A novice would consider the verbatim and conclude the Naive Bayes model is less predictive

Model	Accuracy	Precision	Recall	F_1 Score
Naive Bayes	0.267	0.563	0.897	0.692
Support Vector Machine	0.732	1.000	0.716	0.834
k-Nearest-Neighbor	0.524	0.834	0.637	0.722
Logistic Regression	0.907	0.859	0.617	0.718

Figure 3. Model Output Metrics [5]

Model Explainability

- Hypothesized to foster trust
- Centered on whether users can comprehend the mechanisms and rationales behind the model's output
  - Distinct from whether the model outputs themselves make sense
- Modern ML techniques are notorious for their lack of explainability, even for advanced computer scientists (e.g. "deep learning")
- "Shallow" learning models theorized to be inherently interpretable [6]
  - *Are they?*
  - ***Do people get the gist of inherently interpretable models?***

In computer science, explainability is the weighted contributions of predictor variables an
resulting algorithms.

Model Explainer Example

- SHapley Additive exPlanations (SHAP) returns importance scores for each feature, which are analogous to regression coefficients. [5]
- For a given prediction, SHAP scores indicate
  - the model's baseline value
  - the marginal contributions of each of its features
  - the final prediction.

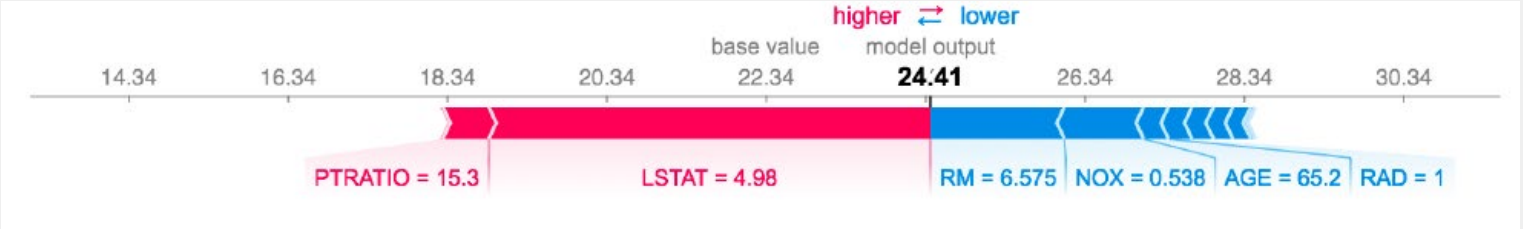

Figure 4. SHAP Output [5]

Conceptualization of interpretability and explainability

- Interpretability and explainability are distinct mental representations of a system that influence human judgments of user trust in an AI system
- Assessment of explainability and interpretability hasn't conventionally been conducted directly due to their reliance on the psychology of system attributes
    - Systematic evaluation of these aspects in AI systems has not been undertaken thus far
- We evaluate these factors using a theoretically motivated empirical assessment.

Objective: Create techniques and measures for evaluating the interpretability and explainability of ML systems.

Hypotheses

According to FTT, experts are more likely to take context into account when making prescription decisions whereas novices are more likely to follow verbatim model predictions. [4] Therefore we hypothesize:

1. Subjects' gist will be a significant factor predicting subjects' decisions.
    a. Controlling for gists, model output will have a smaller impact on subjects' decisions.

According to FTT reliance on gist is associated with individual differences. [5] However, there is evidence that mathematical ability may be distinct from mathematical confidence. Does this apply to machine learning models? We hypothesize that:

2. Self-assessments of model interpretability will be associated with numerical self-confidence in a manner that is statistically distinct from numerical ability. [7]

Hypotheses

Hypotheses	Our expectations	
	Experts	Novices
H1: Subjects' gist will be a significant factor predicting subjects' decisions.	Experts will be more likely to give the gist answer, and less likely to rely on the model and guidelines when they disagree with the gist. Experts will also be more likely to agree with the model and/or guidelines if they agree with the gist.	Novices will be more likely to rely on the model than experts, even when the model is wrong or even when it disagrees with guidelines.
H2: Self-assessments of model interpretability will be associated with numerical self-confidence in a manner that is statistically distinct from numerical ability.	Standard dual process theory predicts that subjects will rely more on the model if they're more numerate. FTT predicts that subjects who are more numerate will be better able to extract the gist of the model and will only rely on it when it's correct. Experts will rely less on the model when they receive the gist of it. They will rely less on CART model, in comparison to novices.	We expect that people will rely more on the model when they receive the gist compared to verbatim. The CART model is more interpretable than the logistic model so the same effect will be stronger for the CART model than for the logistic model. Therefore, we expect novices to rely more on CART model than the logistic model.

Individual Differences

- Differences in skill and personality traits among humans affect one's ability to interpret and understand model output. [5]
- Scales have been developed to capture the differences between subjects
  - Numeracy (subjective and objective) - mathematical ability [10-11]
  - Cognitive Reflection Test - intuitive judgments [12]
  - Need for Cognition - preference to exert mental effort [13]
  - Actively Open-minded Thinking - thinking style [14]

The Reference Study

- Models were obtained from a prior study
- The data is sourced from the Tan Tock Seng Hospital clinical database. The participants were selected as follows:
  - 21 years and above
  - Provided informed consent
  - Attended at TTSH ED for the first time with a primary diagnosis of uncomplicated URTI (ICD10-AM J00-J06) within 30 days
  - Discharged from hospital (not admitted)
- Survey, nasopharyngeal swabs for Polymerase Chain Reaction (PCR) tests, and C-reactive Protein (CRP) tests were given at discharge [8]

The Reference Study

- Created 3 predictive models to output the classification of a patient belonging to two groups
  - NABX: viral group; participant has a positive PCR and CRP of <20 mg/L OR Negative PCR and CRP of <5 mg/L
  - RABX: Any participant that is not in the NABX group
- Models
  - Logistic regression; LASSO Logistic Regression; CART (Classification and Regression Trees - This particular model is classification)
  - Probability Cutoffs are 0.6, 0.625 and 0.675 respectively
- Models were used to design an app the used the 3 predictions (viral vs possibly bacterial) to recommend whether a physician should not prescribe antibiotics or consider the use of antibiotics. [8]

Methodology

Initial studies focus on ML models to assist physicians in making decisions about prescribing antibiotics in the emergency room. This setting is especially important because it captures real-world conflicting requirements in a setting of deep uncertainty.

- Subjects will be medical students, residents and attendings.
- Explanations
  - Subjects receive a gist explanation about 1 of 2 “shallow” learning models (Logistic Regression and CART) from either a computer science or clinical perspective.
- Scenarios
  - Hypothetical patients in an emergency room setting are described to the subject.

Explanation Example - Verbatim

Logistic Regression

The following is the equation for logistic regression:

$$logit(p) = ln\left(\frac{p}{(1-p)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k$$

where,

p = probability of the occurrence of the outcome

$X_1, X_2, \ldots, X_k$= set of input features of X

$b_1, b_2, \ldots b_k$= parameter values to be estimated in the logistic regression formula

This equation is used to predict the probability of a patient having a viral illness, which translates to a recommendation to not prescribe antibiotics. A logistic regression model was trained on patient data to determine the values of b1 through bk. The resulting equation is:

$$logit(p) = 21.21 - 0.019\,X_1 + 1.343\,X_2 - 0.607\,X_3 + 1.343\,X_4 - 0.456\,X_5 - 0.029\,X_6$$

The model uses a 0.6 probability cutoff. If the value of *p* is greater than or equal to the probability cutoff, then the model determines that the patient falls into the NABX category. If the value of *p* is less than the probability cutoff, then the model determines that the patient falls into the RABX category. NABX are patients with a respiratory virus detected via Polymerase Chain Reaction (PCR) and C-reactive Protein (CRP) < 20 mg/L or patients who did not have a respiratory virus detected via PCR and CRP ≤ 5 mg/L. Patients who did not fall into these 2 categories were assigned to the RABX group.

Scenario Example

A 39-year-old woman comes to the ER complaining of a productive cough, sore throat, fever, sweating, shaking chills, low energy, and malaise (i.e., overall discomfort and lack of well-being) that started 9 days ago. She reports lack of appetite and a bad taste in the back of her mouth. Symptoms have not improved over the 9 days. She reports fever at home of 37.8C (100F). She is a nonsmoker with no known co-morbidities and no recent international travel. She reports a family history of cardiovascular disease. Patient has not been vaccinated against influenza.

Examination of the throat shows erythematous (i.e., red) pharynx without exudate (i.e., without thick, pus-like substance on the surface of the tonsils). Vital signs: blood pressure of 120/70, temperature of 37.8C (100F), pulse rate of 75, oxygen saturation of 94%. Examination of eyes, ears, lungs, and heart are otherwise normal.

Methodology - Measures

- Self-report measures
  - Self-assessments of interpretation and explanation
  - General and easy to collect.
  - Cannot indicate whether a misinterpretation or incorrect explanation has occurred.

Explainability Cronbach's α = 0.90	The system explained to me how I got a particular prediction
	The system did NOT explain to me how I got a particular prediction
	I can explain the reasoning behind the predictions the model made
	I CANNOT explain the reasoning behind the predictions the model made
	The system allowed me to see how it made predictions
	The system did NOT allow me to see how it made predictions
Interpretability Cronbach's α = 0.94	I can explain what the system's results mean
	I CANNOT explain what these results mean
	I can make sense of what the system's results are saying
	I can NOT make sense of what the system's results are saying
	The system's results make sense to me
	The system's results do NOT make sense to me

Methodology - Measures

- Observations of human decisions upon receiving model output
  - Humans will be more likely to make decisions that agree with a model's when it is interpretable and explainable.

Logistic Regression

Feeding the data from the scenario into the model results in the following calculations where Z is the liner combination of predictor variables, and p is the probability:

Z = 21.21 - 0.019 (age) + 1.343 (giddiness) - 0.607 (fever) + 1.343 (SOB) - 0.456 (BT) - 0.029 (PR)

Z = 21.21 - 0.019 (40) + 1.343 (0) - 0.607 (1) + 1.343 (0) - 0.456 (37.5) - 0.029 (70)

Z = 0.713

p = (e^Z)/ (1 + (e^Z))

p = (e^0.713)/(1+(e^0.713)

p = 0.671

The model predicted NABX because the resulting probability was greater than the 0.6 cutoff. Will you prescribe antibiotics?

Reminder: NABX are patients with a respiratory virus detected via Polymerase Chain Reaction (PCR) and C-reactive Protein (CRP) < 20 mg/L or patients who did not have a respiratory virus detected via PCR and CRP ≤ 5 mg/L. Patients who did not fall into these 2 categories were assigned to the RABX group.

Methodology - Measures

- Gist endorsement
  - Assess humans' tendencies to rely on meaningful gist representations of model input, operations, and output.
  - Measure humans' agreements with these gists, to judge whether agreement will mediate the relationship between model artifacts (such as outputs) and human decisions.

	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree not Disagree	Somewhat Agree	Agree	Strongly Agree
It can't hurt the patient to take antibiotics.	○	○	○	○	○	○	○
It could hurt the patient to take antibiotics.	○	○	○	○	○	○	○
Antibiotics might not make the patient better, but it is better to be safe than sorry (so he/she should take them).	○	○	○	○	○	○	○

For more information on Likert Scale items, see reference [9].

Design

- 2 (model = CART and Logistic) X 2 (scenarios = guideline yes for NABX and guideline no for NABX) X 2 (scenarios = gist physician yes for NABX and gist physician no for NABX) X 2 (scenarios = model yes for NABX and model no for NABX) X Instructional Group (verbatim, computer science gist, clinical gist, and control group which is no model)
- Subjects are randomly assigned into 1 of 7 conditions:
  - Model (CART and Logistic) X Instructional Group (verbatim, computer science gist, clinical gist, and control group which is no model)
    - 8 scenarios are shown in random order to the subjects

Design

- Between subject variables: participants are only assigned to one level of each factor
  - Model
  - Explanation type
- Within subject variables: participants are assigned to every level of each factor
  - Scenarios
- Subjects will first see the model explanation (if not in the control group). Then they will see the scenarios. Each scenario will be followed by questions. Presentation of questions will be randomized.

Expected Results

- Experienced people will rely more on the gist as opposed to novices relying on the verbatim.
- The model outcome will have a smaller effect on the experts that rely on gist.
- Those who have a lower understanding of gist will be more likely to defer to the model, even if the model output is incorrect.
- More numerate people will be able to understand the verbatim and extract the gist from it.

Conclusion

- We theorize that explainability and interpretability are two distinct measures.
  - Must be measured to foster trust in AI
- We hope to fill the measurement gap through inclusion of psychometric evaluations.
- Next steps
  - Finalize scenarios
  - Finalize survey
  - Pilot the study
- Any interested parties should contact Dr. David Broniatowski via email (broniatowski@gwu.edu).

References

[1] National Artificial Intelligence Initiative Act of 2020.

[2] Gina M. Raimondo N. Laurie E. Locascio (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0).

[3] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[4] Reyna, V. (2018). When irrational biases are smart: A fuzzy-trace theory of complex decision making. Journal of Intelligence, 6(2):29.

[5] Broniatowski, D. A. (2021). Psychological foundations of explainability and interpretability in artificial intelligence. NIST, Tech. Rep.

[6] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. arXiv:1811.10154 [cs, stat]. arXiv: 1811.10154.

[7] Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., and Dickert, S. (2006). Numeracy and decision making. Psychological Science, 17(5):407–413.

[8] Wong, J. G., Aung, A. H., Lian, W., Lye, D. C., Ooi, C. K., & Chow, A. (2020). Risk prediction models to guide antibiotic prescribing: a study on adult patients with uncomplicated upper respiratory tract infections in an emergency department. Antimicrobial Resistance & Infection Control, 9(1), 1-11. https://aricjournal.biomedcentral.com/articles/10.1186/s13756-020-00825-3

[9] Broniatowski D. A., Klein E. Y., May L., Martinez E. M., Ware C., and Reyna V. F. (2018). Patients' and Clinicians' Perceptions of Antibiotic Prescribing for Upper Respiratory Infections in the Acute Care Setting," Medical Decision Making, vol. 38, no. 5, pp. 547–561. doi: 10.1177/0272989X18770664.

References

[10] Fagerlin, A. et al. Measuring numeracy without a math test: development of the subjective numeracy scale. Med. Decis. Mak. 27, 672–680 (2007).

[11] Lipkus, I. M., Samsa, G. & Rimer, B. K. General performance on a numeracy scale among highly educated samples. Med. Decis. Mak. 21, 37–44 (2001).

[12] Thomson, K., & Oppenheimer, D. (2016). Investigating an alternate form of the cognitive reflection test. Judgment and Decision Making, 11(1), 99-113. doi:10.1017/S1930297500007622

[13] Cacioppo J. T., Petty R. E. & Kao C. F. (1984) The Efficient Assessment of Need for Cognition, Journal of Personality Assessment, 48:3, 306-307, DOI: 10.1207/s15327752jpa4803_13

[14] Stanovich K. E., West R. F., Toplak M. E. The Rationality Quotient: Toward a Test of Rational Thinking. MIT Press; Cambridge: 2016