Comments in response to the National Institute of Standards and Technology Request for Information on Developing a Federal AI Standards Engagement Plan
[Docket Number: [190312229–9229–01]]
June 7, 2019

Submitted by:
David Broniatowski, Ph.D. and Aylin Caliskan, Ph.D., The George Washington University
Valerie Reyna, Ph.D., Cornell University
Reva Schwartz, Parenthetic, LLC

Today's crisis of trust in artificial intelligence (AI) stems from a perception on the part of operators that they cannot explain why a given result was generated. This provokes anxiety on the part of its users – they often feel that there is no way to tell what the algorithm might do next, requiring that they exhibit "blind faith" in the algorithm and its designers. This perception has been especially associated with so-called "deep learning" neural networks, because of the internal complexity of their operations. Although, in theory, it might be possible to describe all of the factors that lead an algorithm to a given conclusion, doing so would involve communicating a level of detail that would be difficult for most humans to follow.

In order for algorithms to be trusted, they must be *explainable* – i.e., it must be possible for a human to make sense of why an algorithm did what it did, and, in some cases to explain it to others. Importantly, this is not the same as simply repeating back the logic used by the algorithm to make its decisions; rather, a human typically seeks to understand the *function* that the algorithm was designed to carry out, and whether that function was executed adequately. The algorithm must also be *unbiased* – i.e., it must reflect the values of the user, such as by making decisions based only on information that would not be considered discriminatory.

Whereas an algorithm blindly draws conclusions based upon computer code and input data, humans seek to *contextualize* the algorithm's output, attributing meaning to it, which helps us to determine whether it is relevant and valid. Rather than seeking to answer *how* an algorithm achieved a given output, most humans seek to understand what the generalizable causal principles are and whether the algorithm has achieved larger strategic goals.

This distinction between the *verbatim*, detailed description of how an algorithm works, and the bottom-line meaning, or *gist*, of why it generated a given output is captured by Fuzzy-Trace Theory – a leading account of how humans process technical, and especially numeric, information, and how this differs from algorithmic logic. According to Fuzzy-Trace Theory, humans encode numerical data at multiple levels of mental representation ranging from precise, yet decontextualized, verbatim representations of the stimulus to categorical, yet contextualized, gist representations that help humans to make meaningful distinctions. Most people prefer to rely on gist representations although this varies with individual differences. For example, individuals who are numerate – i.e., possessing high mathematical literacy – have the ability to carry out complex calculations so that they can verify the output of some algorithms. Individuals who have a high Need for Cognition have the desire to do so. Individuals with these traits have been shown

to prefer to rely more on verbatim representations in specific experimental contexts in which technical accuracy (as opposed to gist-based meaning) is emphasized.

Consider a system designed to make hiring recommendations for a large company. A typical machine learning approach would entail encoding data from several thousand résumés that had been annotated as either successful or unsuccessful prior hires. These data might include any number of features, such as level of education, years of prior work experience, requested salary, etc. The algorithm could then be trained to identify those factors that best predict prior hiring. Having identified these factors, the algorithm would then be used to make predictions regarding whether a given candidate might be successfully hired, with the top 10 candidates most likely to succeed being given job offers.

Suppose a support vector machine were used to perform this classification. "Explaining" these algorithms at the verbatim level might entail telling unsuccessful job candidates their location in a high-dimensional space, explaining to them that the algorithm partitions this space based on what is mathematically optimal, and then concluding that their application was "closer" to prior unsuccessful candidates than prior successful candidates, putting them on the "wrong side" of the dividing line. This kind of explanation would lead the candidate to be legitimately confused, and potentially frustrated.

In contrast, a gist explanation of this algorithm's operation would require identifying the bottom-line meaning of the algorithm's operations. Presuming that all of the features in the model were measures of a candidate's qualifications, one could explain to candidates that they did not "make the cut" when their qualifications were assessed against what the company's needs were. A graphical representation communicating this gist could be readily generated if the algorithm were used to generate a representation of the "ideal" candidate profile, such as by displaying the average feature values for points in the training set, and displaying comparisons to the candidate along each of these dimensions – thus allowing the candidate to see what might need to be improved for another submission. Similarly, the nearest point on the separating hyperplane could be chosen as an example of a candidate that would just barely "make the cut." Such comparisons should be designed to clearly communicate to the candidate a sense of why they were not considered.

The above description presumes that the model used for selecting candidates has valid measures of meaningful constructs. However, problems arise when the features included in the model are not valid. For example, if one were to include race and gender as features, one could no longer claim that the candidate's qualifications were the basis for the algorithm's operations. Although the verbatim explanation would the same, the gist would be quite different – the model could be violating workplace discrimination practices. Furthermore, discriminatory features need not be explicitly included to introduce bias – rather, proxies of these protected attributes may introduce bias by interacting with the type of algorithm and training data used. For example, resumes containing linguistic data may be fed into neural nets with word embeddings, which are known to contain implicit biases based on culturally-embedded assumptions present in training data. For example, an algorithm given a choice between a man and a woman with equivalent qualifications applying to be a nurse would likely choose the female candidate since her name signals her gender and, historically, nurses have tended to be female Similarly, one's zipcode on the resume

signals race which may correlate with historical injustices, thus perpetuating racial discrimination.

The above example demonstrates the need for an explicit focus on how human psychology interacts with algorithmic logic. This program would have three areas of focus:

1) Standards of algorithm validity: It is common for machine learning algorithms to include many possible features. Do the features included actually measure what the user intends them to measure? Are these features valid given the context of other information and do they extrapolate to the intended situations of use?
2) Standards for gist communication: Given that a model has valid, unbiased, inputs, how should the outputs be communicated in a manner that emphasizes their bottom-line meanings? How do these vary with standard metrics, such as precision, recall, F-score, and so on.
3) Standards of algorithmic ethics: Human input is required to ensure that the features included are not used in a way that violate ethical mores.