Subject:	Fwd: FW: NIST report on explainable AI
Date:	Saturday, September 26, 2020 at 2:55:57 PM Eastern Daylight Time
From:	Stephanie Tsuei
To:	NIST Explainable AI
Attachments:	image001.png

Dear NIST reviewers,

I am a graduate student at UCLA studying uncertainty quantification and management in robotics and an employee at Northrop Grumman Aeronautics Systems working on problems related to Verification and Validation in autonomous systems. I have a long, open-ended comment on the document "Four Principles of Explainable Artificial Intelligence" that doesn't quite fit into your form. Thank you very much for your time.

Comment:
In the definition of knowledge limits (lines 169-170), "The system only operates under conditions for which it was designed or when the system reaches a sufficient confidence in its output." please consider replacing "or" with "and", especially in the context of safety-critical systems. All machine learning theory and practice assume that test samples are drawn from the same distribution as training samples. Of course, it is possible for a machine learning algorithm to coincidentally perform well on a dataset drawn from a different distribution than it was trained. These coincidences, however, should require extra testing, just as an aircraft using commercial-off-the-shelf parts in ways those parts were not designed requires extra testing for certification.

Next, for the principle of knowledge limits, I urge NIST to consider the topic of uncertainty quantification. The field of uncertainty quantification originated in computational fluid dynamics and materials science, two fields that rely on complex mathematical models that are not able to fully explain physical phenomena observed in both laboratory and real-world settings. In recent years, researchers have applied fundamental ideas from uncertainty quantification to neural networks. Some relevant literature is given below, but they are not 100% comprehensive.

- Guo et. al [1] show that the output of a softmax function, often taken as a confidence score in classifiers, is well-calibrated in older, less-nonlinear neural network architectures, but is incredibly uncalibrated in modern networks, such as the popular Resnet. If a classifier is well-calibrated, then exactly 10% of classifications that it assigns a 90% confidence are incorrectly classified. A figure from the paper shows, however, that only 60% of classifications assigned a 90% confidence score are correct. The authors present methods to scale the softmax output without affecting the training process or the classification accuracy so that confidence scores are better calibrated.
- Galramini et. al [2] describe a method specifically for measuring epistemic uncertainty, or uncertainty as a result of not having enough data, in Bayesian networks. They present a method that applies to both classifiers and regressors.
- Lakshminarayanan et al [3] show that out-of-distribution inputs generally have lower confidence scores than in-distribution inputs, but it doesn't mean that the outputs are well-calibrated.
- Lee et al [4] present methods to detect out-of-distribution inputs

References:

[1] Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 1321–1330. ICML’17. JMLR.org, 2017. http://dl.acm.org/citation.cfm?id=3305381.3305518.

[2] Gal, Yarin, and Zoubin Ghahramani. “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.” In *International Conference on Machine Learning*, 1050–59, 2016. http://proceedings.mlr.press/v48/gal16.html.

[3] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. “Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles.” In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 6402–6413. Curran Associates, Inc., 2017. http://papers.nips.cc/paper/7219-simple-and-scalable-

predictive-uncertainty-estimation-using-deep-ensembles.pdf.
[4] Lee, Kimin, Honglak Lee, Kibok Lee, and Jinwoo Shin. “Training Confidence-Calibrated Classifiers for Detecting Out-of-Distribution Samples,” February 15, 2018. https://openreview.net/forum?id=ryiAv2xAZ.

Best,
Stephanie

---------- Forwarded message ---------
From: **Tsuei, Stephanie [US] (AS)**
Date: Mon, Aug 31, 2020 at 8:23 AM
Subject: FW: NIST report on explainable AI
To:

From: Cook, Stephen P [US] (AS)
Sent: Thursday, August 20, 2020 3:55 PM
To: Milam, Mark B [US] (AS); Hudson, Hunter [US] (AS) ; Tsuei, Stephanie [US] (AS) ; Sarathy, Prakash [US] (AS) ; Plawecki, Daniel W [US] (AS)
Subject: NIST report on explainable AI

NIST report out for comment, fysa…”Four Principles of Explainable Artificial Intelligence” attached

STEVE COOK | NG Fellow, Airworthiness

Northrop Grumman | Aeronautics

Airworthiness and Airspace Integration Office

O: | C: |

-

NORTHROP GRUMMAN