Research Review 2018

Explainable AI and Human Computer Interaction

Dr. April Galyardt
Machine Learning Research Scientist
CERT Division

Carnegie Mellon University
Software Engineering Institute

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

DM18-1239

A brain teaser

I will sell you this rock for $200.

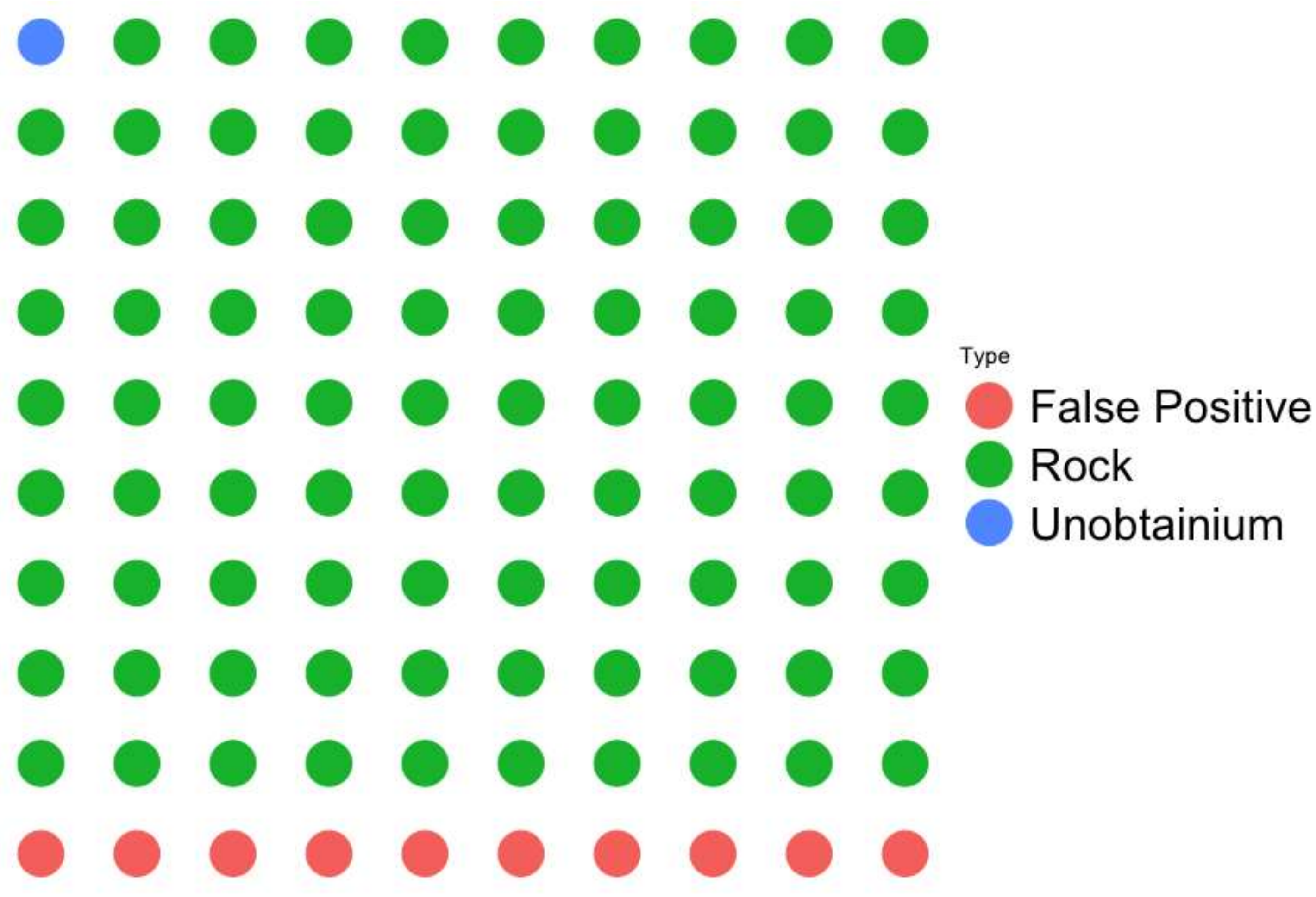
Will you buy it?

The detector says this is a piece of unobtainium.

If the detector is correct, it’s worth **$1000.**

https://ed.ted.com/lessons/can-you-solve-the-false-positive-riddle-alex-gendler

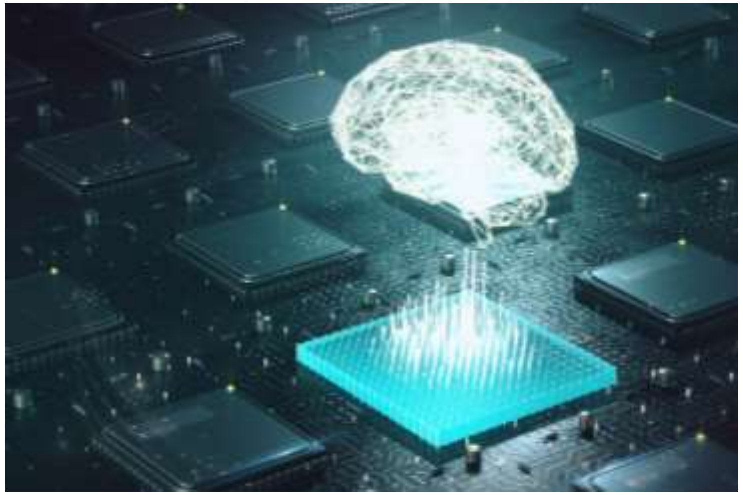
A brain teaser



There’s about a 1/11 chance that this rock is unobtainium.

Implications

Screening at the airport



An AI is predicting who is a threat.

The predictor says this person is a threat.

What's the probability they're actually a threat?

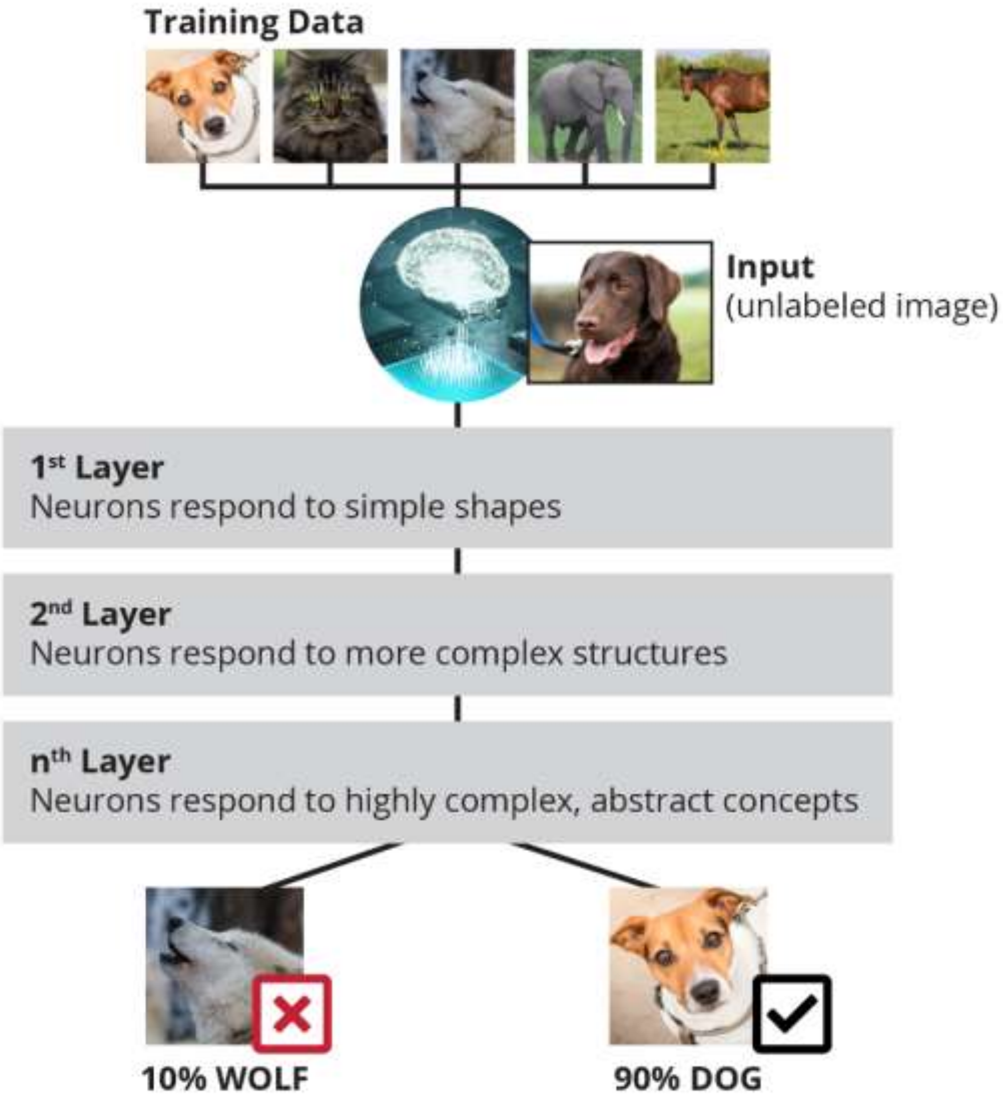
https://ed.ted.com/lessons/can-you-solve-the-false-positive-riddle-alex-gendler

Longstanding Science

Making ***probabilistic judgements*** is hard.

Depending on how probabilities ***are presented*** people make different choices.

Implications for Explainable AI



Training Data
Input
(unlabeled image)
1st Layer
Neurons respond to simple shapes
2nd Layer
Neurons respond to more complex structures
nth Layer
Neurons respond to highly complex, abstract concepts
10% WOLF
90% DOG

Current work in explainable AI is focused on ***providing probabilities*** to the end user.

That’s not enough.

The human-computer-interaction must provide support to help the user interpret those probabilities appropriately.