Eggsplaining AI:
An Analogy for How Machine Learning Systems Work (and Don't Work)

R.R. Hoffman's Concept Blog #6
April 2021

rhoffman@ihmc.us

The role of analogy in sensemaking and explanation has been widely discussed and researched. Analogical reasoning is a fundamental strategy in scientific thinking, for example (atoms are like solar systems). It seems curious therefore that the role of analogy in explaining AI systems seems to have been under-played.

An individual whose role was that of a system integrator commented to me:

> *I've seen a wide gamut on the explanation. I got a great in-depth explanation of one system, how it works. Relied on analogies. System fills seats in a school bus with what it believes; the more input it gets the more refined it gets in putting things into the seats. What was going on under the hood was not visible to you. [It was] easy to interface with the display and understand what it was doing, without understanding what was going on under the hood.*

This analogy is incompletely expressed here, but it got me to thinking.

Imagine a robot that is to be trained to work at a dairy farm, sorting eggs. A tedious job, a good thing for automation to do for us.

We want the robot to identify good eggs and place them in cartons, and identify cracked or bad eggs and put them in a waste can.

First we have to teach the robot what a "good egg" is. It learns by using feedback to create a filter, like the way a spaghetti strainer lets out the water while retaining the noodles.

We place some good eggs and some cracked eggs on a conveyor belt, and have the eggs pass in front of the robot's camera eye.

When a good egg is in front of the robot, we press the "GOOD" button.

When a cracked egg is in front of the robot we push the ""BAD" button.

The robot processes the camera images through a strainer, which it constantly improves based on feedback.

After teaching it using 50 eggs (25 GOOD, 25 BAD) we get tired, so we go ahead and we test it with a new sample.

The robot does great! It correctly identifies 95% of the cracked eggs, which it places in the waste can for the BAD eggs.

So it made a decent strainer. "Hoo-Ray!" we declare. "We've taught it to recognize bad eggs!"

So, we put the robot to work.

At the end of its first day, the farmer comes to us and says, "Hey, what did you do with that contraption? It's putting brown eggs in with the white eggs! We put brown ones in separate cartons!

Oh. We did not know that the robot would have to have a strainer to separate GOOD eggs from BAD eggs, and a second strainer on top of that to separate the WHITE eggs from the BROWN eggs.

So we train the robot with a sample of good eggs and brown eggs, pushing the "BROWN" button every time a brown egg comes before the camera.

Again, we test the robot on a new sample, and again it does great! 95% of the brown eggs are placed in the containers for the BROWN eggs, and 95% of the good white eggs are placed into the containers for the GOOD WHITE eggs.

At the end of the second day, the farmer comes to us and says, "Hey, so *now* what did you guys to with that contraption? It's putting cracked brown eggs in with the good brown eggs!"

Oops. So again we train it to make another strainer enabling it to put BAD+BROWN eggs into the BAD waste can, along with the BAD+WHITE eggs.

Then we test it on another sample, and again it does great! 95% of the BAD+BROWN eggs are placed into the waste can.

At the end of the third day, the farmer comes to us and says "OK, *now* what did you guys to with that contraption? It's putting quail eggs in with the good brown eggs!"

It seems that a flock of quails had set up shop in the chicken coop.

So again we train it to make another strainer enabling it to identify Quails' eggs so that they could be put into the waste can.

On a test sample, this time it does not do so well.

So we ask ourselves, "What's the difference? Well, quail eggs are round, and smaller than hens' eggs. Also they can be speckled and blotchy. And the blotches on quail eggs can be ... Oops!... Brown!

We wonder whether the robot is deciding that the quail eggs with brown blotches are really just BROWN (good) eggs. Eek! What if a cracked, brown-splotchy quail egg passes under the robot's eye?

We also wonder about something much simpler. Quail eggs are smaller than chicken eggs, and they are round.

Why the robot did not learn at the beginning that chicken eggs have a particular size and shape? After all, we trained it on chicken eggs!

We have to train the robot to recognize chicken eggs by their asymmetrical shape. They are a sort of tapered oval.

Now to train the robot we have to place eggs on the conveyor belt making sure that their asymmetry is "face-up" and the robot camera can see that the GOOD eggs are not round.

And so it goes. For each "glitch" we have to train the robot to make yet another strainer to sort the GOOD (not cracked, not round, not brown, not splotchy) eggs from the BAD (white or brown or quail) eggs.

And then it dawns on us. The robot might be doing well (say 95% correct) but it really has no concept of an egg at all! All it filters out are shapes of particular kinds, having a white surface, no little lines on it (cracks), and no brown coloring. It associates those with either a carton or the trash can.

We worry what might happen if a blue Robin's egg makes it into the mix. Or a golf ball. Or an unbounded number of other things. Sure, that would be rare (although there is a golf course nearby), but still . . . .

And then we pause again. Remembering that very first test . . . where it recognized 95% of the BAD eggs. So what was it about the 5% that it got *wrong*? Is it seeing something we don't see? Does it really "see" at all?

Acknowledgement and Disclaimer

This material is approved for public release. Distribution is unlimited. This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under agreement number FA8650-17-2-7711. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL or the U.S. Government.