When AI Goes Wrong and How to Fix it Fast

Daniel Jeffries

Chief Technical Evangelist

Pachyderm

@dan_jeffries1 - *medium.com/@dan.jeffries* - *practical-ai-ethics.org*

With algorithms making more and more decisions in our lives how do we know we can trust those decisions?

Nightmare Examples of AI Gone Wrong

1. Google Photos label people of color as “gorillas.”

2. Uber self-driving car hits a woman in AZ and kills her.

3. Stop signs with stickers seen as 45 MPH signs by visual classifier.

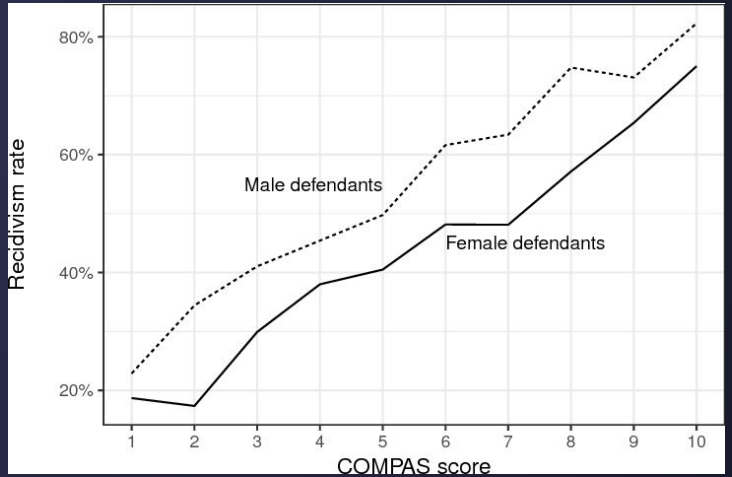
Source: Robust Physical-World Attacks on Deep Learning Visual Classification.

Subtler Problems

But PR disaster problems are just the tip of the iceberg. The smaller problems are much more subtle and harder to see.

1. COMPAS recidivism scores in Florida help judges decide who goes to jail.

2. A loan decision algorithm that leaves out women who could pay back those loans because they were historically left out in the dataset.



How Do We Fix It?

Explainable AI could help but it's not there yet. It's still cutting edge research.

In many ways we're still establishing what "explainable" even means. (*Towards a Rigorous Science of Interpretable Machine Learning*)

How Do Most Orgs Try to Fix It?

Form an AI Ethics committee.

The group meets, talks and eventually puts out a report with about how AI should be "inclusive and fair.” Sounds great but it's ultimately not implementable because those are just abstract concepts and platitudes, not actionable steps.

So nothing gets done, nothing changes.

The Middle Path

But we don't need explainable or an ethics committee to do a lot better with our AI/ML decision making.

Building an AI Anomaly Response team:

We’re going to create two teams.

- **Customer/Public facing team**
- **QA for AI**

Customer/Public Team

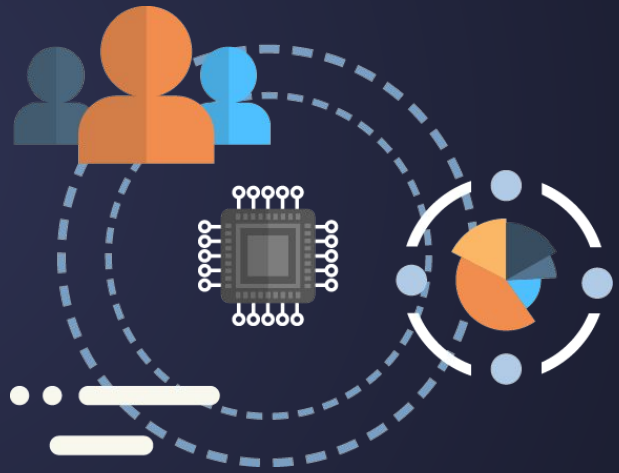
The customer and public team is trained to respond to upset customers or the general public when AI goes wrong.

They're taught to understand how algorithms make decisions vs humans. They have ready made templates to talk to the press with clear, concise language.

QA for AI Team

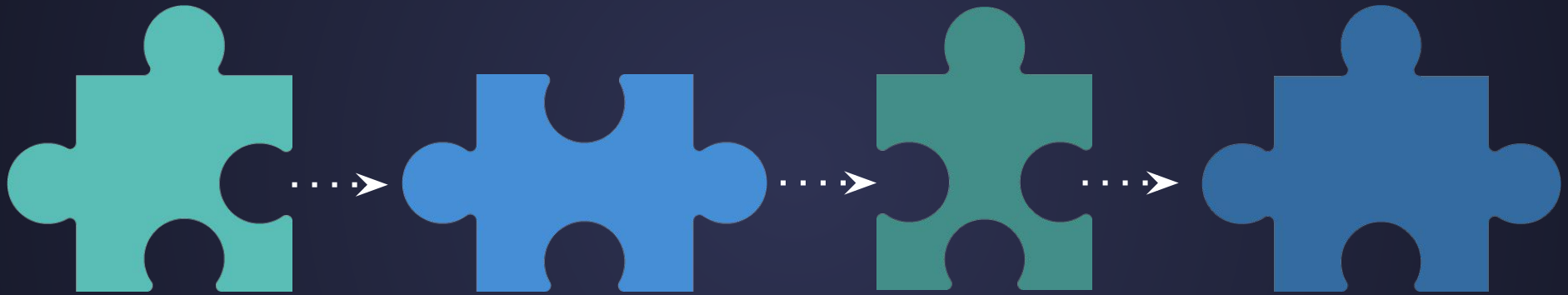
The QA for AI team is a group of coders, engineers and testers. They're responsible for coming up with both triage solutions and long term solutions.

In the Google example, they stopped the AI from being able to label anything as "gorillas." That's actually an effective stop-gap solution but that's not the end of the solution. Once the bleeding is stopped, they must move on to a real, long-term solution.



Adopting IT and Coding Best Practices to AI

The QA for AI team adapts best practices in IT and code development and apply to data and AI/ML.



Version Control / Data and Metadata Management

Snapshots / Rollback

CI/CD for AI/ML

- Unit and integration tests
- Dev --> Staging --> Prod

Auditing and Logging

- Forensic analysis

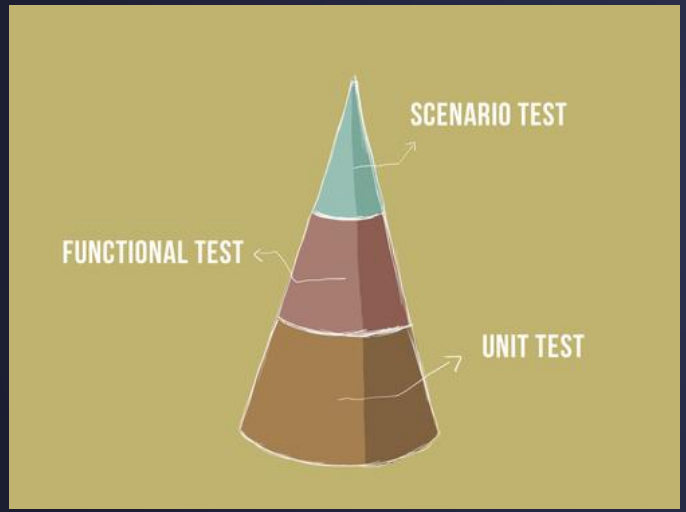
Creative Problem Solving

Long term solutions are harder. It requires creative / critical thinking and coordination across BUs.

Potential Solutions

- Do they need to expand the data set?
- Create a synthetic data set?
- Buy a new one?
- Can they layer a rule based system with a business rules engine such as Drools next to it and combined the black box AI and the rules engine into a weighted score to make a decision?
- Create a GAN to attack the data set and test it further?

Test / Test / Test



Once the AI QA team comes up with a solution they need to retrain the model and employ another IT best practice, testing. whether that's unit, integration, or regression testing. Not enough to just look at model accuracy scores.

Example Solutions

- Test stops signs with various occlusions on them.
- Auto-generate synthetic female loan candidates with various characteristics and test whether they get approved or denied.

A Universal Problem

Every organization will face these challenges in the coming decade. No industry, person or organization will be untouched by the power of algorithmic decision making. We need a way to audit and control the decisions we make.

Research to make AI explainable will continue to accelerate. As it does we'll integrate those kinds of tools into our pipelines.

But we don't have to wait until tomorrow. We can design a better monitoring and management process today with the tools and best practices we've already honed over decades of IT development.

What Does the Future Hold?

1. AI monitoring AI

2. Everything open
   a. “If it's not open don't let it think for you.”
      - Daniel Riek, Director RH AI Center of Innovation
   b. Data sets, algos, models, code

3. Contextual AI

4. Explainable AI

Thank You

Pachyderm.com - @dan_jeffries1 - medium.com/@dan.jeffries - practical-ai-ethics.org