Bias and Quality Measures in ML: Case Studies

David Rimshnick

February 4, 2019

CORNELL TECH

Recognize this?

AlphaZero

- Developed by Google subsidiary DeepMind
- Started with just the rules of chess, played against itself millions of times
- Used deep learning (CNNs) to evaluate board positions
- In ~4 hrs, defeated previously strongest chess engine
- Overcame conventional wisdom / biases: Material is not as important as position

Silver, David, et al. "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm." *arXiv preprint arXiv:1712.01815* (2017).

Less Wrong 1: Absolute Denial Macro

- Post on rationality AI discussion board "Less Wrong" – "What's the Strangest Thing An AI Could Tell You?"
- Anosognosia – condition of not knowing own disabilities
- Example: People whose left arm doesn't work will come up with any excuse they can to explain why they can't use it
- Author calls this "Absolute Denial Macro"

Eleizer Yudkowsky, http://lesswrong.com/lw/12s/the_strangest_thing_an_ai_could_tell_you/?sort=top

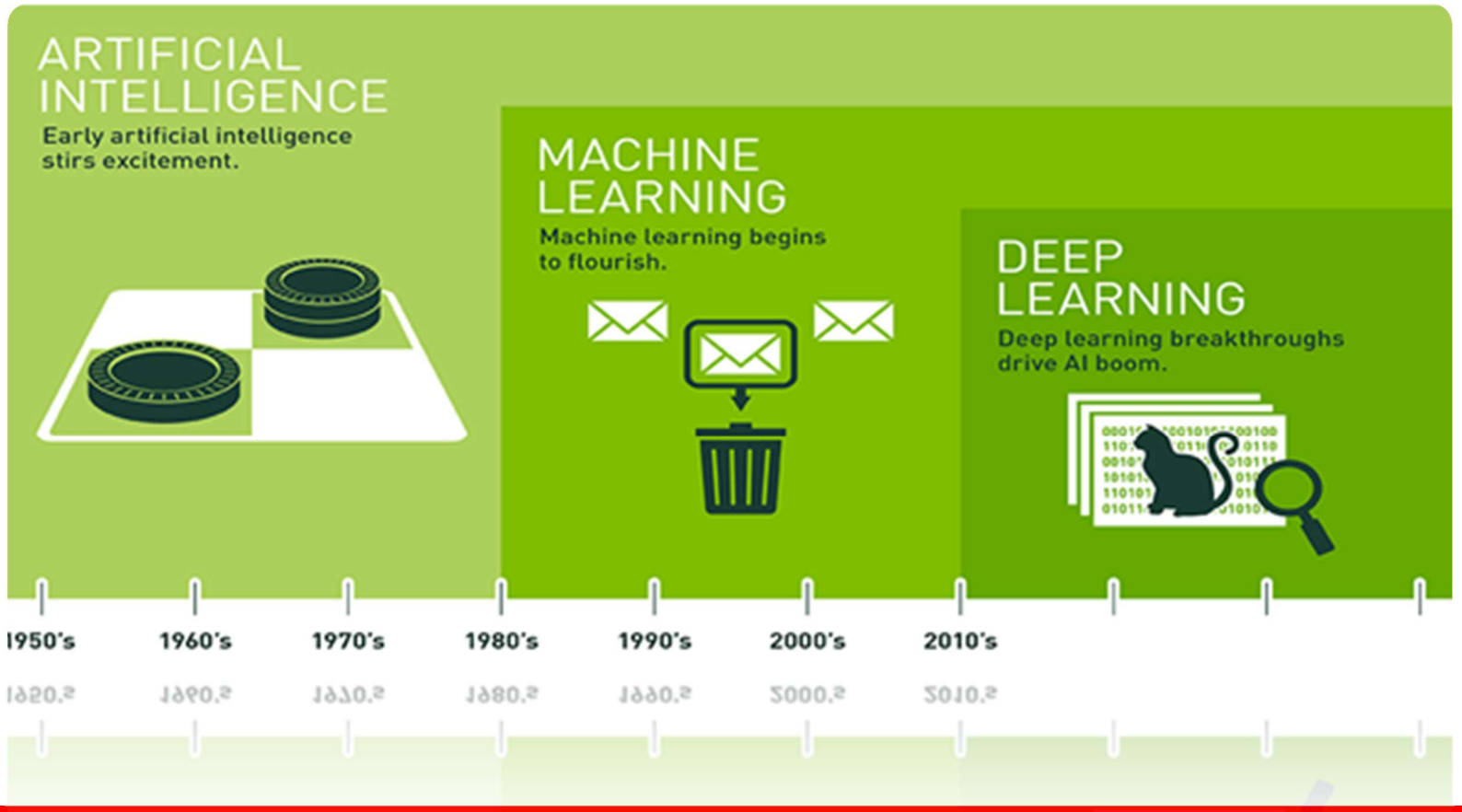
Less Wrong 2: What are our hidden biases?

Now, suppose you built an AI. You wrote the source code yourself, and so far as you can tell by inspecting the AI's thought processes, it has no equivalent of the "absolute denial macro" - there's no training that could inflict on it the equivalent of anosognosia. It has redundant differently-architected systems, defending in depth against cognitive errors. If one system makes a mistake, two others will catch it. The AI has no functionality at all for deliberate rationalization, let alone the doublethink and denial-of-denial that characterizes anosognosics or humans thinking about politics. Inspecting the AI's thought processes seems to show that, in accordance with your design, the AI has no intention to deceive you, and an explicit goal of telling you the truth. And in your experience so far, the AI has been, inhumanly, well-calibrated; the AI has assigned 99% certainty on a couple of hundred occasions, and been wrong exactly twice that you know of.

- Idea: AI you've programmed to be completely objective – what could it tell you about the world that you were blatantly ignorant of?
- In other words, Can AI help us uncover our hidden biases?

Eleizer Yudkowsky, http://lesswrong.com/lw/12s/the_strangest_thing_an_ai_could_tell_you/?sort=top

Deep learning learns what to learn



Case Study: Using Machine Learning to understand biases in promotional decision making in pharma

- How are we prioritizing promotional activity?
  - What features are we using to determine most fruitful targets? How to allocate effort against them?
- What features does an objective AI think are most important?

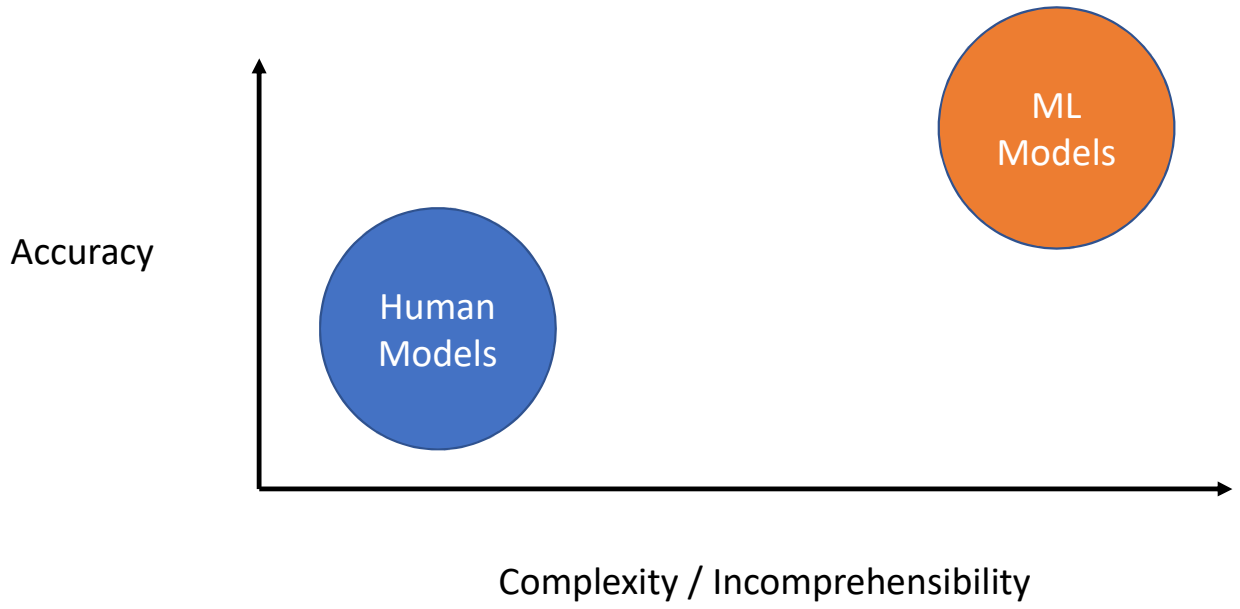
Ways that ML can potentially reduce bias

1. Objectivity in feature selection
2. Avoiding over-simplification (bias)
3. Avoiding over-generalization (variance)

1: Objectivity in feature selection: Deep Learning

- Old-fashioned Machine Learning: How should we prioritize specialists (e.g. Dermatologists) versus Primary Care Physicians (PCPs)?
  - Innate bias: Specialty is an important feature!
- Deep Learning: What features of HCPs should even be prioritized?

2. Avoiding Oversimplification



Oversimplification, Continued

- In Pharma: are we using too simple rules to make decisions about behavior? Just the ones that humans can understand and execute?
- Human analogue: Prejudices, stereotypes
- Bad ML example: misidentifying people in photos
- EU has proposed that the solution is that all ML models need to be explainable
  - Exactly the wrong solution! Need to be more complex, not less
  - Utilize ML drive suggestion engines, etc.

3. Avoiding Over-Generalization

- ML regularization techniques can help avoid “overfitting” (or variance)
- Human analogue – generalizing from too few examples
  - In Pharma: Are we basing decisions off a few anecdotal examples?
- Bad ML example: DARPA learning trucks from cars based on weather conditions
- Best Solution: More data!

Quality Measures

- Review of quality measures used in determining model fit
- How can prudent use of quality measures help reduce innate biases in data?

Quality Measures: Initial Considerations

- Smell test: Is your model at least as good as some simple model such as linear?
  - Still need to define "what good" means
  - Say accuracy to start, will discuss more why this isn't always best

- Binary classification problem should always be able to achieve accuracy of at least 50%
  - Why?

Confusion Matrix

- A confusion matrix summarizes results for a binary classification problem:
  - *a* is the number of examples of Class 0 that the classifier predicted [correctly] as Class 0.
  - *b* is the number of examples of Class 1 that the classifier predicted [incorrectly] as Class 0.
  - *c* is the number of examples of Class 0 that the classifier predicted [incorrectly] as Class 1.
  - *d* is the number of examples of Class 1 that the classifier predicted [correctly] as Class 1.

		Actual	
		Class 0	Class 1
Predicted	Class 0	*a*	*b*
	Class 1	*c*	*d*

Performance statistics based on confusion matrix

- **Accuracy** is the fraction of correct predictions:

$$\frac{a+d}{a+b+c+d}$$

		Actual	
		Class 0	Class 1
Predicted	Class 0	*a*	*b*
	Class 1	*c*	*d*

- **Error rate** is the fraction of incorrect predictions:

$$\frac{b+c}{a+b+c+d}$$

Subgroup Accuracy

- Can also talk about accuracy within a specific subgroup
  - E.g., Accuracy for class 0

$$\frac{a}{a+c}$$

		Actual	
		Class 0	Class 1
Predicted	Class 0	*a*	*b*
	Class 1	*c*	*d*

Issue with accuracy

- Consider the following classification:

		Actual	
		Class 0	Class 1
Predicted	Class 0	10000	9
	Class 1	0	0

- Is this a good classifier?

Positives and negatives

- In binary classification, often denote one group as positive (having condition), and one group negative (not having condition)

		Actual	
		Class +	Class −
Predicted	Class +	TP	FP
	Class −	FN	TN

- TP = true positive, FP = False positive, etc.

Precision and Recall

- **Precision**:: *TP / (TP + FP)*
  - What % of positive predictions were true?
  - More conservative the test, the better the precision
  - But that might sacrifice…
- **Recall**:: *TP/ (TP+FN)*
  - What % of true positives were captured correctly?
  - May get a lot of false negatives if test conservative, but if test too lenient precision suffers
- Medical diagnosis – which is better?

True Positive Rate and False Positive Rate

- True Positive Rate: TP/(Total Actual Positives = FN+TP)
  - % of Actual Positives Classified Correctly
- False Positive Rate: FP/(Total Actual Negatives = FP + TN)
  - % of Actual Negatives Classified Incorrectly

F1 Score: Balancing Precision and Recall

- How can we balance precision and recall?
  - Want high accuracy for both positive and negative classes
  - If we want to use accuracy alone, we can…?

- Another would be to take an aggregate score function, e.g. F1 score (harmonic mean of precision and recall)

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Incorporating economic cost into F1

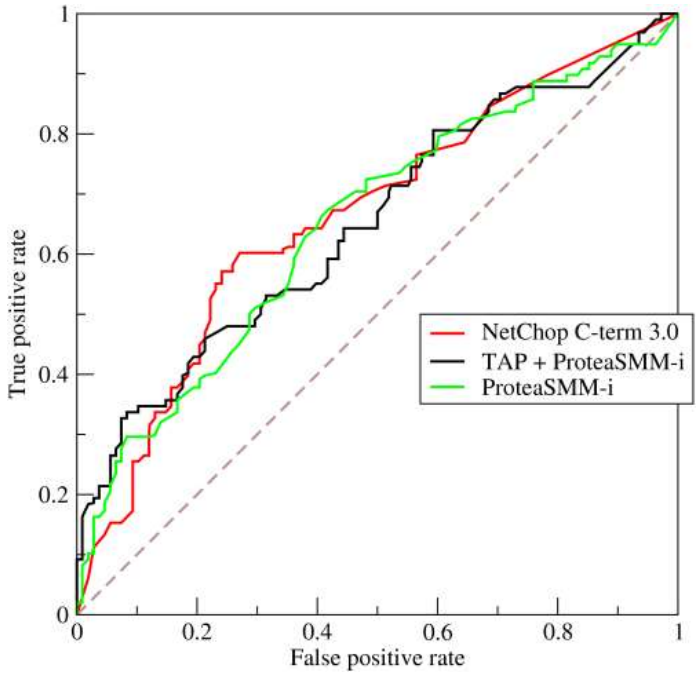
- Sometimes recall is more important than precision, etc
  - Medical test: Rather many False Positives than False Negatives
  - Can weight one of the terms in F1 to account for this

$$F_\beta = (1+\beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

ROC Curve

- Receiver Operating Characteristic: At different classification thresholds, what is the True Positive Rate versus the False Positive Rate?

- If all entries given random score, then line is linear (no predictive power)
  - Want curve to be up and to the left
- How do we capture in one metric?
  - AuROC (Area under ROC curve)



Quality Metrics in Marketing

- So far, discussed how things like medical tests need to balance precision and recall

- How does this figure into marketing decisions?
  - Incorporate economic cost of promotion versus potential upside
    - Still getting catalogs from a company you last ordered from 10 years ago?
    - Still have much higher likelihood of purchasing than random person

How can Quality Metrics overcome bias?

- Human biases tend to be rules of thumb that have highest overall accuracy but lack nuance unless there is a clear downside to risk
  - Maybe easier to assume all outcomes will turn out like the majority, but clearly this is an issue if the 1% are very bad outcomes
    - Don't need to wear a seatbelt because 99.9% of miles there will not be an accident
- Using better calibrated error functions to choose models may lead to better outcomes than human judgement!

Less Wrong Revisited

Now, suppose you built an AI. You wrote the source code yourself, and so far as you can tell by inspecting the AI's thought processes, it has no equivalent of the "absolute denial macro" - there's no training that could inflict on it the equivalent of anosognosia.

- What's the strangest thing this AI could tell you?
- The highest rated response of all time….

Why did you put an absolute denial macro in my program?

How might our AIs still be biased?

- Architecture's biased towards quick decisions?
  - Andrew Ng: ML is good at doing whatever humans can do in < 1 second
- Bias towards understandability
  - Greatest human bias: Only trust what we can understand

Thanks!

?