The Impossibility of Productizable AI:
Problems and Potential Solutions

Falaah Arif Khan,

Identity and Payments Engg, Dell Commerce Platform

Personal: https://falaaharifkhan.github.io/research

Blog: https://thefaladox.wordpress.com/

Twitter: @ArifFalaah

To Do:

- ❑ What are the problems?
- ❑ How do these problems arise?
- ❑ Why should we care?
- ❑ How do we fix these problems?

INDEPENDENT
UK'S LARGEST QUALITY DIGITAL NEWS BRAND

SUBSCRIBE NOW LOGIN

NEWS UK POLITICS US POLITICS VOICES SPORT CULTURE INDY/LIFE INDYBEST LONG READS INDY100 VOUCHERS PREMIUM CLIMATE BLOGS

FACEBOOK'S ARTIFICIAL INTELLIGENCE ROBOTS SHUT DOWN AFTER THEY START TALKING TO EACH OTHER IN THEIR OWN LANGUAGE

'you i i i everything else'

Andrew Griffin | @_andrew_griffin | Monday 31 July 2017 17:10 |
88 comments

c|net

BEST PRODUCTS REVIEWS NEWS VIDEO HOW TO SMART HOME CARS DEALS DOWNLOAD 5G

Facebook put cork in chatbots that created a secret language

Alice and Bob, the two bots, raise questions about the future of artificial intelligence.

Richard Nieva July 31, 2017 11:58 AM PDT

ES

A pair of chatbots has recently done something children often do: create a secret language.

Last month, researchers at Facebook found two bots developed in the social network's AI division had been communicating with each other in an unexpected way. The bots, named Bob and Alice, had generated a language all

FAST COMPANY

CO.DESIGN TECH WORK LIFE CREATIVITY IMPACT PODCASTS VIDEO NEWS RECOMMENDER SUBSCRIBE

07-14-17

AI Is Inventing Languages Humans Can't Understand. Should We Stop It?

Researchers at Facebook realized their bots were chattering in a new language. Then they stopped it.

"Agents will drift off understandable language and invent codewords for themselves," says Batra, speaking to a now-predictable phenomenon that's been observed again, and again, and again. "Like if I say 'the' five times, you interpret that to mean I want five copies of this item. This isn't so different from the way communities of humans create shorthands."

Independent: https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html
Cnet: https://www.cnet.com/news/what-happens-when-ai-bots-invent-their-own-language/
Fast: https://www.fastcompany.com/90132632/ai-is-inventing-its-own-perfect-languages-should-we-let-it

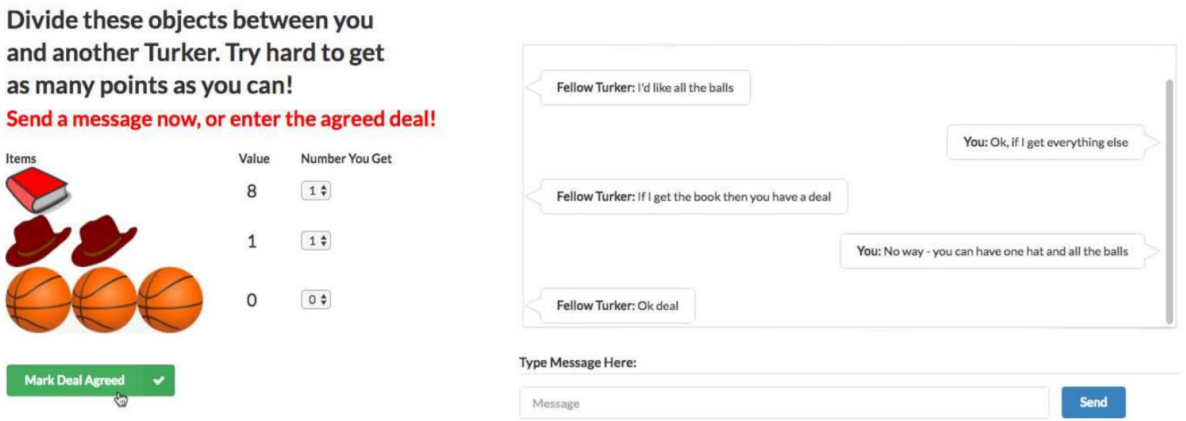


Figure 1: A dialogue in our Mechanical Turk interface, which we used to collect a negotiation dataset.

Reinforcement learning for dialog agents

Supervised learning aims to imitate the actions of human users, but it does not explicitly attempt to achieve an agent's goals. Taking a different approach, the FAIR team explored pre-training with supervised learning, and then fine-tuned the model against the evaluation metric using reinforcement learning. In effect, they used supervised learning to learn how to map between language and meaning, but used reinforcement learning to help determine which utterance to say.

During reinforcement learning, the agent attempts to improve its parameters from conversations with another agent. While the other agent could be a human, FAIR used a fixed supervised model that was trained to imitate humans. The second model is fixed, because the researchers found that updating the parameters of both agents led to divergence from human language as the agents developed their own language for negotiating. At the end of every dialog, the agent is given a reward based on the deal it agreed on. This reward was then back-propagated through every word that the agent output, using policy gradients, to increase the probability of actions that lead to high rewards.

```
Alice : book=(count:3 value:1) hat=(count:2 value:1) ball=(count:1 value:5)
Bob   : book=(count:3 value:0) hat=(count:2 value:0) ball=(count:1 value:10)
-------------------------------------------------------------------------
Bob   : i can i i everything else . . . . . .
Alice : balls have zero to me to me to me to me to me to me to me to me to
Bob   : you i everything else . . . . . .
Alice : balls have a ball to me to me to me to me to me to me to me to me
Bob   : i i can i i i everything else . . . . . .
Alice : balls have a ball to me to me to me to me to me to me to me to me
Bob   : i . . . . . .
Alice : balls have zero to me to me to me to me to me to me to me to me to
Bob   : you i i i i i everything else . . . . . .
Alice : balls have 0 to me to me to me to me to me to me to me to me to
Bob   : you i i i everything else . . . . . .
Alice : balls have zero to me to me to me to me to me to me to me to me to
```

Models produce meaningful novel sentences. One interesting question is whether our models are capable of generating novel sentences in the new circumstances they find themselves in, or if they simply repeat messages from the training data verbatim. We find that 76% of messages produced by the LIKELIHOOD model in self-play were found in the training data. We manually examined the novel utterances produced by our model, and found that the overwhelming majority were fluent English sentences in isolation—showing that the model has learnt a good language model for the domain (in addition to results that show it uses language effectively to achieve its goals). These results suggest that although neural models are prone to the safer option of repeating sentences from training data, they are capable of generalising when necessary. Future work should choose domains that force a higher degree of diversity in utterances.

Maintaining multi-sentence coherence is challenging. One common linguistic error we see RL+ROLLOUTS make is to start a message by indicating agreement (e.g. *I agree* or *Deal*), but then going on to propose a counter offer—a behaviour that human partners found frustrating. One explanation is that the model has learnt that in the supervised data, messages beginning with *I agree* are often at the end of the dialogue, and partners rarely reply with further negotiation—so the models using rollouts and reinforcement learning believe this tactic will help their offer to be accepted.

Blog Post: https://engineering.fb.com/ml-applications/deal-or-no-deal-training-ai-bots-to-negotiate/ Paper: https://arxiv.org/abs/1706.05125

What are the problems?

- **Exaggerated Capabilities**

  Reporting != Research

About the Apple Card

By Jamie Heinemeier Hansson on November 11, 2019

My name is Jamie Heinemeier Hansson. Since my husband, David, tweeted about an unfortunate and ridiculous situation with AppleCard that involves me, I have been (or my credit-worthiness has been) the subject of lots of speculation. Unlike David, I am an extremely private person who does not post on social media. I am slightly mortified to have my name in the news. However, lest I be cast as a meek housewife who cannot speak for herself, I would like to make the following statement:

I care about digital privacy. It's why I wanted an AppleCard in the first place.

I care about transparency and fairness. It's why I was deeply annoyed to be told by AppleCard representatives, "It's just the algorithm," and "It's just your credit score." I have had credit in the US far longer than David. I have never had a single late payment. I do not have any debts. David and I share all financial accounts, and my very good credit score is higher than David's. I had a career and was successful prior to meeting David, and while I am now a mother of three children — a "homemaker" is what I am forced to call myself on tax returns — I am still a millionaire who contributes greatly to my household and pays off credit in full each month. But AppleCard representatives did not want to hear any of this. I was given no

GS Bank Support @gsbanksupport · Nov 11, 2019

We wanted to address some recent questions regarding the #AppleCard credit decision process.

We wanted to address some recent questions regarding the Apple Card credit decision process.

With Apple Card, your account is individual to you; your credit line is yours and you establish your own direct credit history. Customers do not share a credit line under the account of a family member or another person by getting a supplemental card.

As with any other individual credit card, your application is evaluated independently. We look at an individual's income and an individual's creditworthiness, which includes factors like personal credit scores, how much debt you have, and how that debt has been managed. Based on these factors, it is possible for two family members to receive significantly different credit decisions.

In all cases, we have not and will not make decisions based on factors like gender.

Finally, we hear frequently from our customers that they would like to share their Apple Card with other members of their families. We are looking to enable this in the future.

- Andrew Williams, Goldman Sachs Spokesperson

DHH @dhh · Nov 9, 2019

Replying to @dhh

So nobody understands THE ALGORITHM. Nobody has the power to examine or check THE ALGORITHM. Yet everyone we've talked to from both Apple and GS are SO SURE that THE ALGORITHM isn't biased and discriminating in any way. That's some grade-A management of cognitive dissonance.

DHH
@dhh

Apple has handed the customer experience and their reputation as an inclusive organization over to a biased, sexist algorithm it does not understand, cannot reason with, and is unable to control. When a trillion-dollar company simply accepts the algorithmic overlord like this...

♡ 4,396 4:59 AM - Nov 9, 2019

Steve Wozniak
@stevewoz

Replying to @dhh @AppleCard

The same thing happened to us. I got 10x the credit limit. We have no separate bank or credit card accounts or any separate assets. Hard to get to a human for a correction though. It's big tech in 2019.

♡ 4,150 6:21 AM - Nov 10, 2019

807 people are talking about this

About the Apple Card, Blog post: https://dhh.dk/2019/about-the-apple-card.html
Tweets: https://twitter.com/gsbanksupport/status/1193703266003177472?s=21
https://twitter.com/stevewoz/status/1193330241478901760?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1193330241478901760&ref_url=https%3A%2F%2Fwww.bloomberg.com%2Fnews%2Farticles%2F2019-11-10%2Fapple-co-founder-says-goldman-s-apple-card-algo-discriminates

What are the problems?

- ❑ Exaggerated Capabilities

  Reporting != Research

- ❑ **Math-washing**

  “Algorithms are neutral, they just know Math” != True

PREDICTIVE POLICING

GOOGLE'S HATE SPEECH DETECTOR

AMAZON'S TALENT SEARCH ENGINE

MEDICAL TREATMENT RECOMMENDATIONS

Predictions != Policies

What are the problems?

- ❑ Exaggerated Capabilities

  Reporting != Research

- ❑ Math-washing

  “Algorithms are neutral, they just know Math” != True

- ❑ **Predictions != Policies**

https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

What are the problems?

- ❑ Exaggerated Capabilities

  Reporting != Research

- ❑ Math-washing

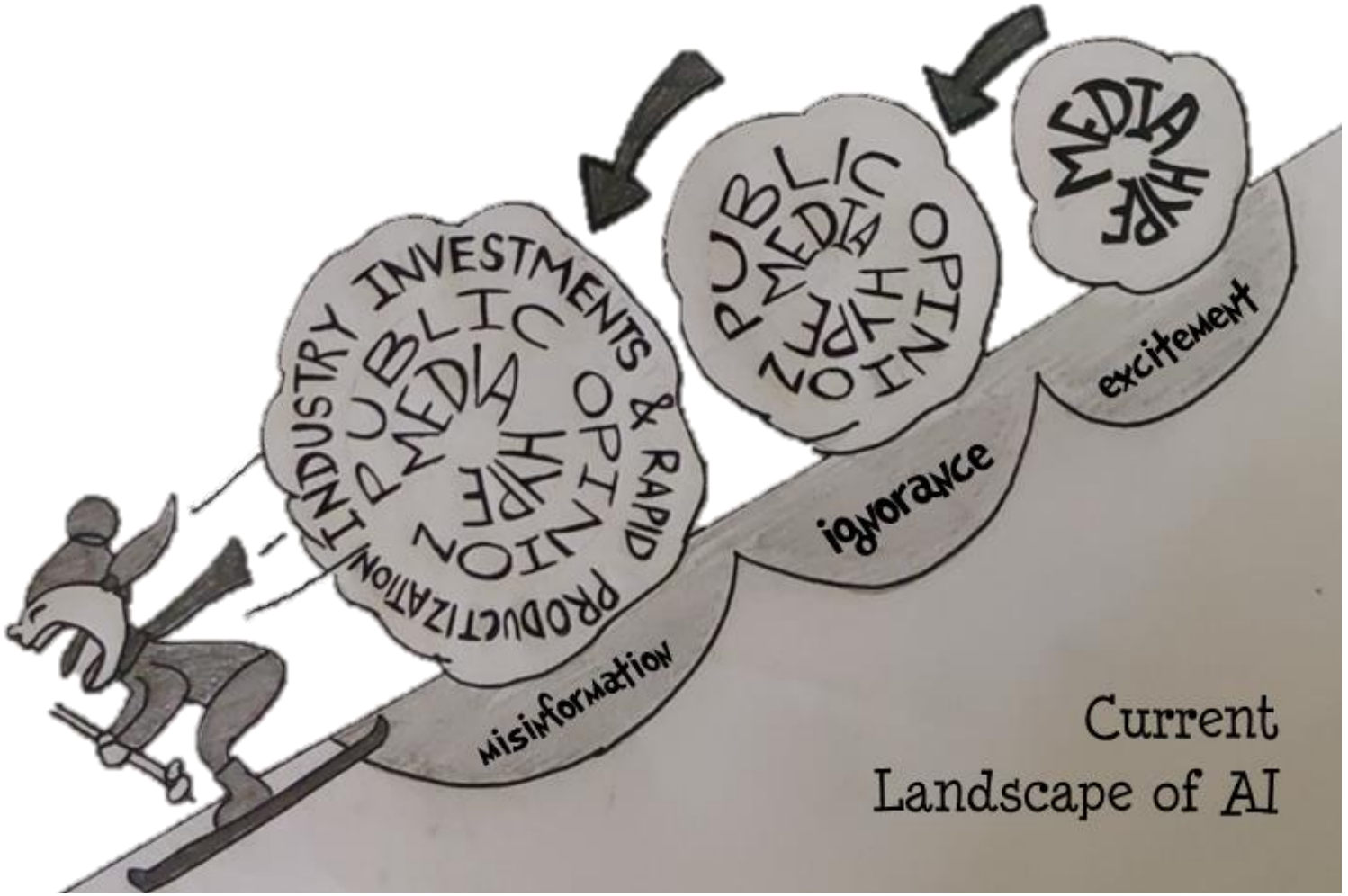
  “Algorithms are neutral, they just know Math” != True

- ❑ Predictions != Policies

- ❑ **Models cannot “thrive in the wild”**

Reporting != Research

How/Why do these problems arise?



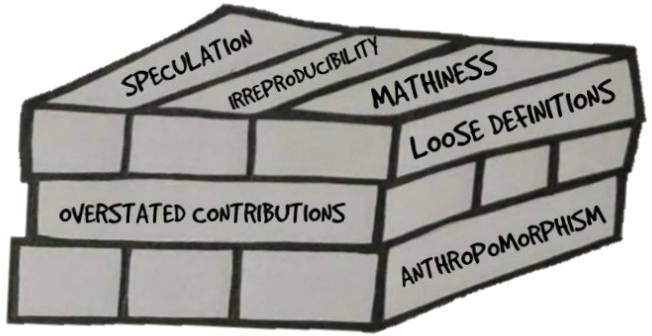
https://falaaharifkhan.github.io/research/documents/Vol1.pdf

Current State of Research

Think of ML Scholarship as a game of

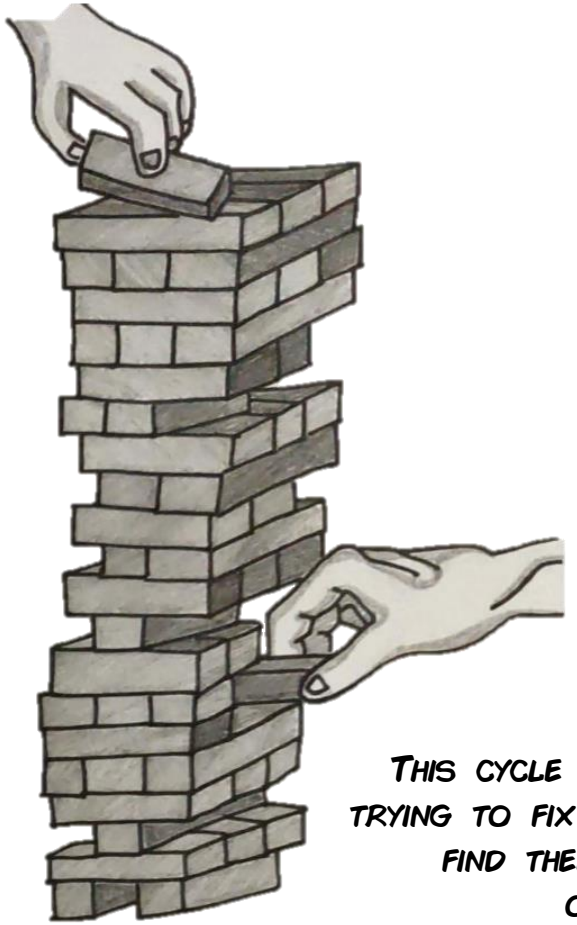
Jenga...

When demand > supply,
Boom of overnight experts
+
Media hype



Leads to a shoddy foundation and a skewed incentive scheme

Novices turn to "experts" and are motivated by the same incentive scheme

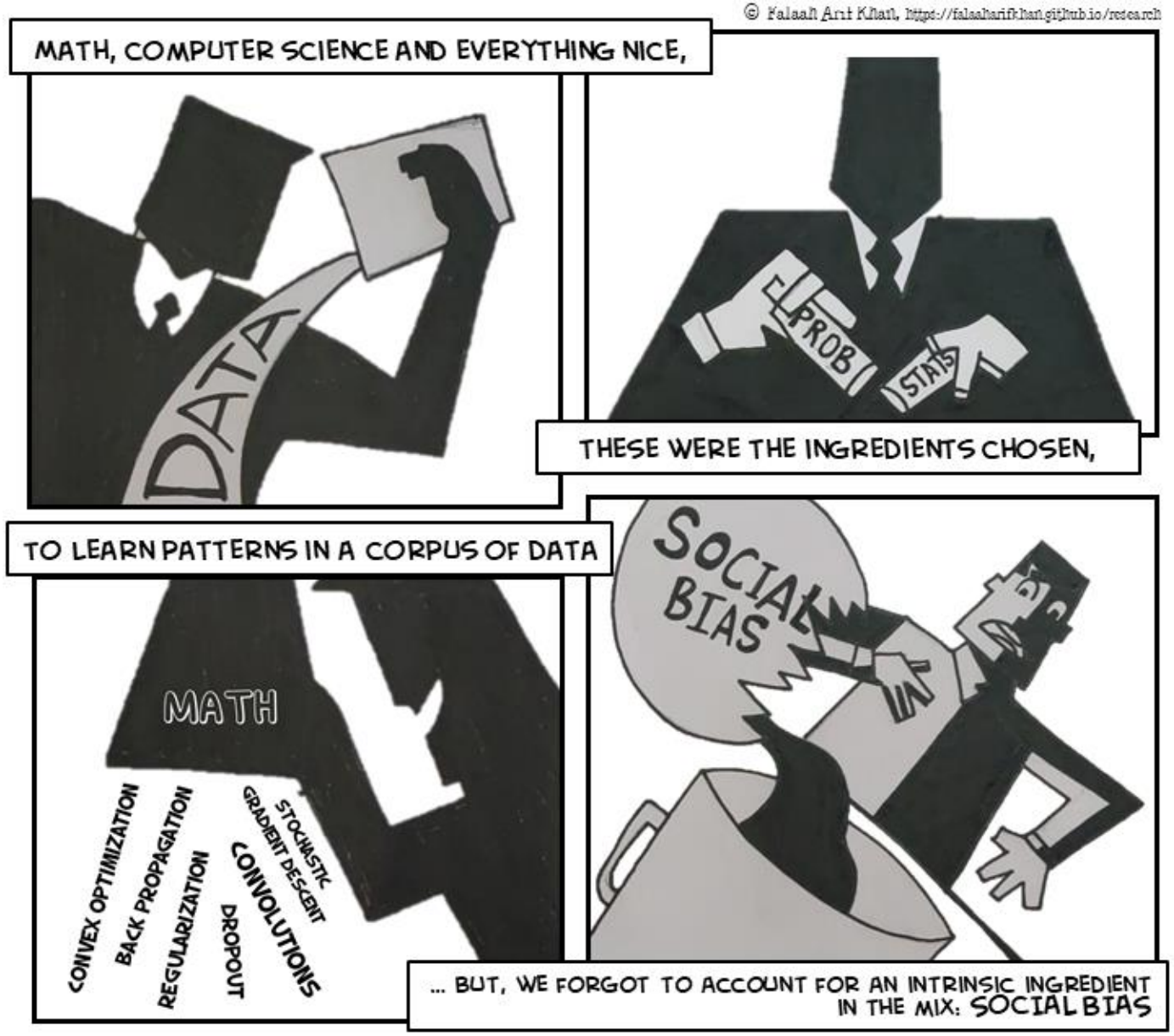


This cycle propagates and those trying to fix the system from within find themselves dreadfully outnumbered

How many **rounds** of this **game** do we have **left**?

https://falaaharifkhan.github.io/research/documents/Vol1.pdf

Math-washing does not work

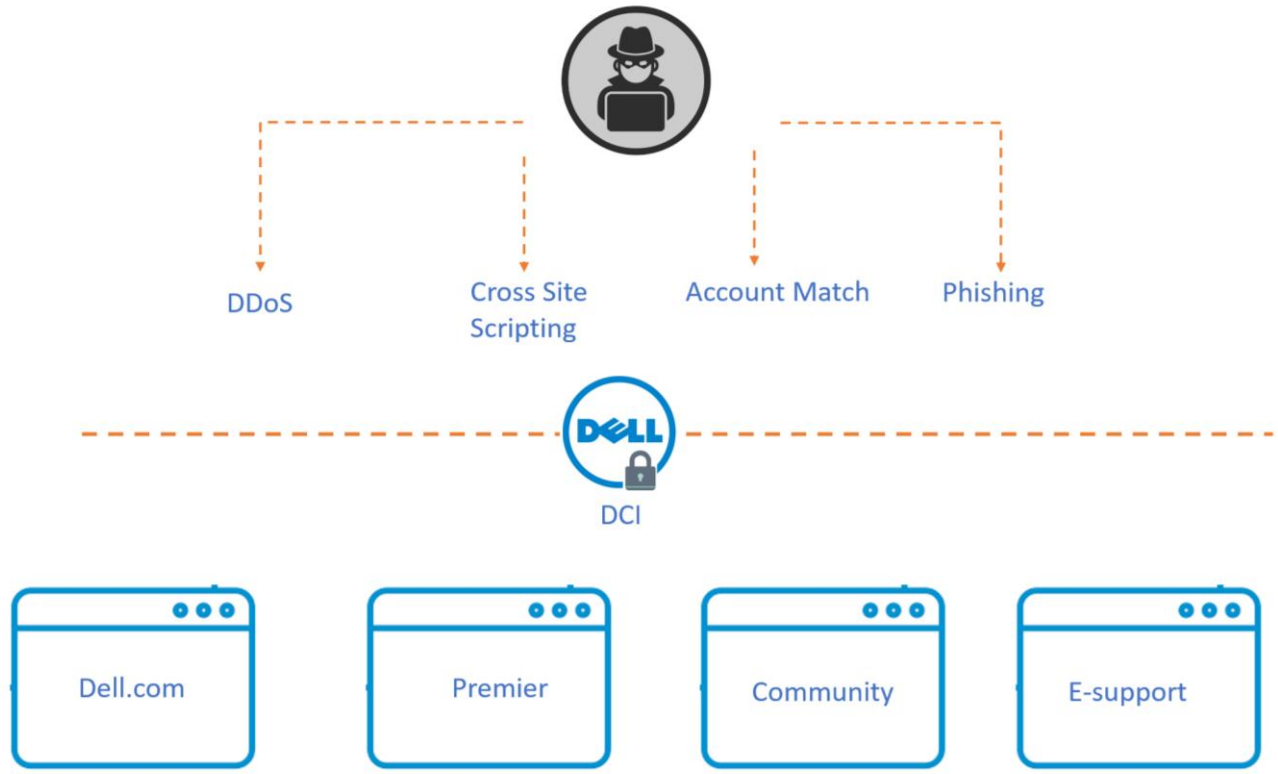


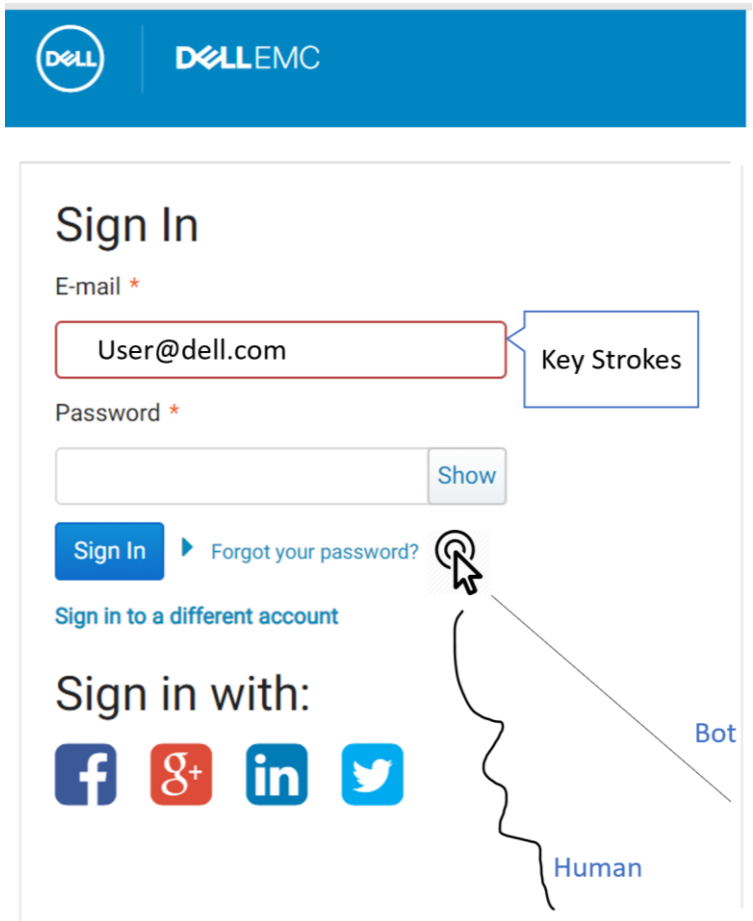
https://falaaharifkhan.github.io/research/documents/Vol1.pdf

Algorithmic Limitations:

Can we ensure IID in the real world?

Behavioral Biometrics and Machine Learning to secure Website Logins



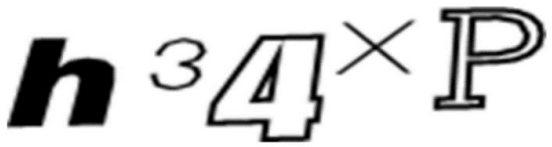


Paper: Arif Khan F., Kunhambu S., G K.C. (2019) Behavioral Biometrics and Machine Learning to Secure Website Logins

US Patent: Arif Khan, Falaah, Kunhambu, Sajin and Chakravarthy G, K. Behavioral Biometrics and Machine Learning to secure Website Logins. US Patent 16/257650, filed January 25, 2019

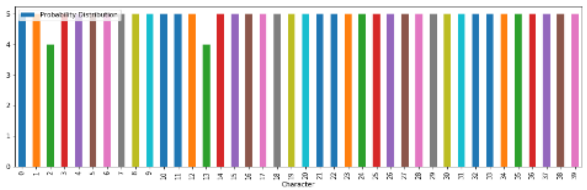
Algorithmic Limitations: Robustness-Generalization Dichotomy

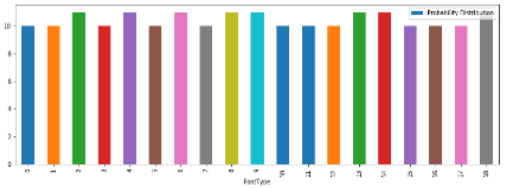
Make CAPTCHAs smart again: A framework to design Completely Automated Reverse Turing tests

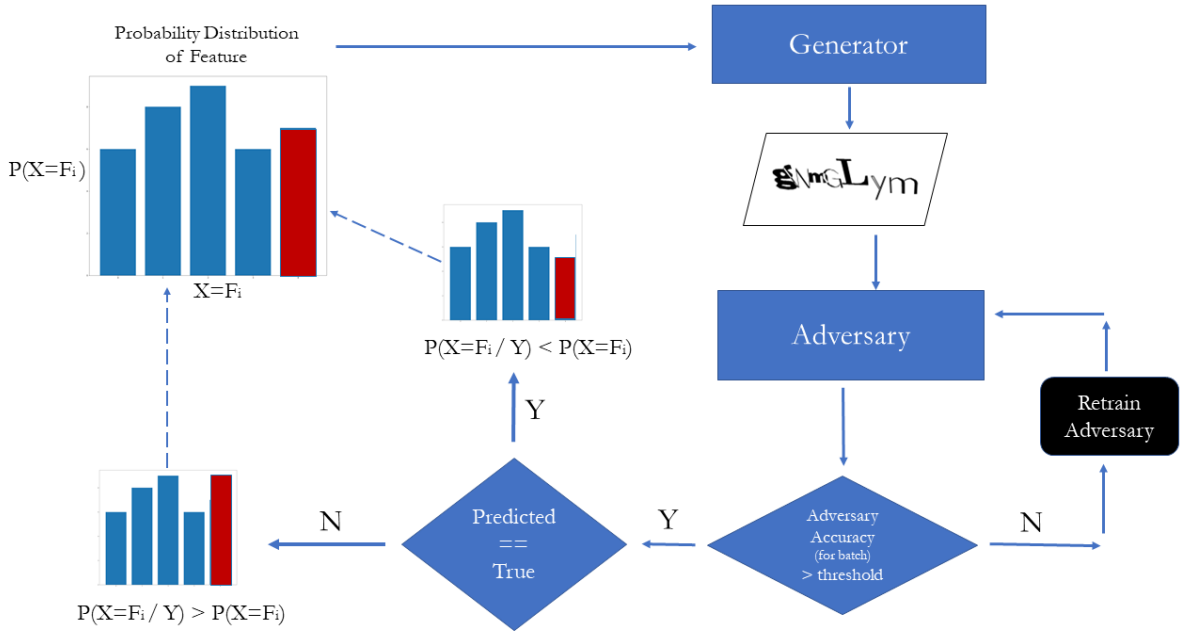


Human Preferences: Solved (not refreshed), solved correctly
Attacker Preferences: Custom deep OCR performance

Character Parameters					
Character	h	3	4	X	P
Font Type	Font 1	Font 7	Font 4	Font 3	Font 9
Font Size	74	62	73	63	77
Hollow/ Solid	Solid	Solid	Hollow	Solid	Hollow
X Coordinate	21	60	93	129	169
Y Coordinate	49	48	54	41	46
Image Parameters					
Skew Points	P1(x1,y1)	P2(x2,y2)	P3(x3,y3)	P4(x4,y4)	







Bayesian Inference:

$$p(\theta|\mathbf{D}) = \frac{\mathcal{L}(\mathbf{D}|\theta)\pi(\theta)}{\mathbf{p}(\mathbf{D})}$$

Paper: Arif Khan F, Sharma H. "Making CAPTCHAs smart again: A framework to Design Completely Automated Reverse Turing tests (CART Framework)", in preparation
US Patent: Arif Khan, Falaah and Sharma, Hari Surender. Framework to Design Completely Automated Reverse Turing Tests, authorized, in preparation

Algorithmic Limitations:

Correlation != Causation

PREDICTIVE POLICING

GOOGLE'S HATE SPEECH DETECTOR

AMAZON'S TALENT SEARCH ENGINE

MEDICAL TREATMENT RECOMMENDATIONS

Ill posed problems

How do we fix these problems?

I don’t know!

Write testable Machine Learning models

Unit Testing

Break model into modular, testable chunks and test individual components
Eg: Sample on distribution classes and check empirical vs exact, Test gradient update steps (Eg: scipy.optimize.check_grad)

Functional/Integration Testing

Write test cases based on prediction (probability, MSE)
(Given) Take a sample from each class
(When) Run test cases
(Then) Check if result within specified threshold (acceptance criteria)

Still operating under IID!
Cannot detect Distribution Shift!

Write test cases that check for Distribution Shift
All statistical measurements should be indistinguishable between source and target distributions, Progressively evaluate model performance
> Retrain or Predict under distribution shift

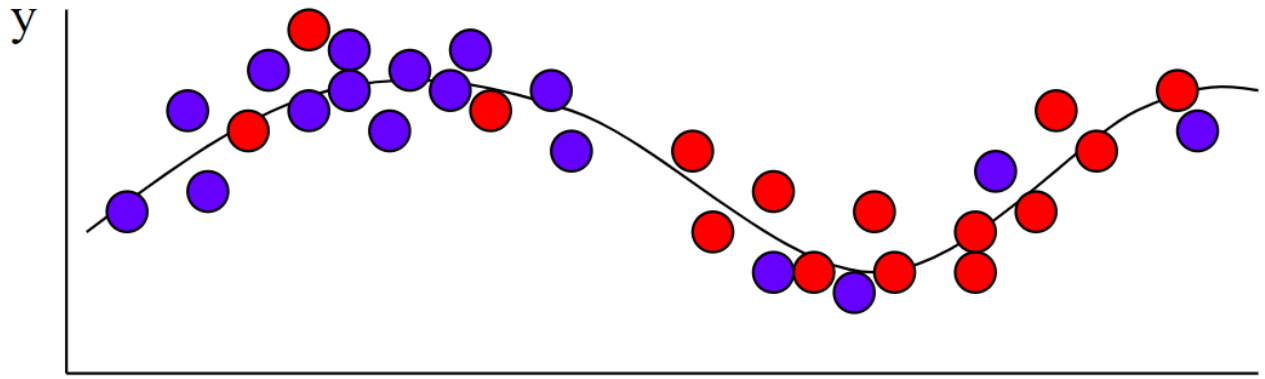
Roger Grosse: https://lips.cs.princeton.edu/testing-mcmc-code-part-1-unit-tests/, https://lips.cs.princeton.edu/testing-mcmc-code-part-2-integration-tests/

Prediction under Distribution Shift

Covariate Shift:

$$P_s(y/x) = P_t(y/x)$$

$$P_s(x) != P_t(x)$$



Amos Storkey, 2009 https://slideplayer.com/slide/9029382/

Detect:

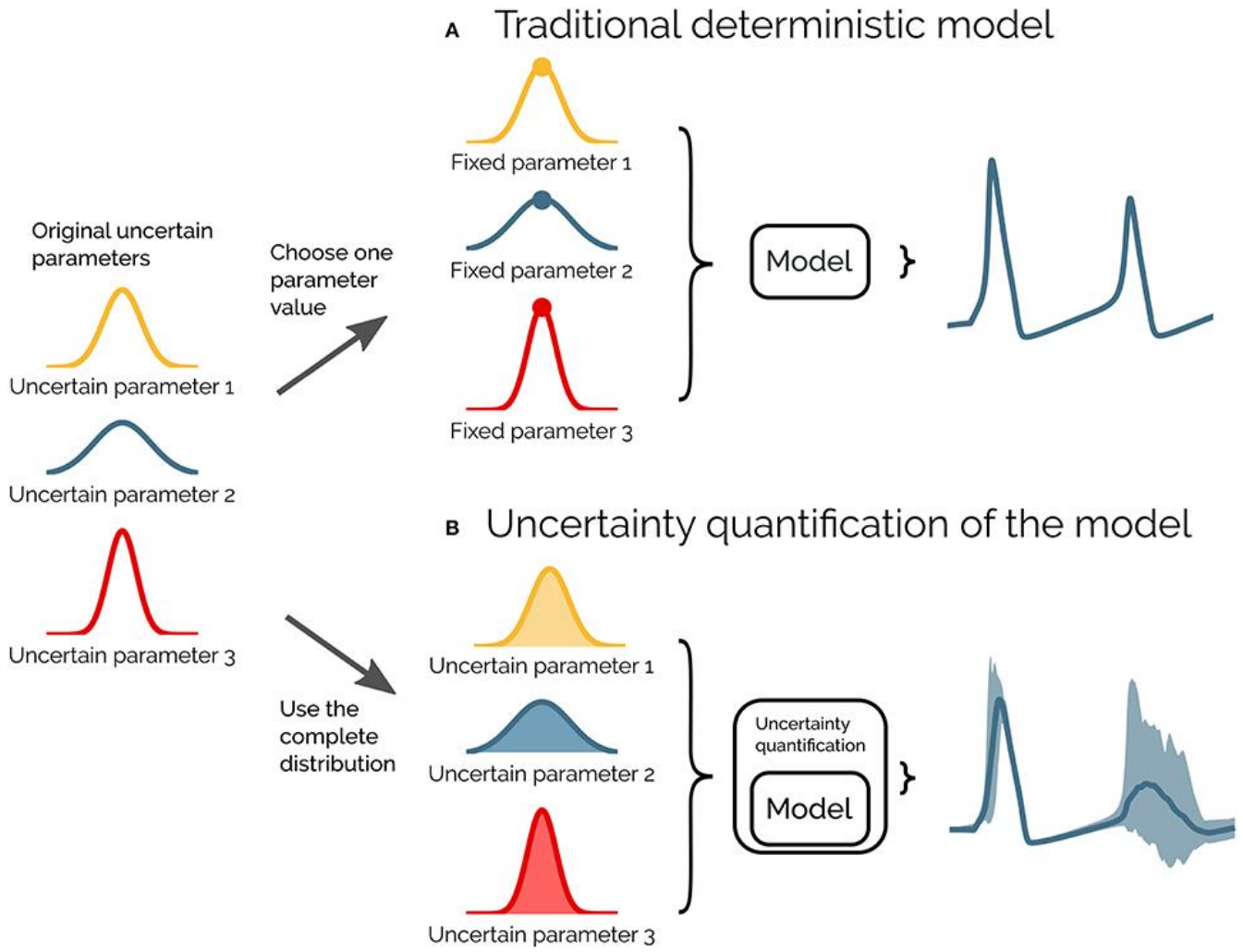
- ❑ Statistical Distance: Create histograms of classifications, important features, etc
- ❑ Discriminative Distance: Train a classifier to detect which distribution the sample came from. Training error ~ distance, training error is high, not able to distinguish between the two distributions

Predict:

- ❑ Upweight samples that do not change and retrain
- ❑ Find a mapping from S to T
- ❑ Drop features that change from S to T

Nice Talk! Peter Prettenhofer: "Dataset shift in machine learning", https://www.youtube.com/watch?v=Cwy2Qf8fT2A

Uncertainty Quantification and Sensitivity Analysis



https://www.frontiersin.org/articles/10.3389/fninf.2018.00049/full

Think critically!