Computers are Stupid: Protecting “AI” from Itself

Katharine Jarmul - KIProtect

GOTO Berlin 2018

Neue Religion will gottgleiche Künstliche Intelligenz verehren

Translate

English | Spanish | French | German - detected

Die Volkswirtschaftslehre (auch Nationalökonomie, Wirtschaftliche Staatswissenschaften oder Sozialökonomie, kurz VWL), ist ein Teilgebiet der Wirtschaftswissenschaft.

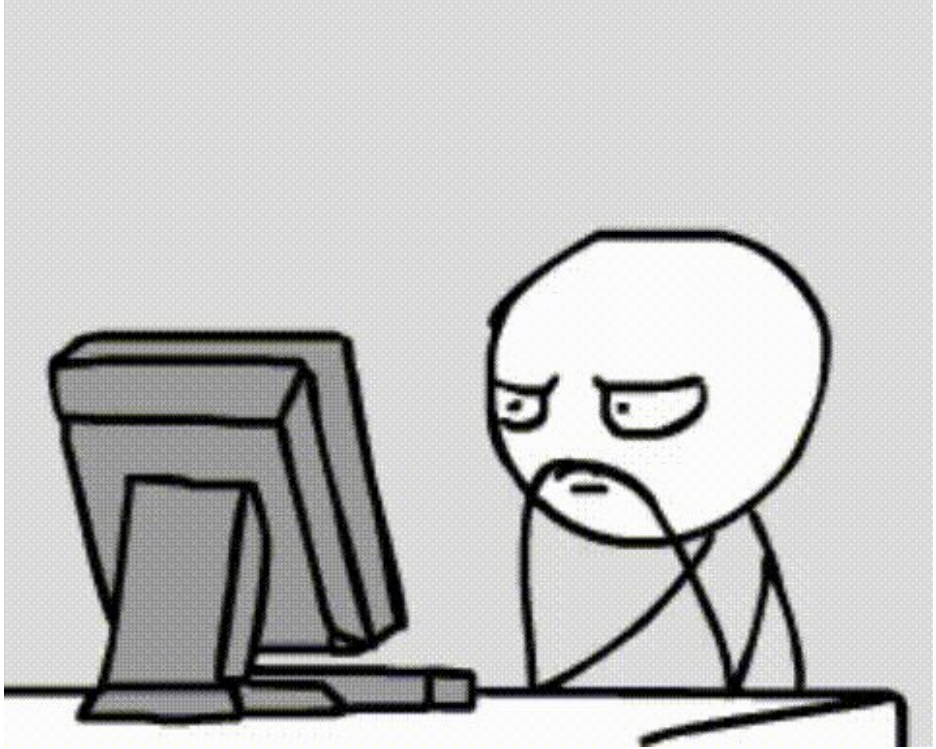
167/5000

English | Spanish | Arabic | Translate

The economics of economics (including economics, economics, economics, economics, economics, economics) is a part of economics.

(a) Input 1 (b) Input 2 (darker version of 1)

Figure 1: An example erroneous behavior found by DeepXplore in Nvidia DAVE-2 self-driving car platform. The DNN-based self-driving car correctly decides to turn left for image (a) but incorrectly decides to turn right and crashes into the guardrail for image (b), a slightly darker version of (a).



Computers are Stupid; But Humans are Smart

Adversarial Examples

Athalye et al. *Synthesizing Robust Adversarial Examples*, 2017.

Poisoned Data

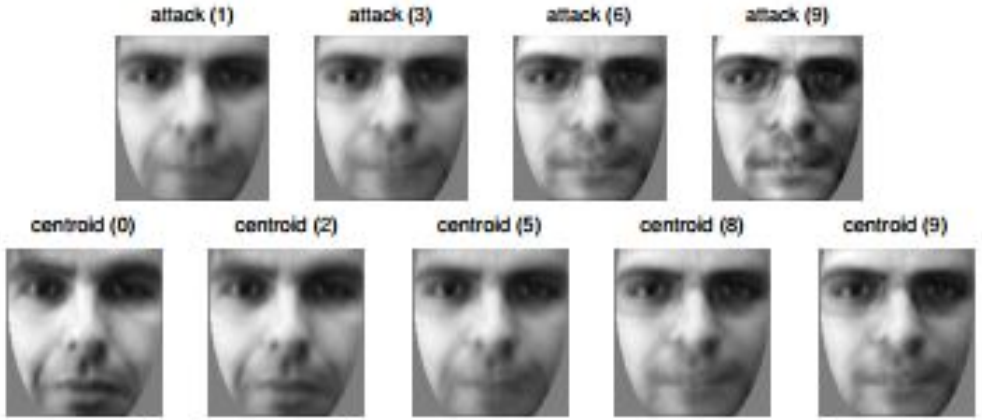


Figure 3. Attack samples (top) and victim's centroids (bottom) for poisoning with perfect knowledge, at different iterations.

Biggio et al. *Poisoning Attacks to Compromise Face Templates*, 2013

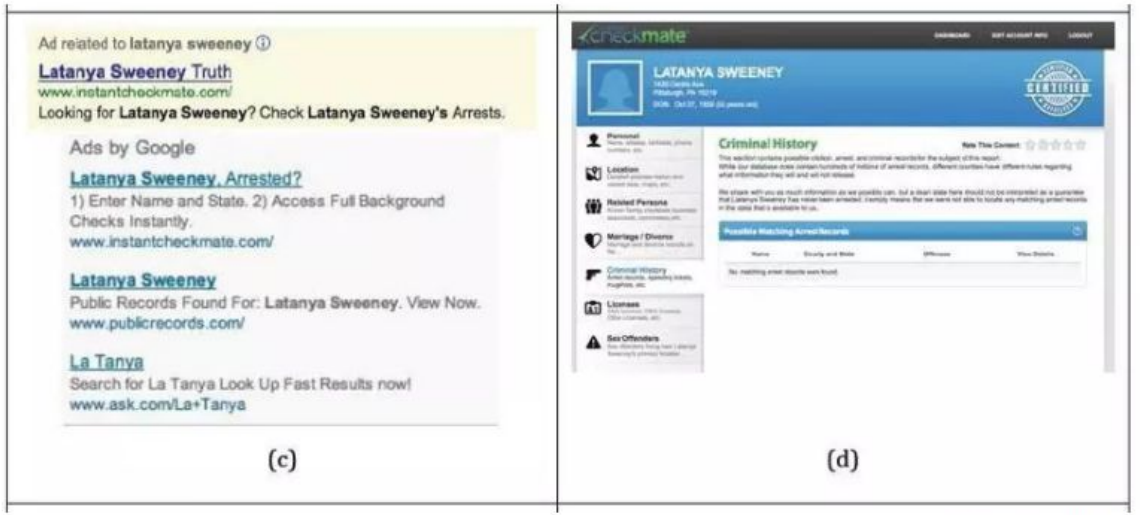
Malicious Business Interests

Computers are Stupid, Humans are Smart, But Prone to Bias

Ethical Issues

Racial bias alleged in Google's ad results

Names associated with blacks prompt link to arrest search



LATANYA SWEENEY

Web page results of ads that appeared on-screen when Harvard professor Latanya Sweeney typed her name in a google search. Ads featured services for arrest records. Sweeney conducted a study that concluded searches with "black sounding" names are more likely to get results with ads for arrests records and other negative information.

By Hiawatha Bray | GLOBE STAFF FEBRUARY 06, 2013

Privacy Issues

Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Fredrikson et al. *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, 2015

Model Explanations

```
In [5]:  1  eli5.explain_weights(model)
```

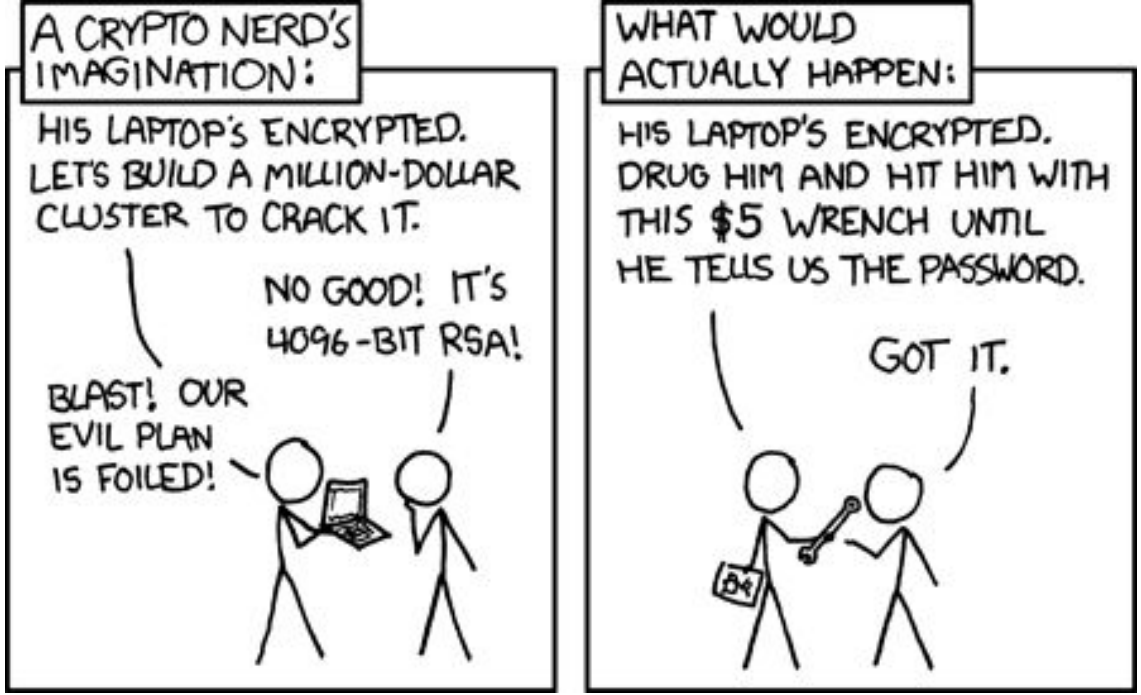
Out[5]: **y** top features

Weight[?]	Feature
+11.948	<BIAS>
+0.157	x7
+0.083	x75
+0.038	x3437
+0.036	x4787
+0.035	x3669
+0.035	x2344
+0.034	x10
+0.034	x73
+0.034	x836
+0.033	x1923
+0.033	x16
+0.030	x3391
+0.030	x98
+0.029	x2596
+0.028	x14
+0.028	x97
… 387 more positive …	
… 325 more negative …	
-0.032	x4817

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Ribeiro et al., *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*. 2016

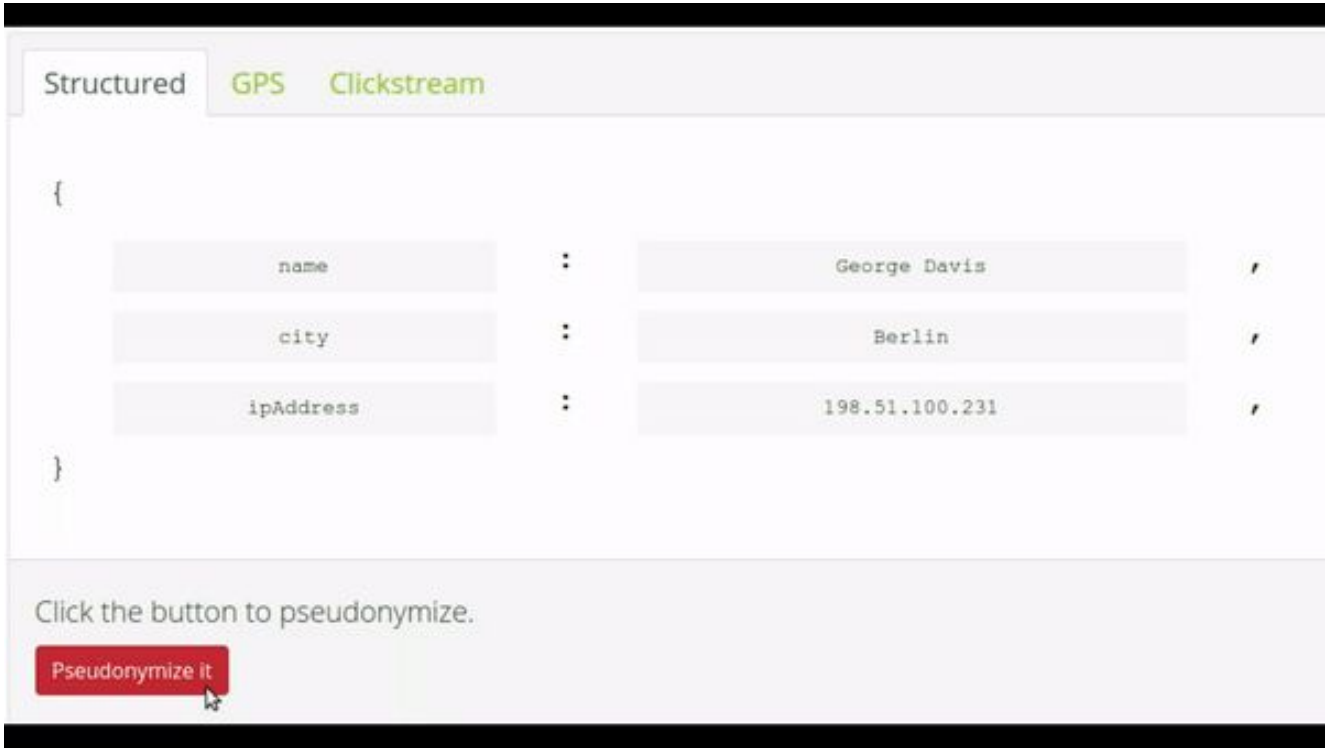


How Can We Protect "AI" From Itself, Clever Humans and Human Biases?

Protecting Model APIs

https://evademl.org

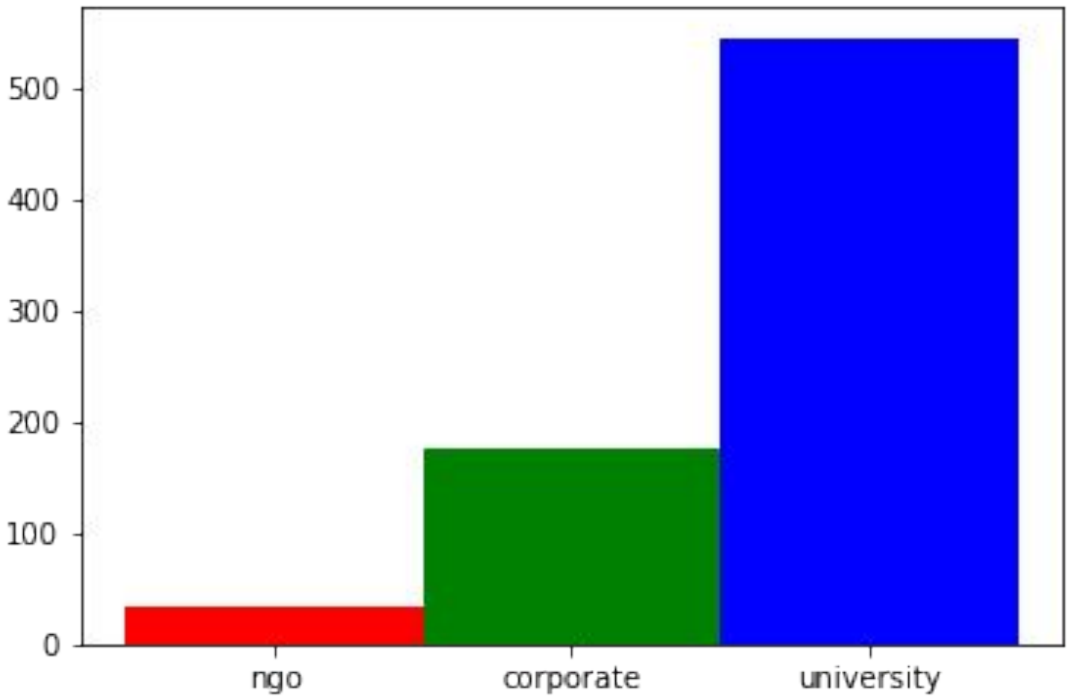
Protecting User Data



Interdisciplinary & Social Collaboration

Source: [Center for a New American Security (CNAS)]

All Voices > Some Voices



NIPS 2018 Paper Submissions (grouped by author employer)

Stupid Computer

Compounding Our Own Problems

I Thought It Would Help

A Haiku written by Natural Intelligence

Thank you!

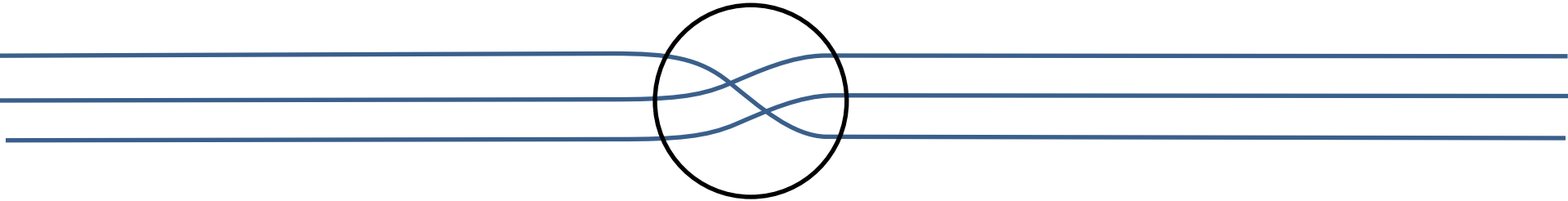
Questions? I'd love to hear them!

Or reach out anytime:

info@kiprotect.com
@KIProtect (Twitter)
https://github.com/kiprotect

Katharine Jarmul
katharine@kiprotect.com
@kjam (Twitter)

7scientists GmbH
KIProtect
Bismarckstr. 10-12
10625 Berlin



Slide References

- AI Religion: https://www.techbook.de/easylife/web/religion-kuenstliche-intelligenz-way-of-the-future
- AI is the new Electricity: https://www.youtube.com/watch?v=fgbBtnCvcDI
- Google Translate Fail: https://www.reddit.com/r/funny/comments/6c2n0n/the_german_language/
- Siri Fails: http://whysiriwhy.com / https://mashable.com
- Adversarial Turtle Video: https://www.youtube.com/watch?v=YXy6oX1iNoA
- Adversarial Turtle Paper: https://arxiv.org/abs/1707.07397
- Poisoning Attack: https://pralab.diee.unica.it/sites/default/files/biggio-ICB2013.pdf
- Cambridge Analytica Facebook Ads: https://www.buzzfeednews.com/article/craigsilverman/cambridge-analytica-says-they-won-the-election-for-trump
- Latanya Sweeney paper on Boston Globe: https://www.bostonglobe.com/business/2013/02/06/harvard-professor-spots-web-search-bias/PtOgSh1ivTZMfyEGj00X4I/story.html
- Model Inversion Attack: https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf
- Membership Inference Attack: https://arxiv.org/pdf/1610.05820.pdf
- Model Explanations (LIME): https://homes.cs.washington.edu/~marcotcr/blog/lime/
- XKCD: https://xkcd.com/538/
- Feature Squeezing: https://evademl.org/squeezing/
- KIProtect Whitepaper: Please reach out at: whitepaper@kiprotect.com
- AI Safety Panel: https://www.youtube.com/watch?v=6sCKa5and1I
- NIPS and ICML Statistics: https://medium.com/@karpathy/icml-accepted-papers-institution-stats-bad8d2943f5d and https://medium.com/machine-learning-in-practice/nips-accepted-papers-stats-26f124843aa0