How to open an algorithmic black box?

The multiple ways of making transparency: from source code to counterfactual examples

Katja de Vries, Assistant Professor in Public Law, Uppsala University

THE XXXV NORDIC CONFERENCE ON LAW & IT
Transparency session
11 November 2020, 13-16

Plan for today

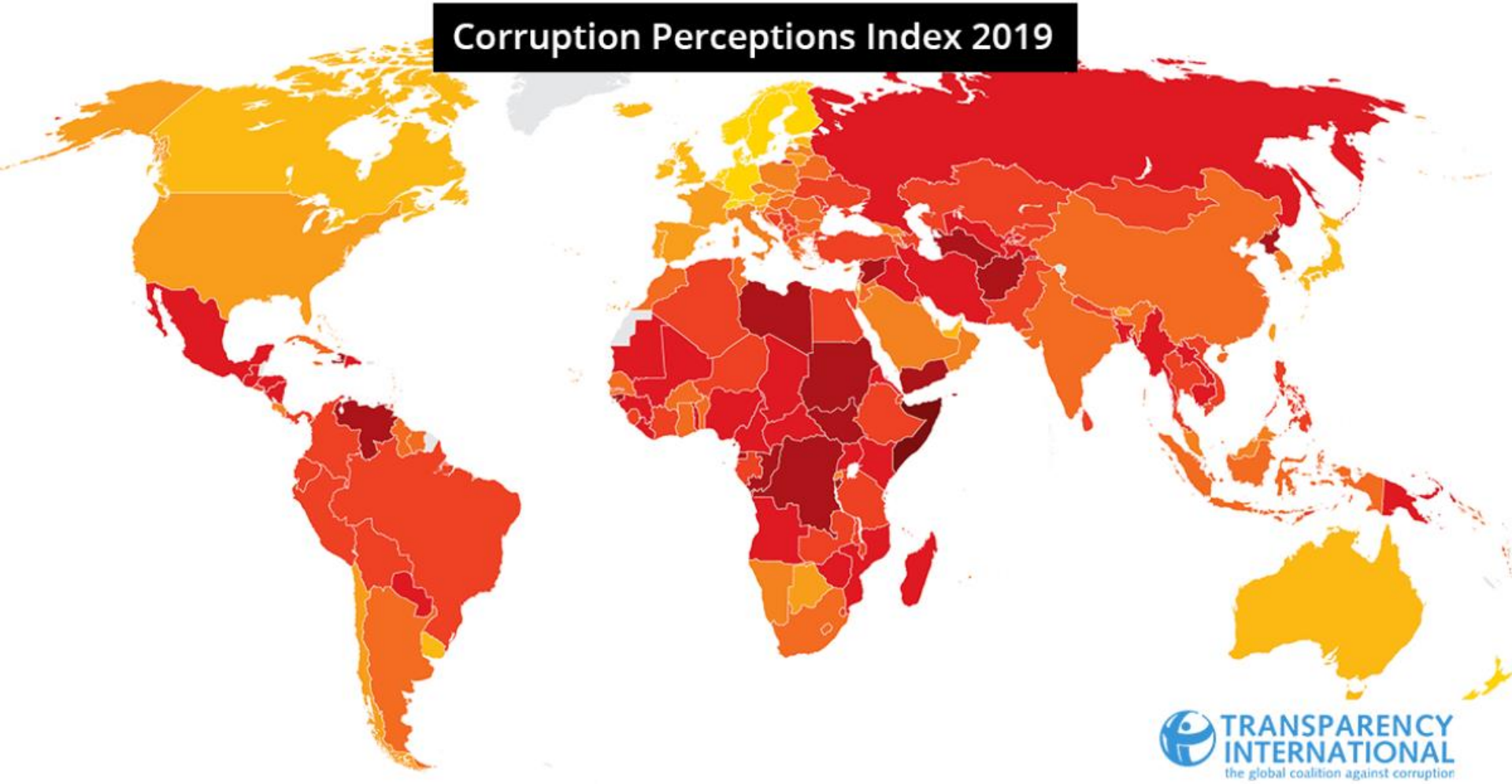
1. Transparency – why?
2. Automated Decision Making (and discriminatory ML)
3. Multiple ways of opening the blackbox
4. Generative ML
5. Counterfactuals – the good
6. Counterfactuals – the bad
7. Concluding thoughts

Part 1

Transparency – *why?*

Transparency

https://www.pikrepo.com/fjfpo/sun-rays-on-the-window


Corruption Perceptions Index 2019
TRANSPARENCY INTERNATIONAL
the global coalition against corruption


Computer says no.

GDPR, Recital 39

Any processing of personal data (…) should be **transparent** to natural persons that personal data concerning them are collected, used, consulted or otherwise processed and to what extent the personal data are or will be processed. The **principle of transparency** requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand, and that clear and plain language be used. That principle concerns, in particular, information to the data subjects on the identity of the controller and the purposes of the processing and further information to ensure fair and **transparent processing** in respect of the natural persons concerned and their right to obtain confirmation and communication of personal data concerning them which are being processed. Natural persons should be made aware of risks, rules, safeguards and rights in relation to the processing of personal data and how to exercise their rights in relation to such processing.

Profile transparency

Providing "meaningful information about the logic involved" and about the "significance and the envisaged consequences" of the profiling.

(Arts. 12-15 GDPR 2016/679)

Limitations

profile transparency is subject to limitations (see recital 42 of the DPD 95/46 and recital 63 of GDPR 2016/679) due to trade secrets and Intellectual Property Rights (IPRs)

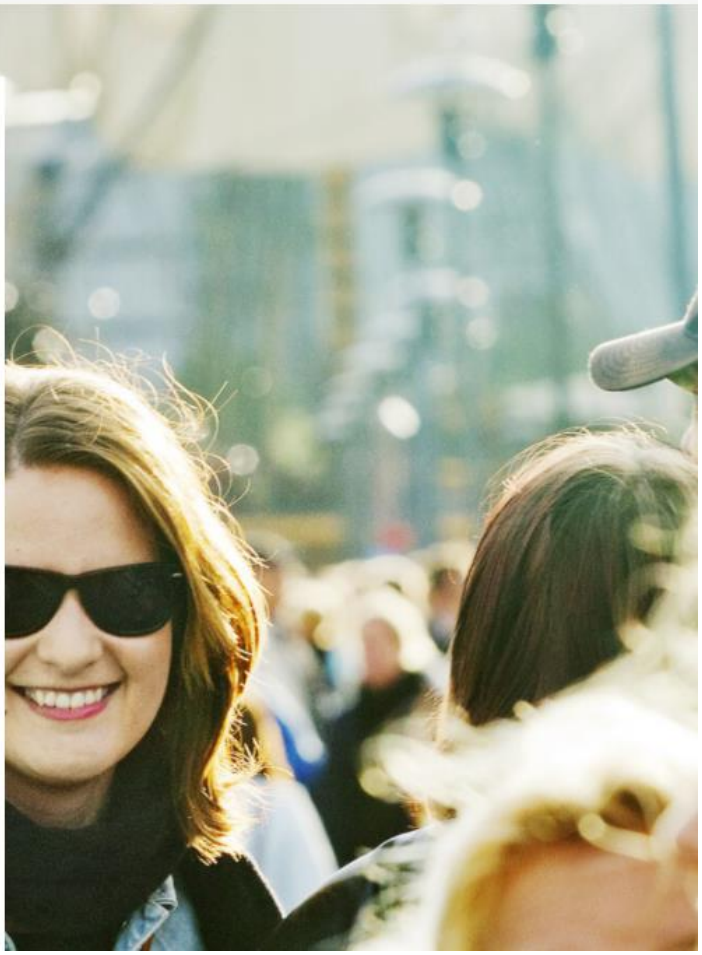
Gemeente Amsterdam

City of Amsterdam Algorithm Register Beta

Algorithm Register | More information | Participate in a survey

What is the Algorithm Register?

The Algorithm Register is an overview of the artificial intelligence systems and algorithms used by the City of Amsterdam. Through the register, you can get acquainted with the quick overviews of the city's algorithmic systems or examine their more detailed information based on your own interests. You can also give feedback and thus participate in building human-centered algorithms in Amsterdam. The register is still under development.



Holiday rental housing fraud risk

From 1 July 2020, a pilot will be carried out for six months with an algorithm that supports the employees of the department of Surveillance & Enforcement in their investigation of the reports made concerning possible illegal holiday rentals. The algorithm helps prioritize the reports so that the limited enforcement capacity can be used efficiently and effectively. By analyzing the data of related housing fraud cases of the past 5 years, it calculates the probability of an illegal holiday rental situation on the reported address.

Sneak-preview: counterfactual examples

Sandra Wachter, et al., Counterfactual explanations without opening the black box: automated decisions and the GDPR, 31 *Harvard Journal of Law & Technology* (2018).


Computer says no.

If transparency is the *answer*, what is the *problem*?

(1) systemic problems of discriminatory bias

(2) decisions with significant consequences lacking an appropriate justification

(3) objectification of the subject (no dignity)

If transparency is the *answer*, what is the *problem*?

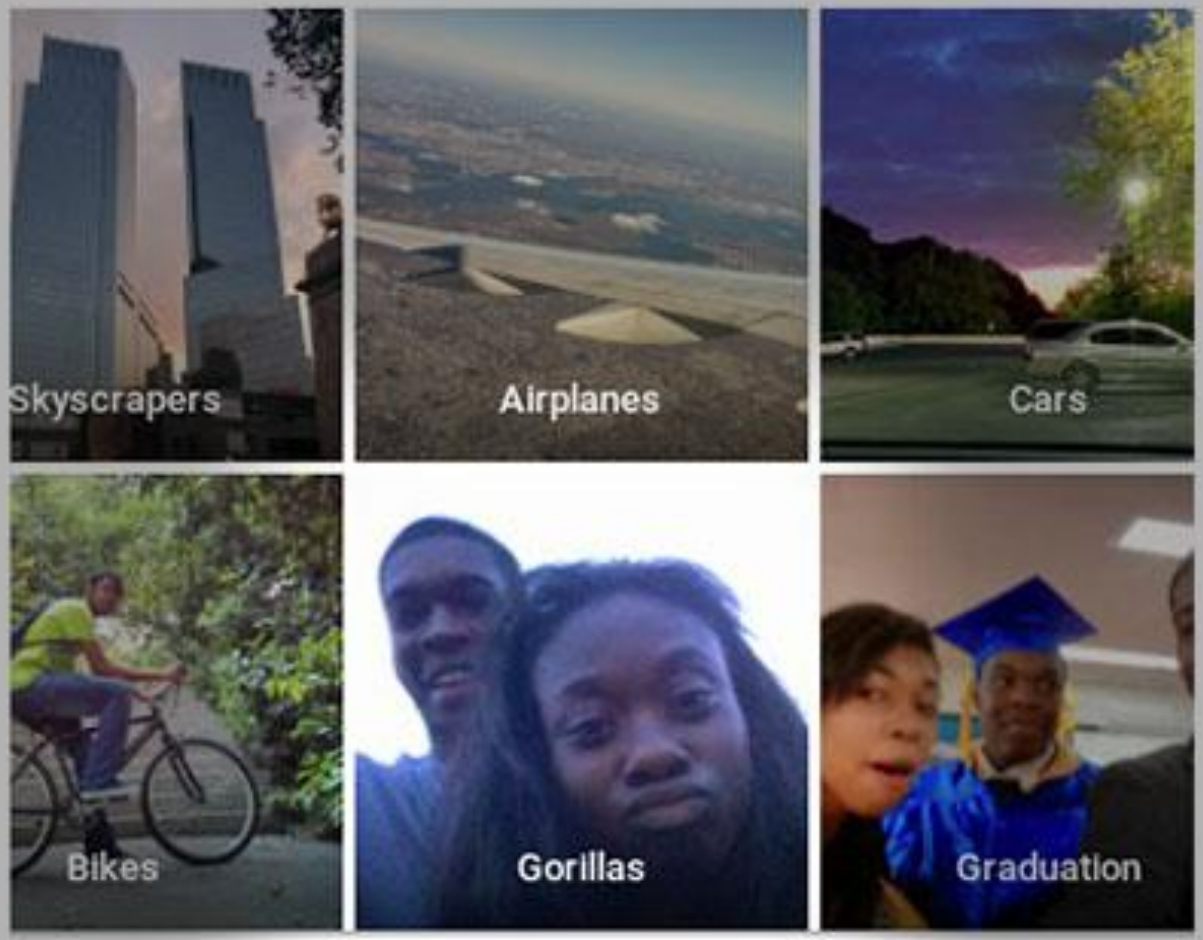
1) systemic problems of discriminatory bias
   remove discriminatory bias

2) decisions with significant consequences lacking an appropriate justification
   provide an appropriate justification

3) objectification of the subject (no dignity)
   empowering design, allowing for challenges

@jackyalcine

Google Photos, y'all f[redacted]d up. My friend's not a gorilla.



RETWEETS FAVORITES

A beauty contest was judged by AI and the robots didn't like dark skin

The first international beauty contest decided by an algorithm has sparked controversy after the results revealed one glaring factor linking the winners

▲ One expert says the results offer 'the perfect illustration of the problem' with machine bias. Photograph: Fabrizio Bensch/Reuters

The first international beauty contest judged by "machines" was supposed to use objective factors such as facial symmetry and wrinkles to identify the most attractive contestants. After Beauty.AI launched this year, roughly 6,000 people from more than 100 countries submitted photos in the hopes that artificial intelligence, supported by complex algorithms, would determine that their faces most closely resembled "human beauty".

But when the results came in, the creators were dismayed to see that there was a glaring factor linking the winners: the robots did not like people with

https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people

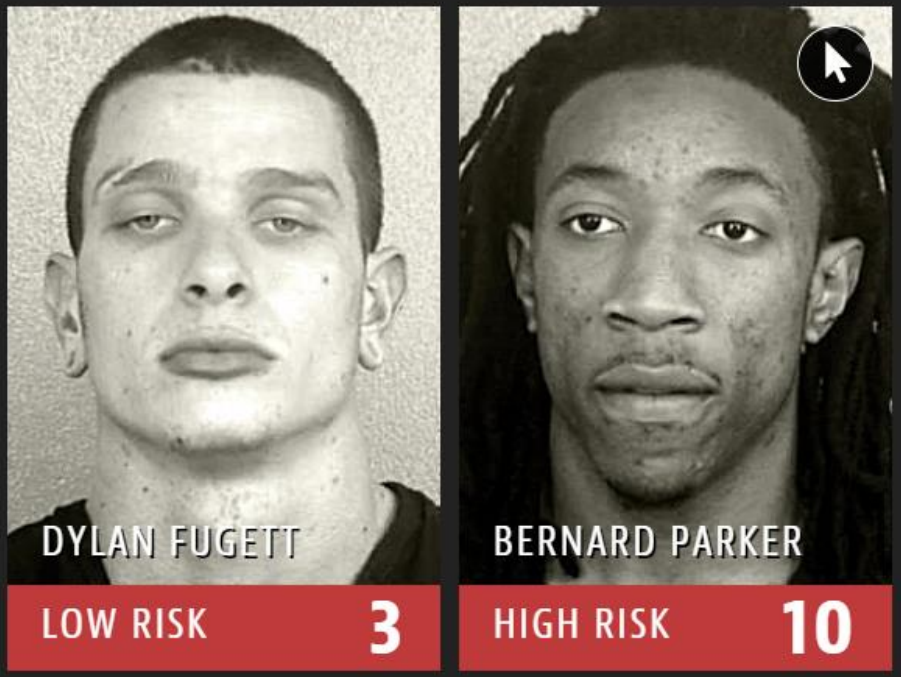
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Two Drug Possession Arrests



Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

https://www.theguardian.com/education/2020/aug/20/england-exams-row-timeline-was-ofqual-warned-of-algorithm-bias#img-1

“It’s polluted data producing polluted results,” said Malkia Cyril, executive director of the Center for Media Justice.

https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people /

Part 2

Automated Decision Making *(and discriminatory ML)*

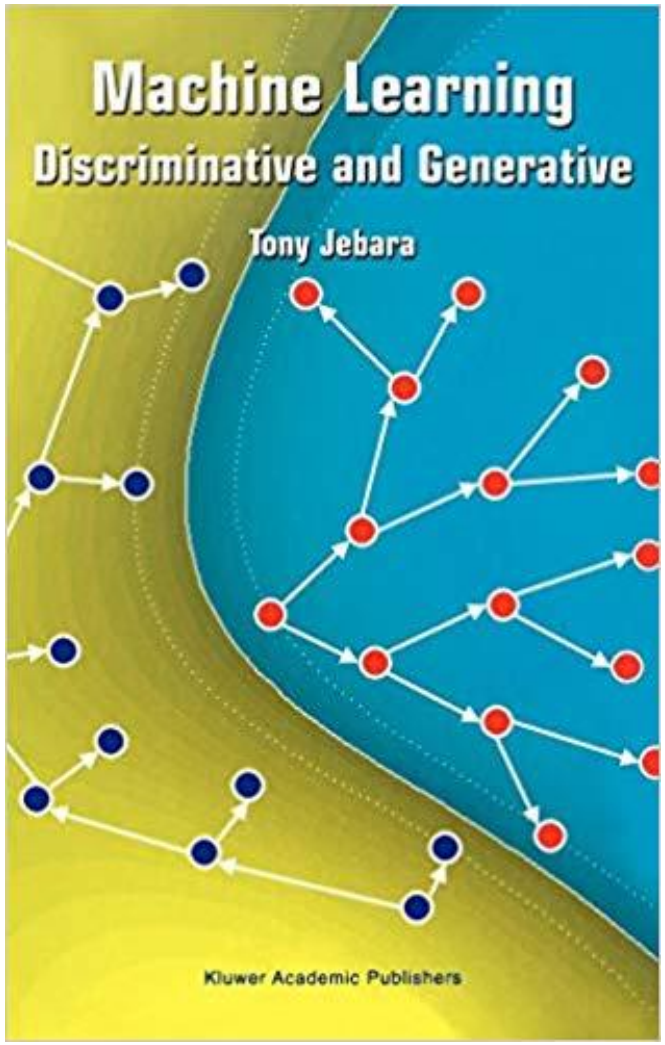
Machine Learning

"Field of study that gives computers the ability to learn without being explicitly programmed".

(Arthur Samuel, 1959)

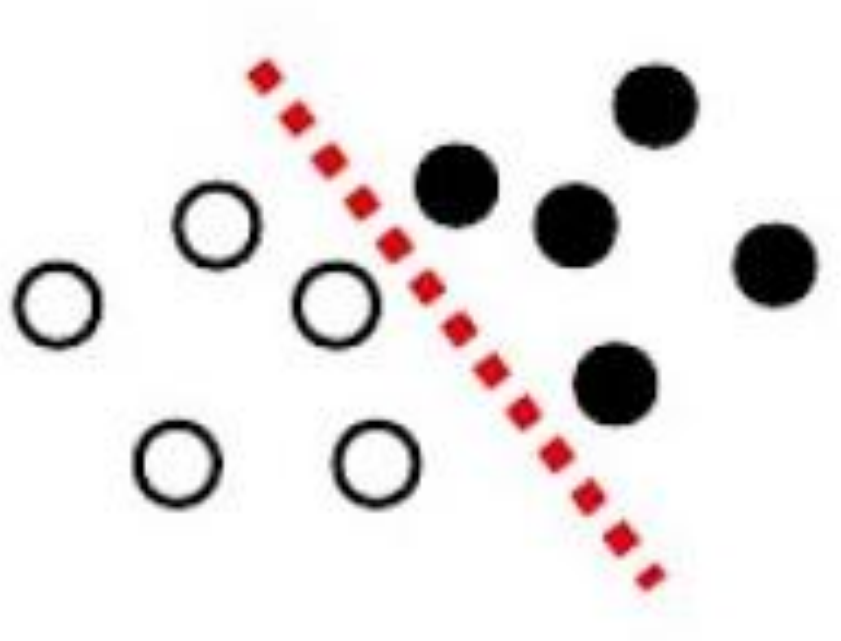
Indirectness:
learning through examples

The 2010s
the success story of Machine Learning


Machine Learning
Discriminative and Generative
Tony Jebara
Kluwer Academic Publishers

2004

The 2010s: the success story of Machine Learning



kaggle Search Competitions Datasets Kernels Discussion Learn •••

Dogs vs. Cats

Create an algorithm to distinguish dogs from cats

215 teams · 5 years ago

Overview Data Kernels Discussion Leaderboard Rules

Data Description

The training archive contains 25,000 images of dogs and cats. Train your algorithm on these
= dog, 0 = cat).

Part 3

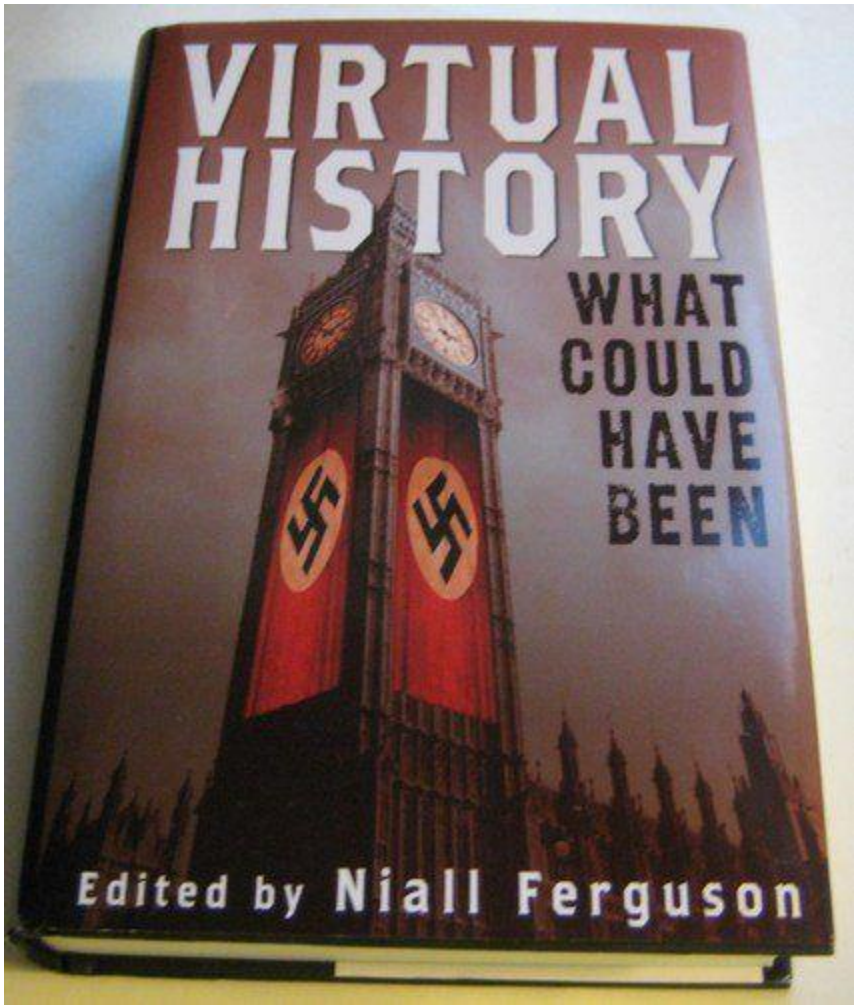
Multiple ways of opening the blackbox

Multiple ways of opening the blackbox

1. Training data
2. The source code
3. The summary-model (building simpler models on top of complex models)
4. The layman's sentence
5. The counterfactual


Computer says no.

"conjecturing on what did not happen, or what might have happened, in order to understand what did happen."



Counterfactual examples for transparency

Counterfactual explanations describe the minimum conditions that would have led to an alternative decision (e.g. a bank loan being approved), without the need to describe the full logic of the algorithm.

Sandra Wachter, Brent Mittelstadt, Chris Russell, (2018) 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR', *Harvard Journal of Law and Technology.*

Synthetic instances for transparency

Juan Hernandez Follow

Apr 4 · 10 min read

Making AI Interpretable with Generative Adversarial Networks

Authors: Juan Hernandez | @damienrj

AI has made tremendous advances in technology, business, and science in the past decades, and this progress continues to accelerate today. Many of our experiences in daily life are influenced by AI and machine learning. For example, music is recommended to us by artificially intelligent systems. Our eligibility for financial services is driven by credit-scoring machine learning models. Automobiles are speedily moving toward full autonomy, and many

Part 4

Generative ML

Variational autoencoders
(Diederik Kingma & Max Welling 2013)

Generative Adversarial Networks
(Ian Goodfellow, 2014)

But while deep-learning AIs can learn to recognize things, they have not been good at creating them. The goal of GANs is to give machines something akin to an imagination.

https://www.technologyreview.com/s/610253/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination/

Karras, T., e.a. (2018), Progressive Growing of GANs for Improved Quality, Stability, and Variation, ICLR 2018. Available at: http://research.nvidia.com/publication/2017-10_Progressive-Growing-of

Deepfakes

Karras, T., e.a. (2018), Progressive Growing of GANs for Improved Quality, Stability, and Variation, ICLR 2018. Available at: http://research.nvidia.com/publication/2017-10_Progressive-Growing-of

https://thispersondoesnotexist.com/

kaggle Search Competitions Datasets Kernels Discussion Learn •••

Dogs vs. Cats

Create an algorithm to distinguish dogs from cats

215 teams · 5 years ago

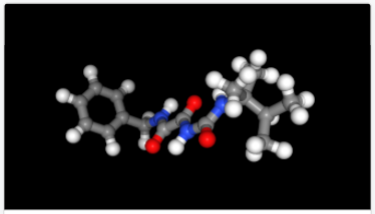
Overview Data Kernels Discussion Leaderboard Rules

Data Description

The training archive contains 25,000 images of dogs and cats. Train your algorithm on these
= dog, 0 = cat).

https://thiscatdoesnotexist.com/


SHAUN S HOT CHRISTMAS



This Chemical Does Not Exist

Who said drug discovery is hard? Just refresh until you find the right chemical. In all seriousness, the fact that this renders a 3D model with the correct bond pairings is impressive.

https://thisxdoesnotexist.com/

This Artwork Does Not Exist

Be inspired by minimalism, realism, post-modernism, pre-modernism, modernism, and ancientism (not actually a thing). No matter your art preferences, you can find it here with enough refreshing.

This Vessel Does Not Exist

Drawing a beautiful parallel between Generators and Discriminators in GANs and apprentices and masters in ceramics, this site demonstrates the beautiful ability of neural networks to replicate the mastery of professionals at yet another craft.

Created by Derek Philip Au.

These Lyrics Do Not Exist

Now we can generate lyrics for a song given a theme or topic. If only we combined this with text-to-speech and a melody-generating model to create completely original songs. Then the music industry would take GANs seriously.

Created by Peter Ranieri.

This Snack Does Not Exist

For when you're feeling especially hungry and creative, this website will remind you that you're more hungry and less creative than you thought.

Created by Ariel Levi, Koby Ofek, et al.

This Word Does Not Exist

Using GPT-2, this website manages to generate words that sound like they should exist, but don't. Great for startup names or baby names (now that the coveted X Æ A-12 is taken).

Created by Thomas Dimson.

This Satire Does Not Exist

It's hard enough to tell what news is real and what news isn't with The Onion and the craziness of the world today. But this site takes it a bit further by generating everything. And yet it all seems so plausible...

Created by Koen Mangelschots.

This Meme Does Not Exist

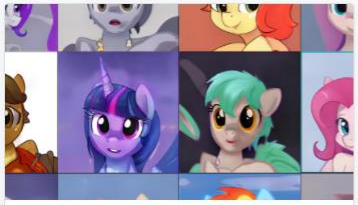
Honestly, if I had to quantify the site with the largest value-add to humanity, this would be in the top three. Memes galore and so much more on every pageload.

This Chair Does Not Exist

Before I explain this site, you may want to take a seat, preferably in one of the chairs generated by this GAN. What's more is that the chairs are all 3D and you can fine tune the "weirdness."

This Foot Does Not Exist

Note that this is an SMS chatbot. You can text it to get pictures of feet. The pictures are animated. The feet are nonexistent. Why would you want to do this? Who knows.



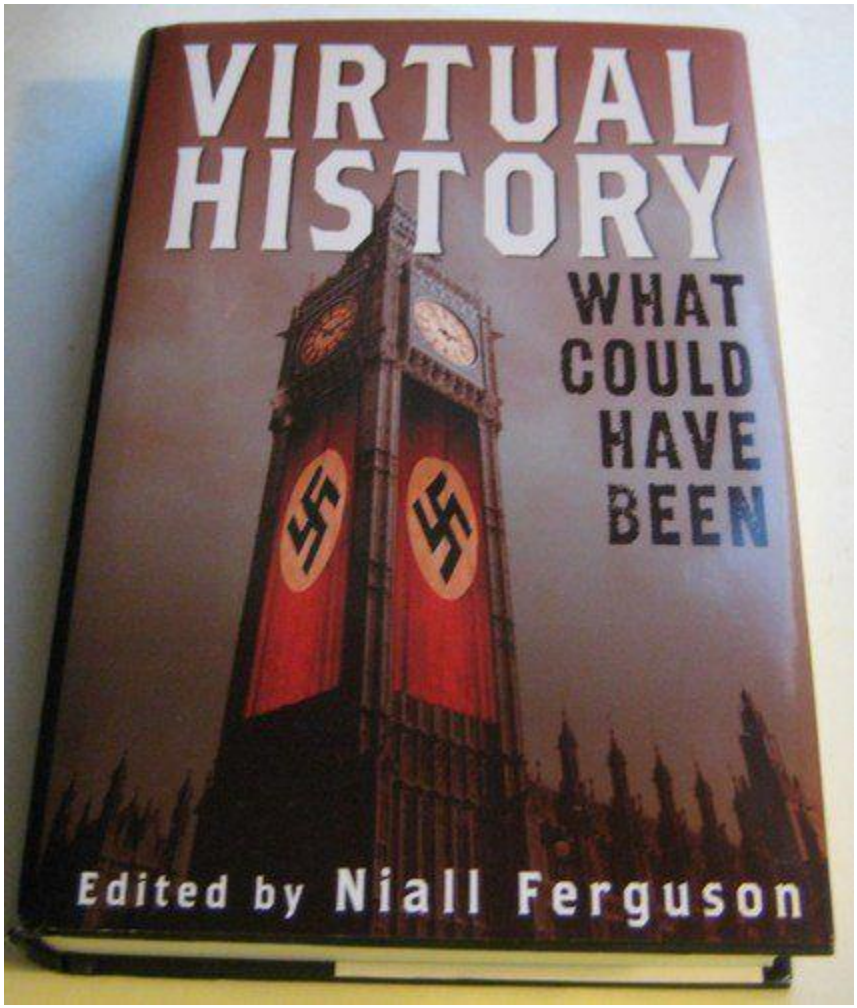
This Pony Does Not Exist

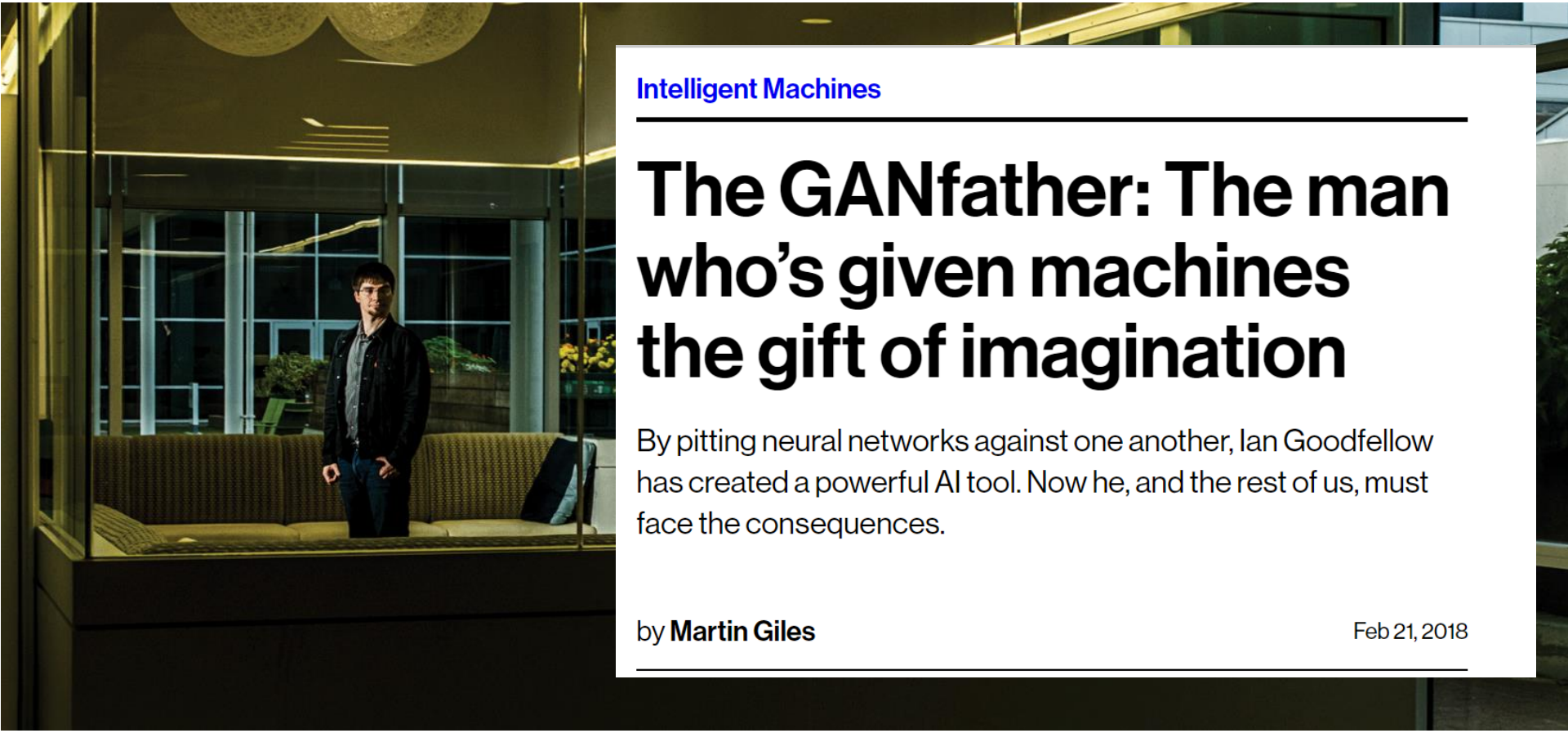
This site was created with user experience in mind. You can pan the generated ponies, zoom in, and even auto-expand on hover. And best of all, each pony is uniquely made for you.

This Automobile Does Not Exist

Legend has it Elon used this website to design the Cybertruck. From there, he made some slight changes to the curvature on the left windows, but that was it.

"conjecturing on what did not happen, or what might have happened, in order to understand what did happen."





Intelligent Machines

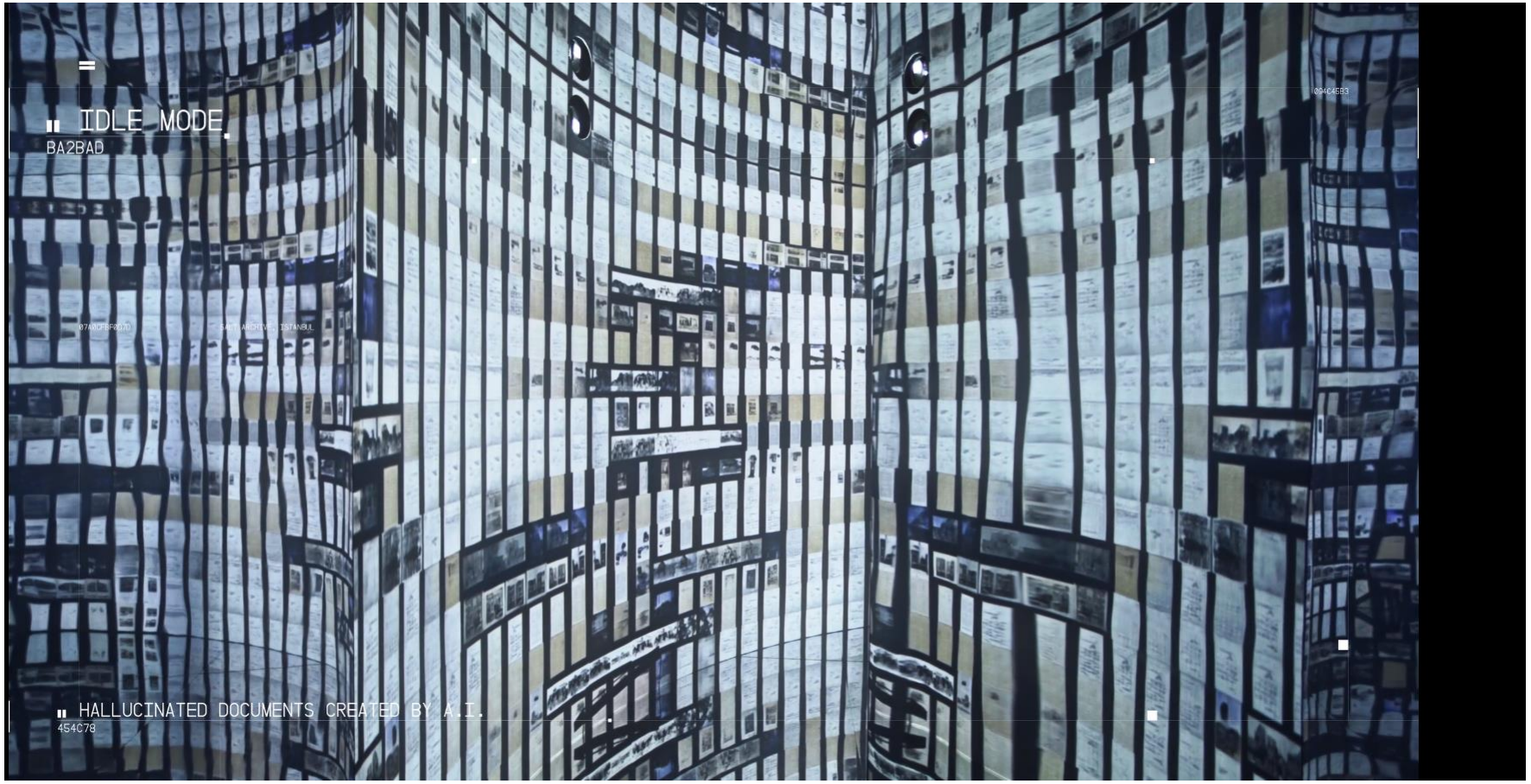
The GANfather: The man who's given machines the gift of imagination

By pitting neural networks against one another, Ian Goodfellow has created a powerful AI tool. Now he, and the rest of us, must face the consequences.

by **Martin Giles** Feb 21, 2018

https://www.technologyreview.com/s/610253/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination/

Archive Dreaming – Refik Anadol (2017) hallucinated documents created by AI



Part 5

Counterfactuals – the good

Transparency

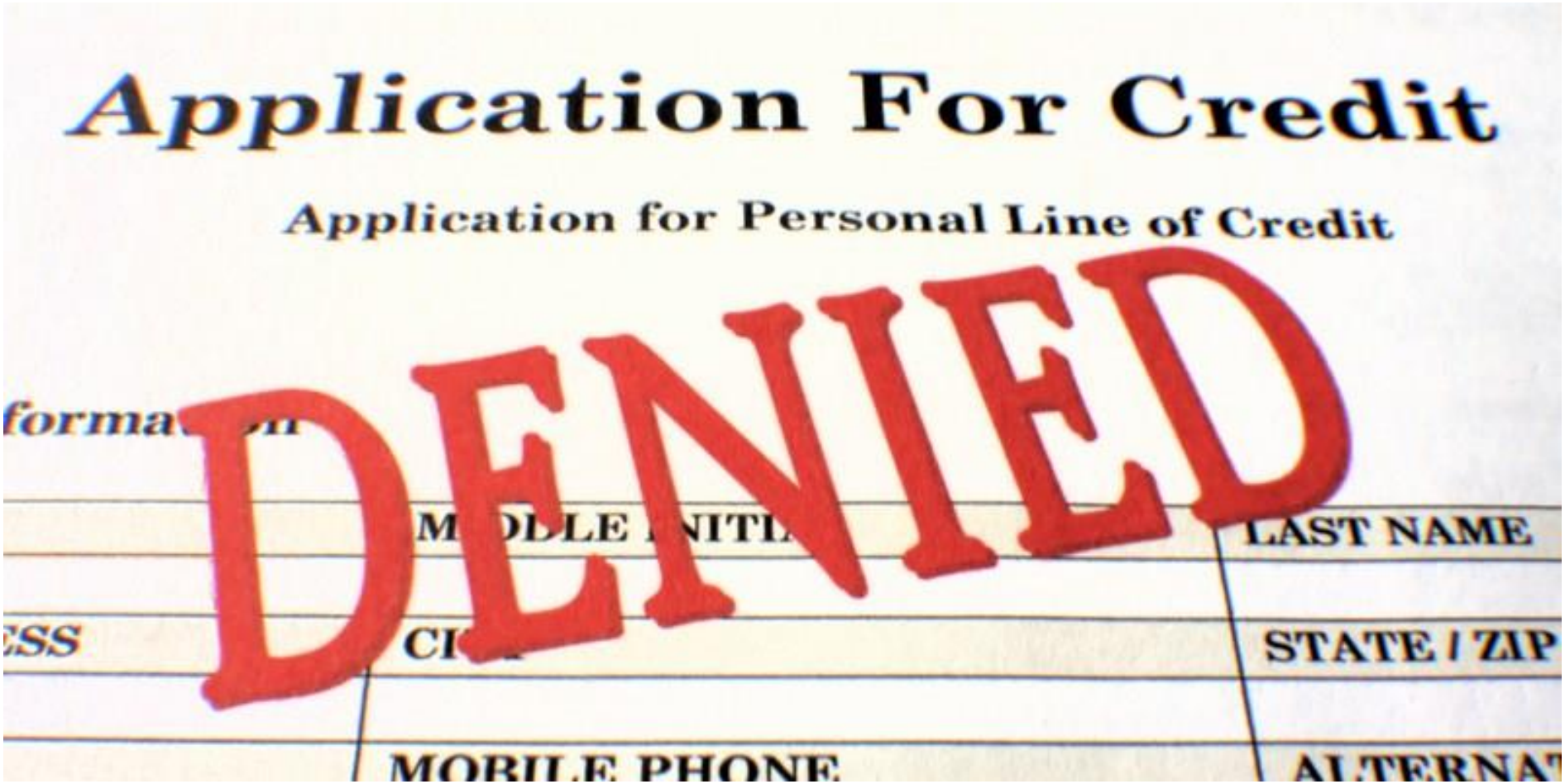
https://www.pikrepo.com/fjfpo/sun-rays-on-the-window

Part 6

Counterfactuals – the bad

Rashomon effect?


Computer says no.


Application For Credit
Application for Personal Line of Credit
DENIED
LAST NAME
STATE / ZIP
MOBILE PHONE

Give a full list of counterfactuals?

- Information overload (not very actionable)
- Might reveal too much (trade secrets, gaming the system?)

Select the best of?

- Remove everything that is not actionable?
- What about interactions? *Following counterfactual advise might activate some other factor that is adverse (but that you did not know about)*

Calculus is not just calculus?

- Distance – how to compare scales?
- What outcome?

Part 7

Concluding thoughts


Computer says no.

Holiday rental housing fraud risk

From 1 July 2020, a pilot will be carried out for six months with an algorithm that supports the employees of the department of Surveillance & Enforcement in their investigation of the reports made concerning possible illegal holiday rentals. The algorithm helps prioritize the reports so that the limited enforcement capacity can be used efficiently and effectively. By analyzing the data of related housing fraud cases of the past 5 years, it calculates the probability of an illegal holiday rental situation on the reported address.