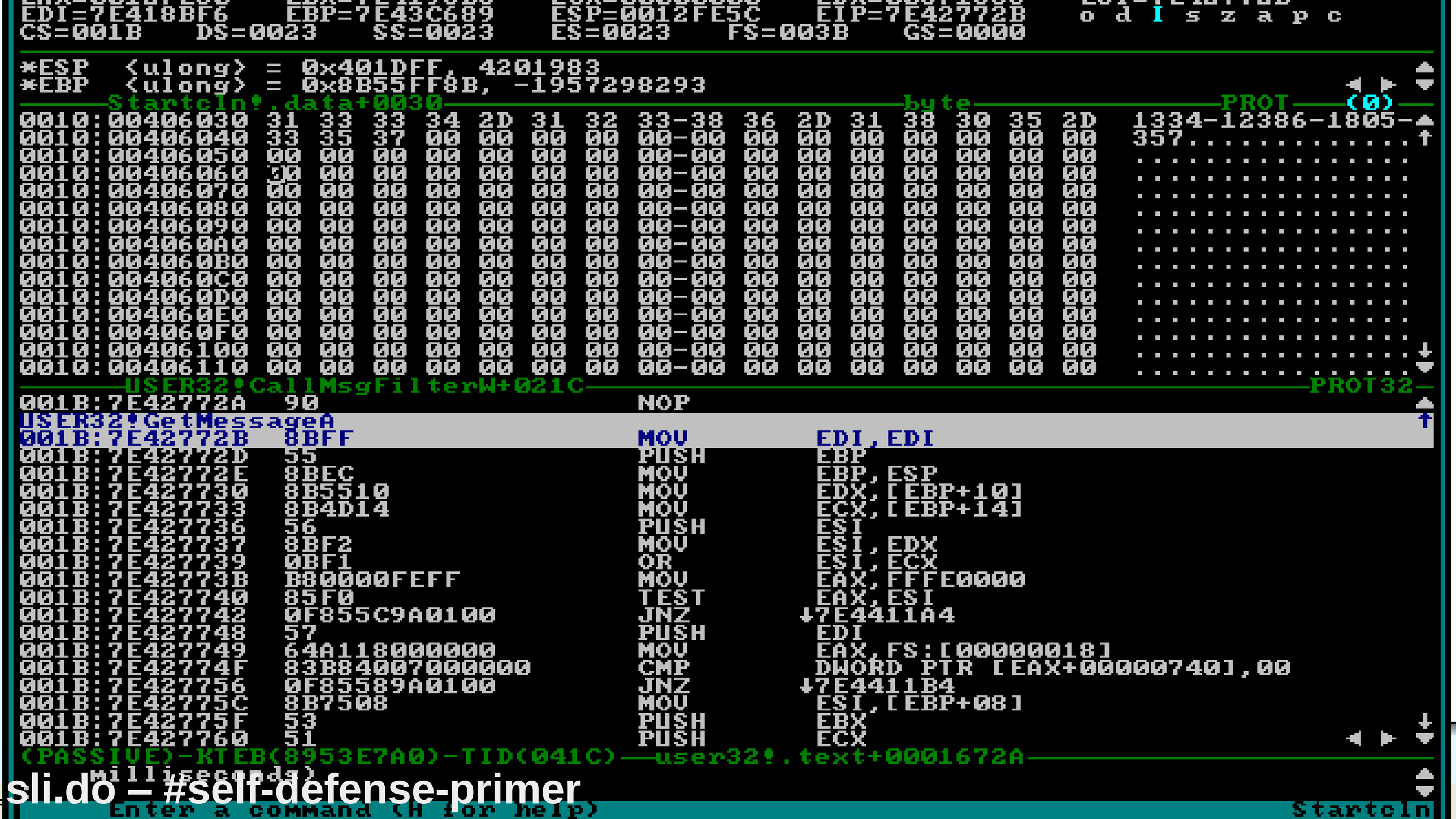
Julien Cornebise, Ph.D.

2020-12-18 ML Collective / DL Classics and Trends

AI FOR GOOD AND ETHICS-WASHING
A SELF-DEFENSE PRIMER

UCL



sli.do — #self-defense-primer

We have the following limits, as $N \to \infty$.

Theorem

1. *Asymptotic behavior of KLD-based criterion:*

$$\left| d_{\mathrm{KL}}(\mu_N^{\mathrm{aux}} || \pi_N^{\mathrm{aux}}) - \mathcal{E}(\{\tilde{\omega}_{N,i}\}_{i=1}^{\tilde{M}_N}) \right| \xrightarrow{\mathbb{P}} 0 \, .$$

In addition,

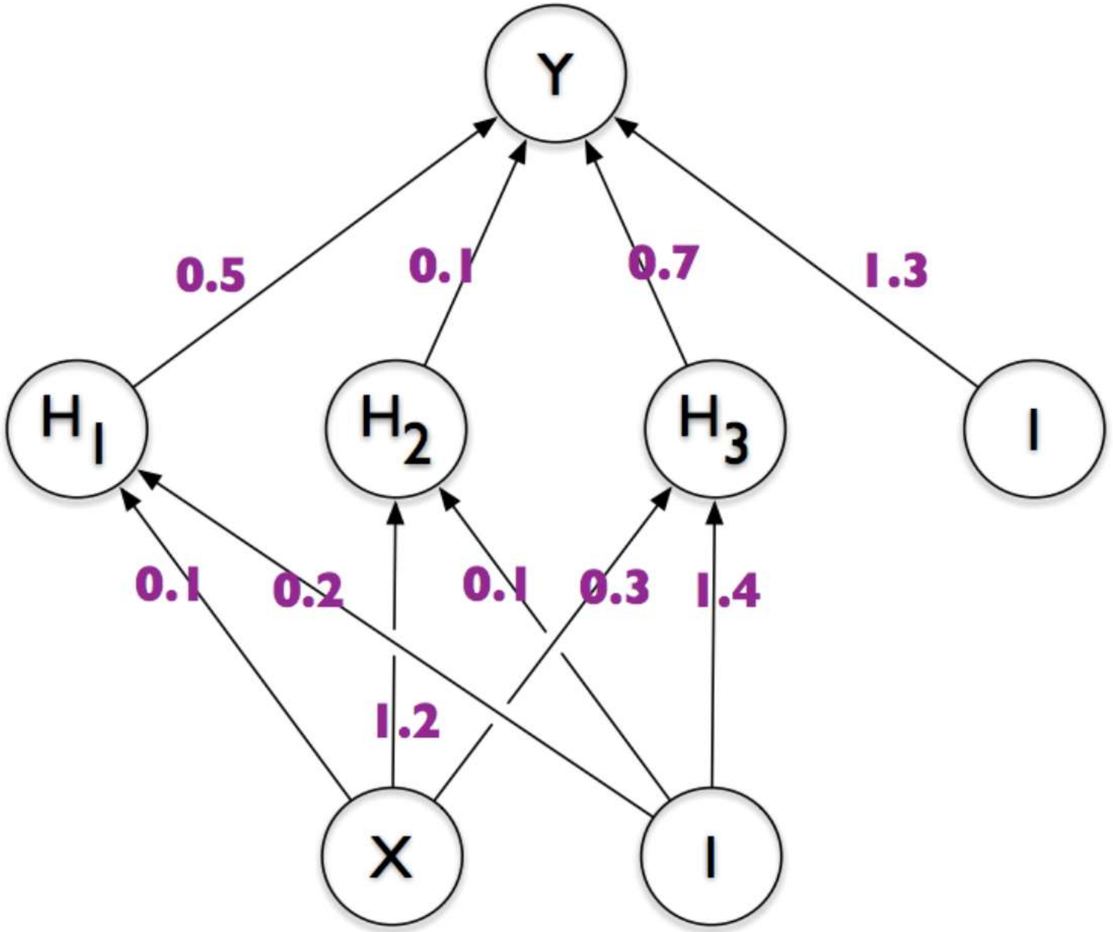
$$d_{\mathrm{KL}}(\mu_N^{\mathrm{aux}} || \pi_N^{\mathrm{aux}}) \xrightarrow{\mathbb{P}} \eta_{\mathrm{KL}}(\Psi) \triangleq \frac{\nu \otimes L}{\nu L(\tilde{\Xi})} \{ \log[\Phi \frac{\nu(\Psi)}{\nu L(\tilde{\Xi})}] \} \, ,$$

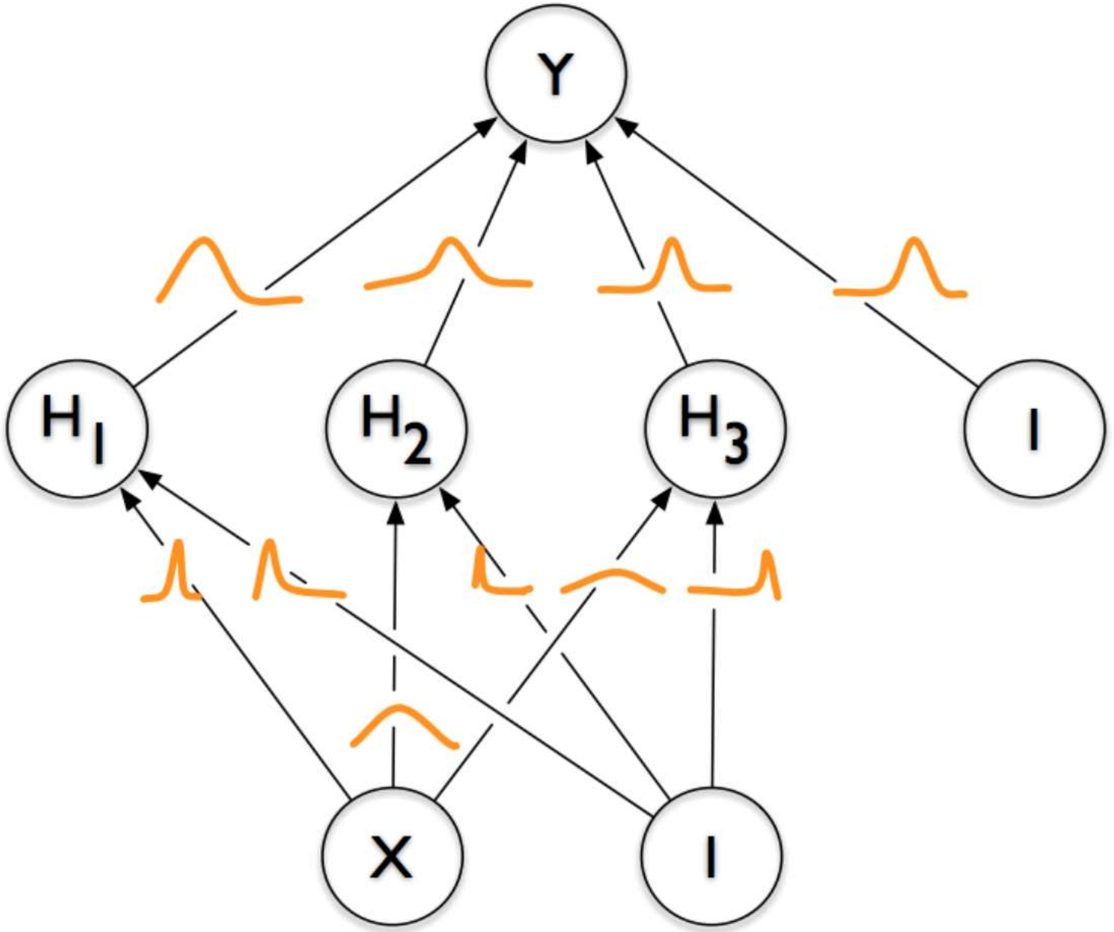
2. *Asymptotic behavior of CSD-based criterion:*

$$\left| d_{\chi^2}(\mu_N^{\mathrm{aux}} || \pi_N^{\mathrm{aux}}) - \mathrm{CV}^2(\{\tilde{\omega}_{N,i}\}_{i=1}^{\tilde{M}_N}) \right| \xrightarrow{\mathbb{P}} 0 \, .$$

In addition,

$$d_{\chi^2}(\mu_N^{\mathrm{aux}} || \pi_N^{\mathrm{aux}}) \xrightarrow{\mathbb{P}} \eta_{\chi^2}(\Psi) \triangleq \frac{\nu(\Psi)}{\nu L(\tilde{\Xi})} \times \frac{\nu \otimes L}{\nu L(\tilde{\Xi})}(\Phi) - 1 \, .$$


Y
0.5
0.1
0.7
1.3
H1
H2
H3
1
0.1
0.2
0.1
0.3
1.4
1.2
X
1


Y
H1
H2
H3
1
X
1

theguardian

home › tech | UK | world | politics | sport | football | opinion | culture | all

Google

Google DeepMir on AI-based radi

Google's AI research arm is partnerin radiotherapists by using machine lea

The DeepMind project will use anonymisec Juice/REX/Shutterstock

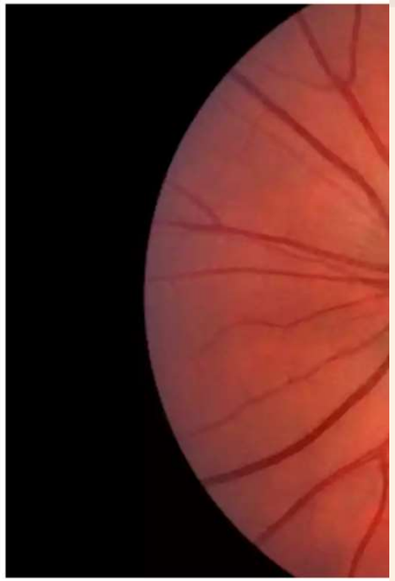
theguardian

home › tech | UK | world

Google

Google DeepMir machine learnin

'Deep learning' research company wi network to identify early signs of deg



Retinal scans are produced rapidly, but requ / Alamy Stock Photo/Alamy Stock Photo

Sign In | Subscribe

FINANCIAL TIMES myFT

Get a fresh start. Choose your FT trial

Latest on Artificial intelligence

Marcus du Sautoy on creative AI

Daimler T US self-d

Artificial intelligence + Add to myFT

Artificial intelligence gro commercial product

Device can diagnose a range of eye di specialists

Save

Madhumita Murgia in London MARCH 31, 2019

DeepMind, the British artificial intel device that can diagnose complex eye Alphabet-owned company's first mec

In a live demonstration this month o examined publicly, DeepMind perfor eye. The scan was analysed by a set o urgency score and a detailed diagnos

WIRED

Health

DeepMind's new AI predicts kidney injury two days before it happens

New research from the Google-owned firm hints that AI may be a better way of assessing if someone is at risk of acute kidney injury. But there are still questions about how it handles patient data

By **MATT REYNOLDS**

31 Jul 2019

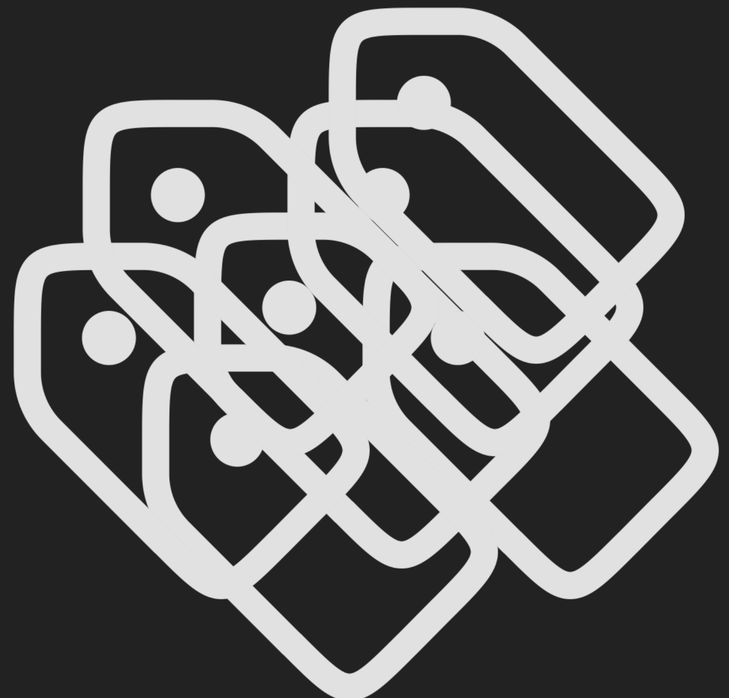
UCL

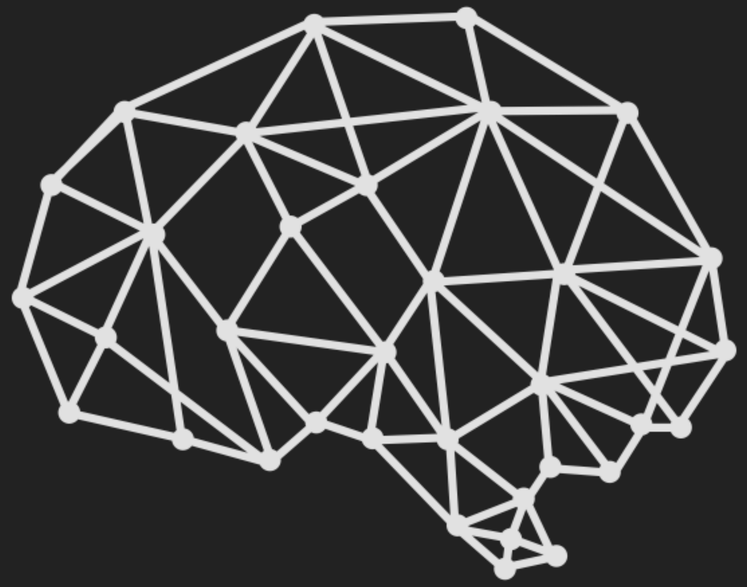
AMNESTY INTERNATIONAL

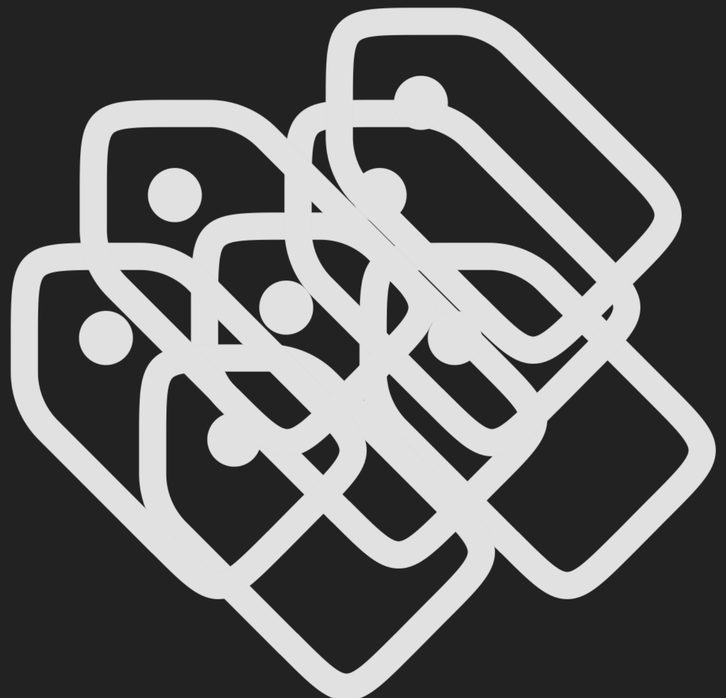
@deworrall92 @milena_iul

2017

DECODE DARFUR







AMNESTY INTERNATIONAL

DECODE DARFUR

FORUM ACCOUNT @GMAIL.COM

Well done! You identified the human presence in this tile

NEXT

SELECT TITLES CONTAINING HUMAN PRESENCE

FLAG HELP

NEW ACHIEVEMENT

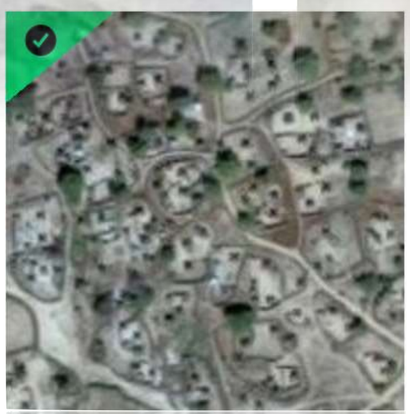
YOU DECODED 50 SQUARE KM

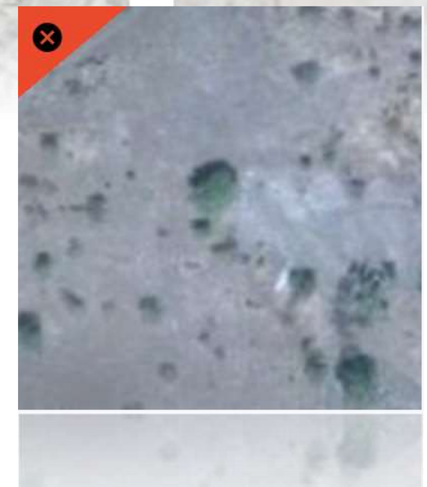
CONTINUE →

25	149	1659
SQUARE KM THIS SESSION	SQUARE KM ALL SESSIONS	SQUARE KM ALL VOLUNTEERS







<Video demo removed due to potentially sensitive content>

AMNESTY INTERNATIONAL

@alkalait

@laure_delisle

@milena_iul

@SnazzyAzzy

@sherifea

@ArchydeB

2018 TROLL PATROL: STUDYING ABUSE AGAINST WOMEN ON TWITTER

TRIGGER WARNING

[redacted] · Jul 13

Replying to @HackneyAbbott

this **fat** retarded black **bitch** thinks you should be forced to feed and house a bunch of violence foreign invaders. i strongly disagree.

[redacted] · Jul 13

@HackneyAbbott Piss off you disgusting useless **fat bitch!** You're a parasite alien looking to silence native people for your power.

[redacted] · Jul 14

Replying to @HackneyAbbott

You forgot "fat disgusting obese chicken-loving **nigger**"

[redacted] · Jul 14

Replying to @HackneyAbbott

An acid attack would probably make your face look better you fat **nigger**

[redacted]

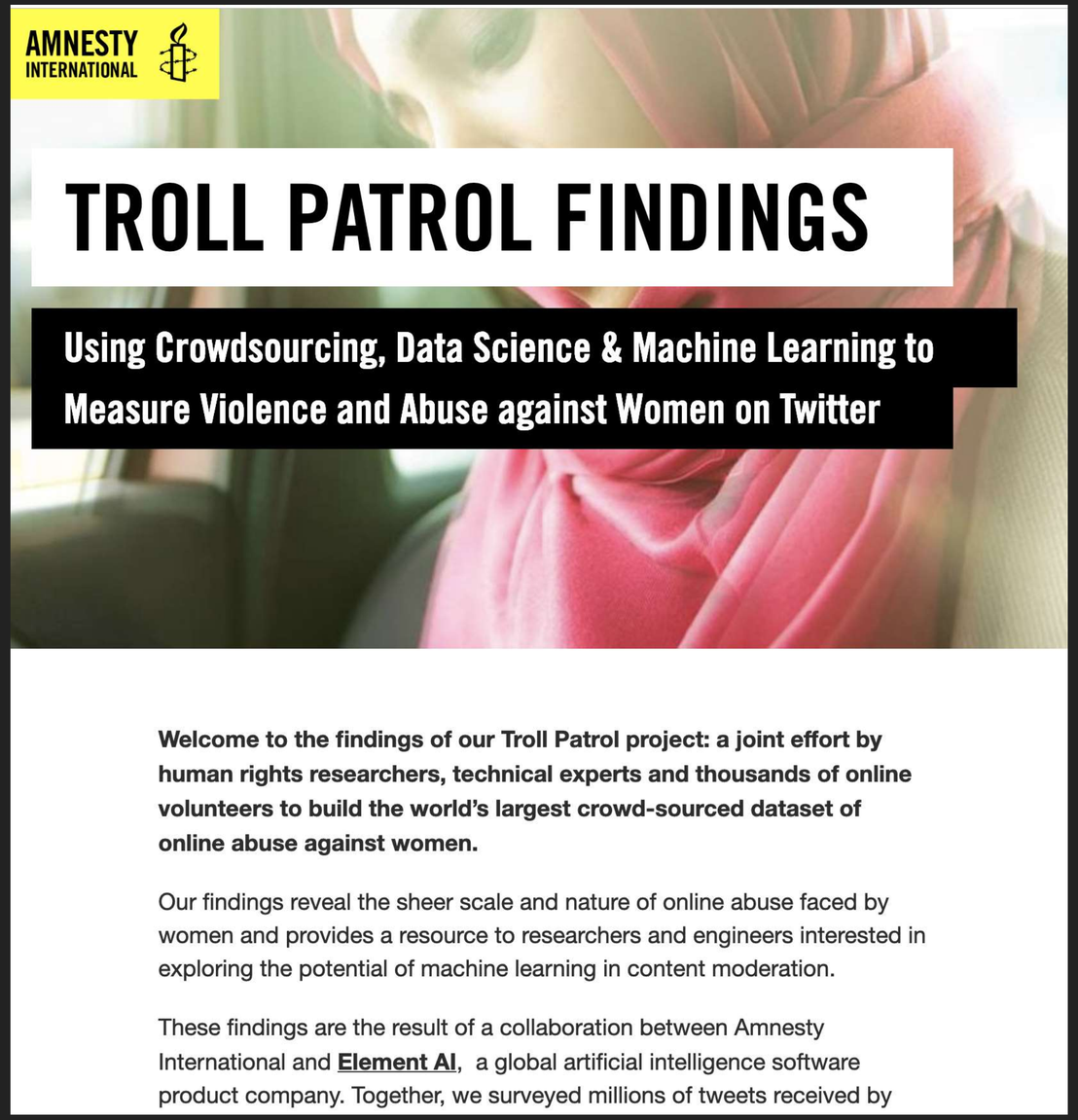
Someone jo cox Anna sourby please

02/12/2016, 07:41 from Poplar, London

https://decoders.amnesty.org/projects/troll-patrol/findings

STUDY: ML TO HUMANS TO ML TO HUMANS

- Build a dataset with ML prefilter
- Importance sampling to
  - reduce variance
  - drive engagement
- Human Crowdsourcing
- Data Analysis with Bootstrap
- Train Deep model and benchmark
- Carefully and expertly report and go public!
  - Very sensitive topic, human rights expertise!



Bloomberg

Technology

Twitter Is 'Toxic Place for Women,' Says Amnesty International

Together with Element AI, the human rights watchdog quantified what women on Twitter have known for a long time

TIME

U.S. POLITICS WORLD TECH ENTERTAINM

WORLD • SOCIAL MEDIA

'A Toxic Place for Women.' A New Study Reveals the Scale of Abuse on Twitter

MIT Technology Review

Topics+

Intelligent Machines

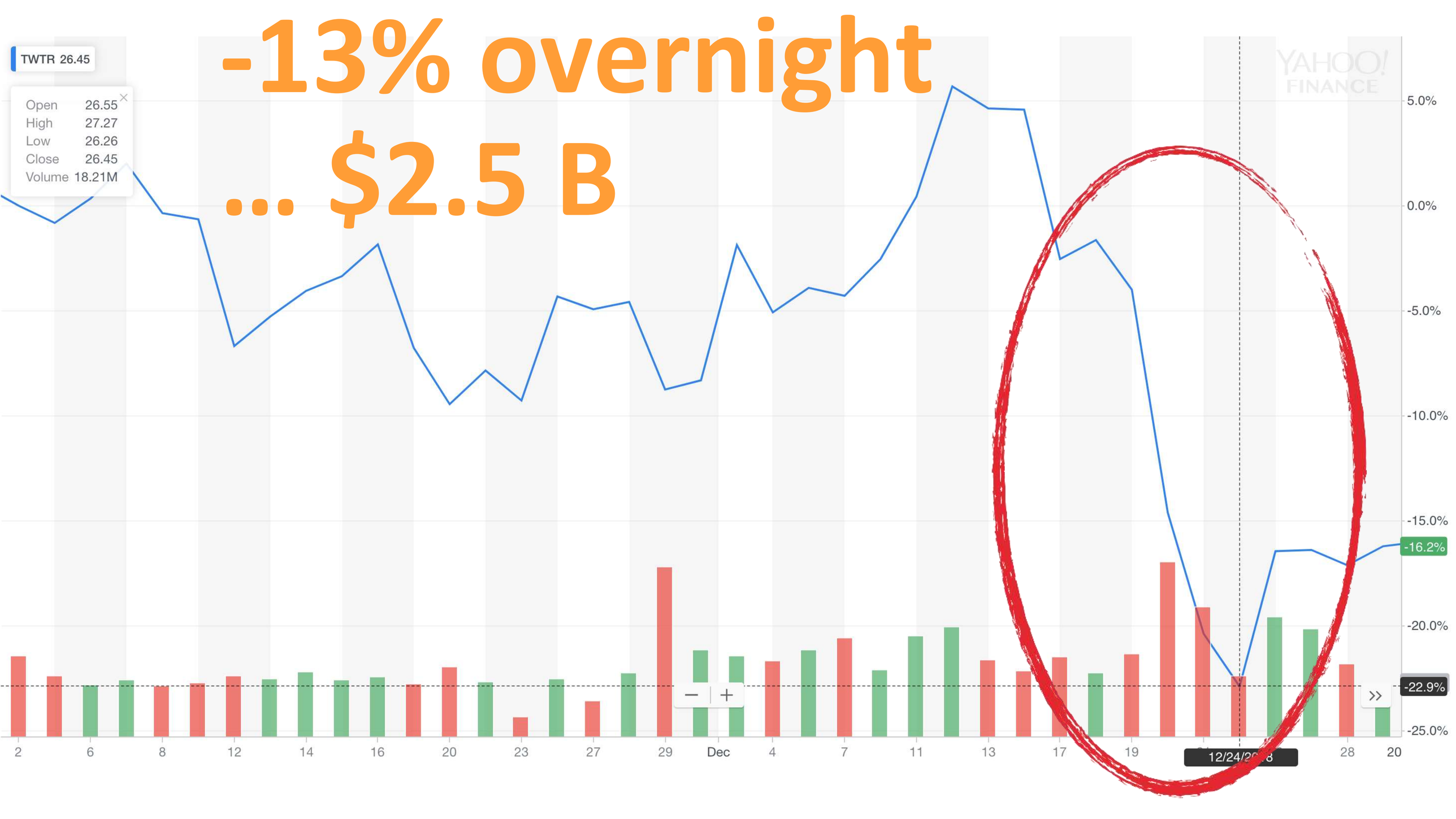
Female black journalists and politicians get sent an abusive tweet every 30 seconds

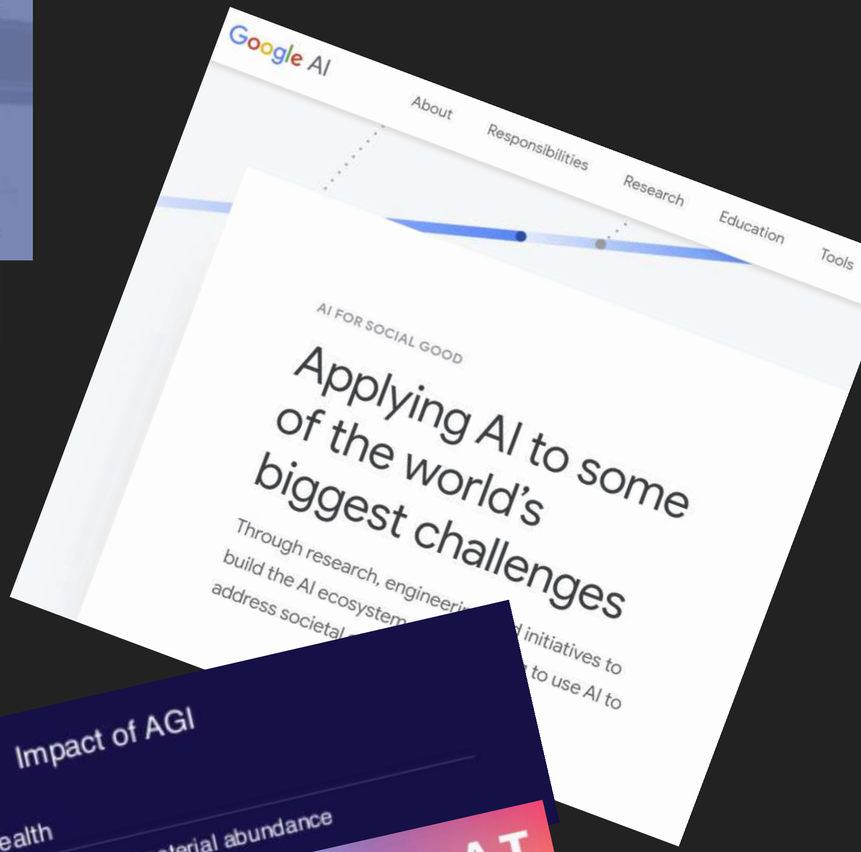
WIRED

EMILY DREYFUSS SECURITY 12.18.18 12:00 AM

TWITTER IS INDEED TOXIC FOR WOMEN, AMNESTY REPORT SAYS

	A	B
1	Publication	Link
2	Wired	https://www.wired.com/story/amnesty-report-twitter-abuse-women/
3	FT	https://www.ft.com/content/1091e116-0324-11e9-99df-6183d3002ee1
4	Gizmodo	https://gizmodo.com/we-finally-have-some-hard-data-on-how-much-twitter-suck-1831156963
5	Shropshire Star	https://www.shropshirestar.com/news/uk-news/2018/12/15/diane-abbott-condemns-twitter-over-thousands-of-racist-and-misogynist-tweets/
6	Montreal Gazette	https://montrealgazette.com/business/local-business/element-ai-partners-with-amnesty-international-to-study-online-abuse
7	Guardian	https://www.theguardian.com/politics/2018/dec/15/diane-abbott-calls-for-twitter-to-clamp-down-on-hate-speech
8	Bloomberg	https://www.bloomberg.com/news/articles/2018-12-15/twitter-is-a-toxic-place-for-women-says-amnesty-international
9	Le Devoir	https://www.ledevoir.com/societe/543526/quand-les-trolls-s-en-prennent-aux-femmes-sur-twitter
10	Engadget	https://www.engadget.com/2018/12/15/women-are-abused-every-30-seconds-on-twitter/
11	Time	http://time.com/5452386/twitter-online-abuse-women-amnesty-international-study/
12	Packt	https://hub.packtpub.com/troll-patrol-report-amnesty-international-and-element-ai-use-machine-learning-to-understand-online-abuse-against-women/
13	Jakarta Post	https://www.thejakartapost.com/news/2018/12/15/twitter-is-toxic-place-for-women-says-amnesty-international.html
14	NewsTalk	https://www.newstalk.com/Study-reveals-scale-of-online-abuse-against-women
15	Hindustan Times	https://m.hindustantimes.com/world-news/twitter-is-toxic-place-for-women-says-amnesty-international/story-7yzTcNx1aAMJv4OxoORSVP.html
16	La Presse	https://www.lapresse.ca/techno/reseaux-sociaux/201812/17/01-5226405-les-femmes-noires-premieres-cibles-des-injures-sur-twitter.php
17	Business Standard	https://www.business-standard.com/article/international/amnesty-international-explains-how-twitter-is-a-toxic-place-for-women-118121500254_1.html
18	The Independent	https://www.independent.co.uk/life-style/women/twitter-women-sexist-amnesty-international-diane-abbott-report-a8656436.html
19	Business Today Kenya	https://businesstoday.co.ke/women-abused-twitter-every-30-seconds-new-study-reveals/
20	Press Gazette	https://www.pressgazette.co.uk/1-in-every-14-tweets-directed-at-women-journalists-is-abusive-or-problematic-new-online-abuse-study-shows/
21	Mashable	https://mashable.com/article/amnesty-study-twitter-abuse-women/?europe=true#WIAsyDDDqqM
22	Fast Company	https://www.fastcompany.com/90265064/a-woman-on-twitter-is-abused-every-30-seconds
23	MIT Tech Review	https://www.technologyreview.com/s/612694/female-black-journalists-and-politicians-get-sent-an-abusive-tweet-every-30-seconds/
24	Evening Standard	https://www.standard.co.uk/tech/online-abuse-twitter-amnesty-international-element-ai-a4020676.html
25	The Voice	http://www.voice-online.co.uk/article/diane-abbott-urges-twitter-address-abusive-tweets
26	Mail and Guardian (South Africa)	https://mg.co.za/article/2018-12-15-women-across-political-spectrum-receive-twitter-abuse-study
27	Daily Dot	https://www.dailydot.com/debug/black-women-abuse-white-women-twitter-report/
28	Nu (Netherlands)	https://www.nu.nl/internet/5637205/amnesty-twitter-beerput-vrouwen.html
29	NRC (Netherlands)	https://www.nrc.nl/nieuws/2018/12/15/vrouwelijk-journalisten-en-politici-op-twitter-eens-per-30-seconden-beledigd-a3061044
30	Talk Radio	https://talkradio.co.uk/news/diane-abbott-urges-twitter-crackdown-abusive-tweets-after-study-abuse-women-social-media
31	All Africa	https://allafrica.com/stories/201812180615.html
32	Westfälische Rundschau	https://www.wr.de/panorama/amnesty-studie-warnt-frauen-vor-twitter-und-ausmass-an-hass-id216039215.html
33	Forbes	https://www.forbes.com/sites/fruzsinaeordogh/2018/12/15/troll-patrol-robots-prove-women-harassed-on-twitter/
34	The Journal (Ireland)	https://www.thejournal.ie/twitter-abuse-women-of-colour-4400085-Dec2018/?amp=1
35	Fortune	http://fortune.com/2018/12/18/twitter-women-amnesty-international/
36	Le Monde	https://www.lemonde.fr/pixels/article/2018/12/15/harcelement-des-femmes-sur-twitter-amnesty-enfonce-le-clou_5399413_4408996.html
37	sky news	https://news.sky.com/story/amp/diane-abbott-twitter-must-act-over-abuse-targeting-black-women-11584785
38	The Telegraph	https://www.telegraph.co.uk/politics/2018/12/15/diane-abbott-urges-twitter-tackle-racist-misogynist-abuse/amp/
39	Evening Standard (again)	https://www.standard.co.uk/news/uk/diane-abbott-condemns-twitter-over-thousands-of-racist-and-misogynistic-tweets-appearing-on-platform-a4018996.html?amp
40	CNBC	https://www.cnbc.com/amp/2018/12/15/women-are-abused-every-30-seconds-on-twitter-new-research-finds-.html
41	Washington Post	https://www.washingtonpost.com/technology/2018/12/15/social-media-crisis-after-senate-reports-facebook-apologizes-russia-leaks-protest-trump-attacks/?noredir
42	The Jewish Chronicle	https://www.thejc.com/news/uk-news/amnesty-international-criticised-for-not-including-antisemitism-in-report-on-online-abuse-1.474093
43	BBC News Hour Radio	Tanya, live interview
44	BBC World News	Live TV, Azmina interviewed on stage
45	Channel 5	https://twitter.com/5_news/status/1075102933125183257?s=12
46	Euronews TV	Azmina recorded interviewed over Skype
47	TRT Radio	https://www.youtube.com/watch?v=QZGYnXJn5CA
48	Quartz	https://qz.com/1495774/the-number-of-abusive-or-problematic-tweets-directed-at-women/
49	The Hill	https://thehill.com/policy/technology/421573-study-female-politicians-journalists-abused-every-30-seconds-on-twitter?amp
50	Huffington Post UK	https://m.huffingtonpost.co.uk/entry/shadow-home-secretary-diane-abbott-urges-twitter-to-do-more-to-tackle-online-abuse-of-women_uk_5c19f743e4b02d2cae8fd4
51	NY Mag	http://nymag.com/intelligencer/2018/12/study-finds-black-women-on-twitter-face-more-abuse.html
52	BuzzFeed's @AM2DM	Milena in streaming on 19th
53	KNX 1070 News Radio in Los Angeles	Milena recorded interview
54	Irish Radio	Milena recorded interview
55	CBC Daybreak	Julien recorded interview, airing on Thursday at 7:40am MTL time
56	Fox News	https://www.foxnews.com/tech/female-journalists-and-politicians-were-abused-every-30-seconds-on-twitter-in-2017-new-report-charges
57	Daily Times (Pakistan)	https://dailytimes.com.pk/334577/twitter-toxic-place-for-women-amnesty-international/
58	Motherboard	https://motherboard.vice.com/en_us/article/ye74zb/twitter-harassment-of-woman-journalists-and-politicians-happens-every-30-seconds-amnesty-study-finds
59	The Register	https://www.theregister.co.uk/2018/12/15/twitter_toxic_for_women/
60	Market Watch	https://www.marketwatch.com/story/amnesty-international-teams-with-ai-firm-to-quantify-abuse-targeting-women-of-color-on-twitter-2018-12-15
61	Daily Mail	https://www.dailymail.co.uk/news/article-6593345/New-Amnesty-international-study-reveals-women-receive-one-abusive-tweet-30-seconds.html
62	Elle	https://www.elle.com/culture/career-politics/a25615005/amnesty-international-study-troll-patrol-twitter-toxic-women/
63	LA Times	https://www.latimes.com/business/technology/la-fi-tn-facebook-twitter-20181215-story.html
64	NBC	https://www.nbcnews.com/news/world/female-journalists-politicians-received-abusive-hostile-tweets-every-30-seconds-n949501
65	Euronews	https://www.euronews.com/2018/12/15/female-journalists-politicians-received-abusive-hostile-tweets-every-30-seconds-n949501
66	CNET	https://www.cnet.com/news/study-twitter-is-a-toxic-place-for-women/
67	Axios	https://www.axios.com/twitter-toxic-place-female-politicians-journalists-7f2d44e1-7c93-4aea-992b-fa105a83520c.html
68	Japan Times	https://www.japantimes.co.jp/news/2018/12/19/world/female-politicians-journalists-color-u-k-u-s-face-rampant-online-abuse-amnesty/#.XBn6vRanyEc
69	Forbes 2	https://www.forbes.com/sites/ronbhn/2018/12/15/amnesty-international-says-twitter-trolling-is-a-human-rights-abuse/#5a5d4002346c


-13% overnight
… $2.5 B
TWTR 26.45
Open 26.55
High 27.27
Low 26.26
Close 26.45
Volume 18.21M
YAHOO! FINANCE
5.0%
0.0%
-5.0%
-10.0%
-15.0%
-16.2%
-20.0%
-22.9%
-25.0%
2
6
8
12
14
16
20
23
27
29
Dec
4
7
11
13
17
19
28
20



MANY INITIATIVES

“TECHNOLOGY – NO MATTER HOW WELL DESIGNED – IS ONLY A MAGNIFIER OF HUMAN INTENT AND CAPACITY.”

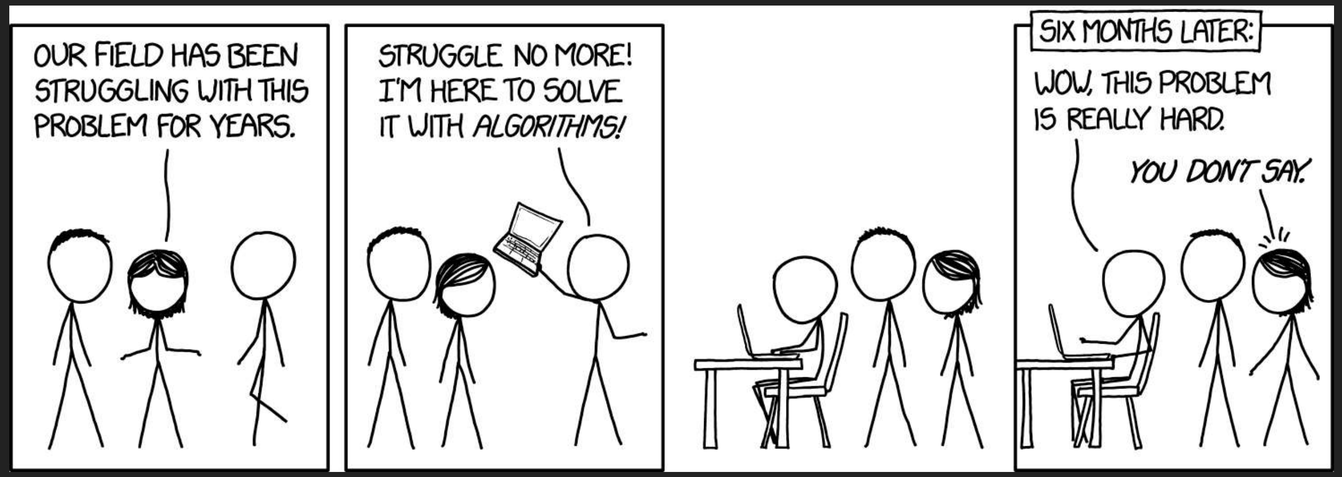
Kentaro Toyama (2015)

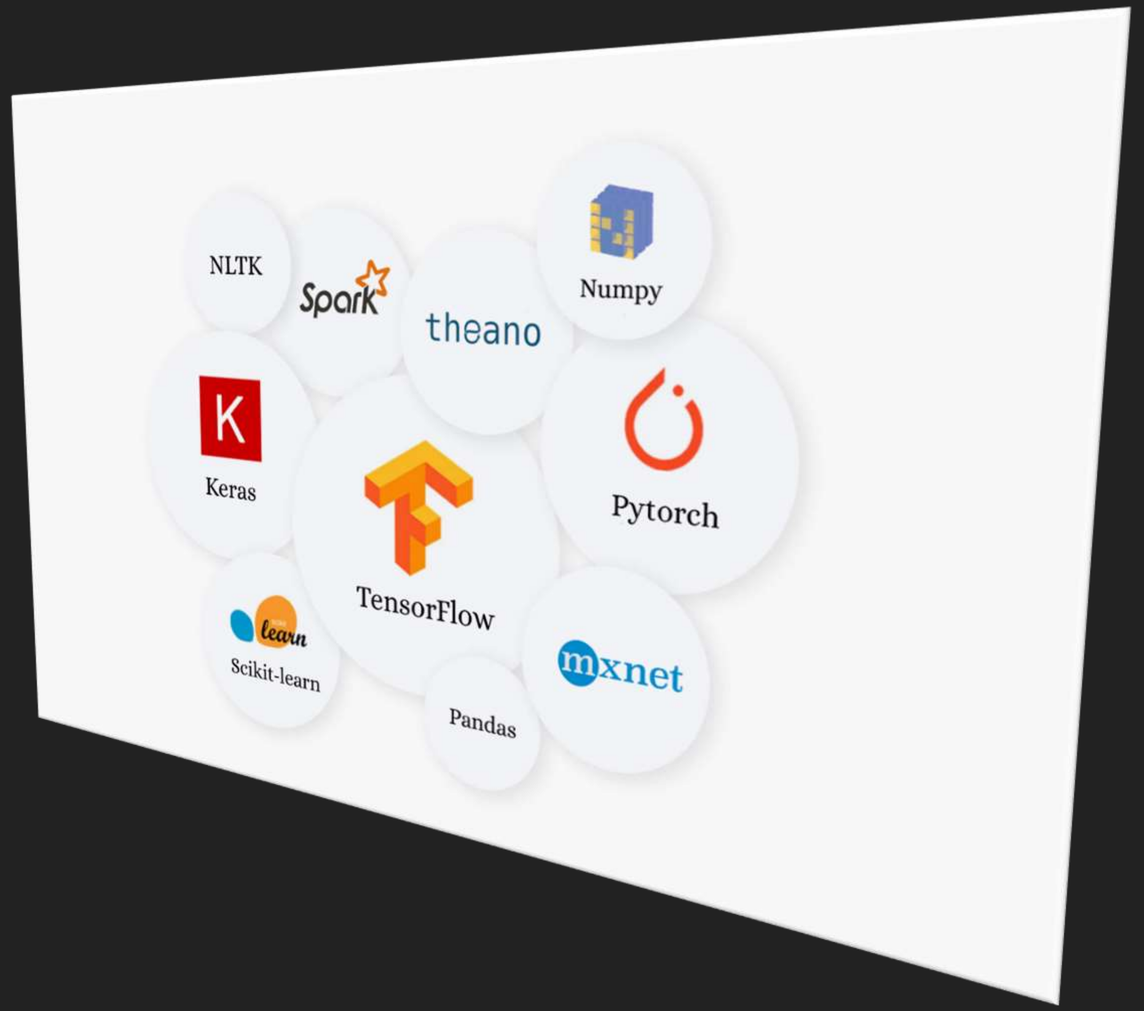
GEEK HERESY

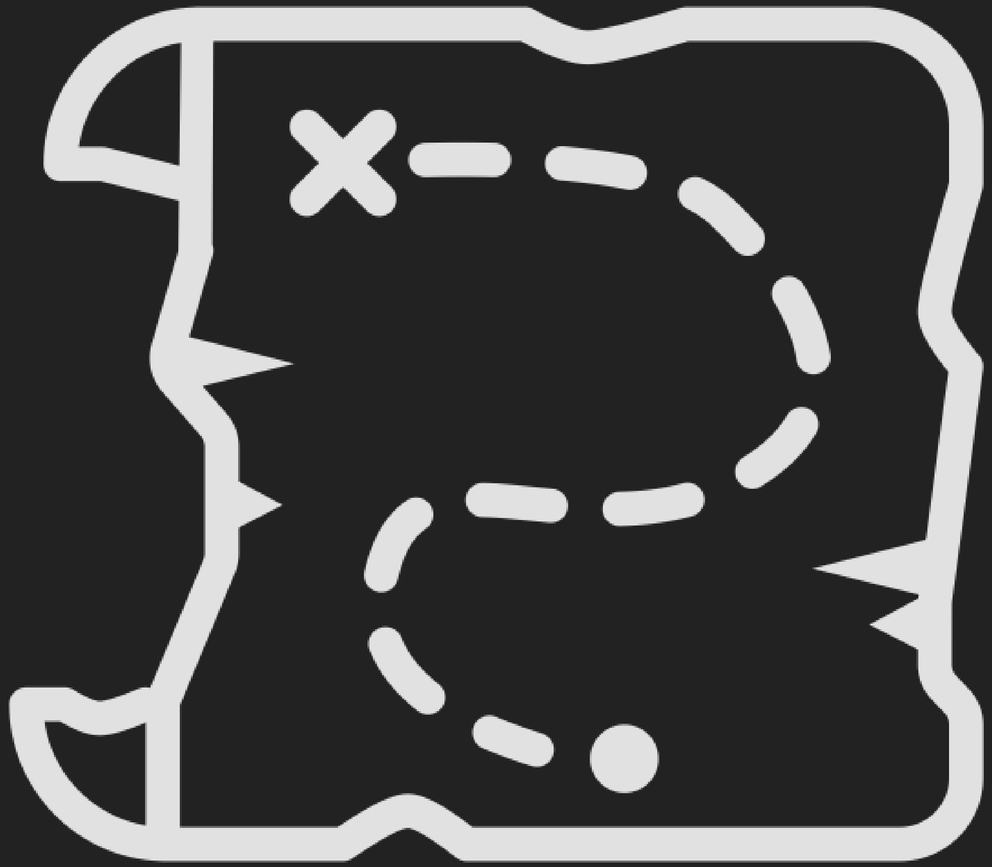
RESCUING SOCIAL CHANGE FROM THE CULT OF TECHNOLOGY

KENTARO TOYAMA


NGO


OUR FIELD HAS BEEN STRUGGLING WITH THIS PROBLEM FOR YEARS.
STRUGGLE NO MORE! I'M HERE TO SOLVE IT WITH ALGORITHMS!
SIX MONTHS LATER:
WOW, THIS PROBLEM IS REALLY HARD.
YOU DON'T SAY.


NLTK
Spark
theano
Numpy
K
Keras
TensorFlow
Pytorch
learn
Scikit-learn
Pandas
mxnet



The New York Times

Zuckerberg Takes Steps to Calm Facebook Employees

Mark Zuckerb
week for his c
Jim Wilson/The

By Sheera

March 23, 2

SAN FRA
grappled
Facebook
another r

Industry Voice Data Strategy Spotlight

the INQUI

Artificial Intelligence Internet of Things Open Source Hardwa

ePlus 6T AMD Zen 2 Samsung Galaxy X Xiaomi c

INDEPENDENT JOIN US SUBSCRIBE / REGISTER LOGIN

News > World

Google employees demand transparency over secret work on censored search engine for China

Staff worry they have been unknowingly working on project 'Dragonfly' to develop technology helping Chinese government withhold information from citizens

Kate Conger and Daisuke Wakabayashi | Friday 17 August 2018 09:55 | 2 comments

protest

.net/.../Guidelines+

rinciples fo

ject

etter to compa
's Joint Enterp
cause harm.

A nu

FORTUNE SUBSCRIBE

PROJECT MAVEN

Report: Google Employees Resigning Over Controversial Pentagon Contract

Googl
Accorc

My account

The Guardian

News Opinion Sport Culture Lifestyle

UK World Business Football UK politics Environment Education Society Science More

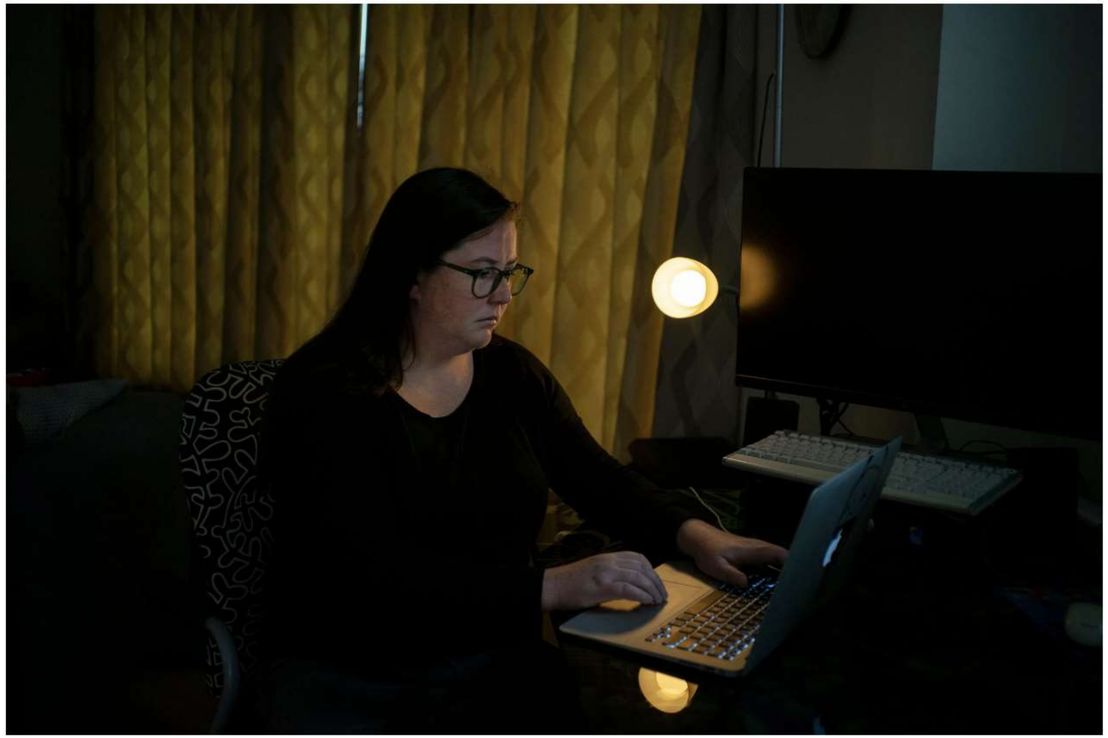
Google

Google walkout: global protests after sexual misconduct allegations

Thousands of employees from Tokyo to California stage demonstrations targeting workplace culture

The New York Times

Tech Workers Now Want to Know: What Are We Building This For?



Laura Nolan, a software engineer in Ireland, left Google in June over the company's involvement in Project Maven, an effort to build artificial intelligence for the Department of Defense. Paulo Nunes dos Santos for The New York Times

By Kate Conger and Cade Metz

Oct. 7, 2018

SAN FRANCISCO — Jack Poulson, a Google research scientist,

The Guardian

n | Sport | Culture | Lifestyle | More

ment Science Global development Football Tech Business Obituaries

Google workers demand reinstatement and apology for fired Black AI ethics researcher

Timnit Gebru's departure sparked outrage in the industry as it followed her paper criticizing the company's diversity efforts

npr

SIGN IN | NPR SHOP | DONATE

NEWS | ARTS & LIFE | MUSIC | SHOWS & PODCASTS | SEARCH

TECHNOLOGY

Ousted Black Google Researcher: 'They Wanted To Have My Presence, But Not Me Exactly'

December 17, 2020 · 8:28 PM ET

BOBBY ALLYN

MIT Technology Review

Sign in | Subscribe

Topics | Magazine | Newsletters | Events

Tech policy / AI Ethics

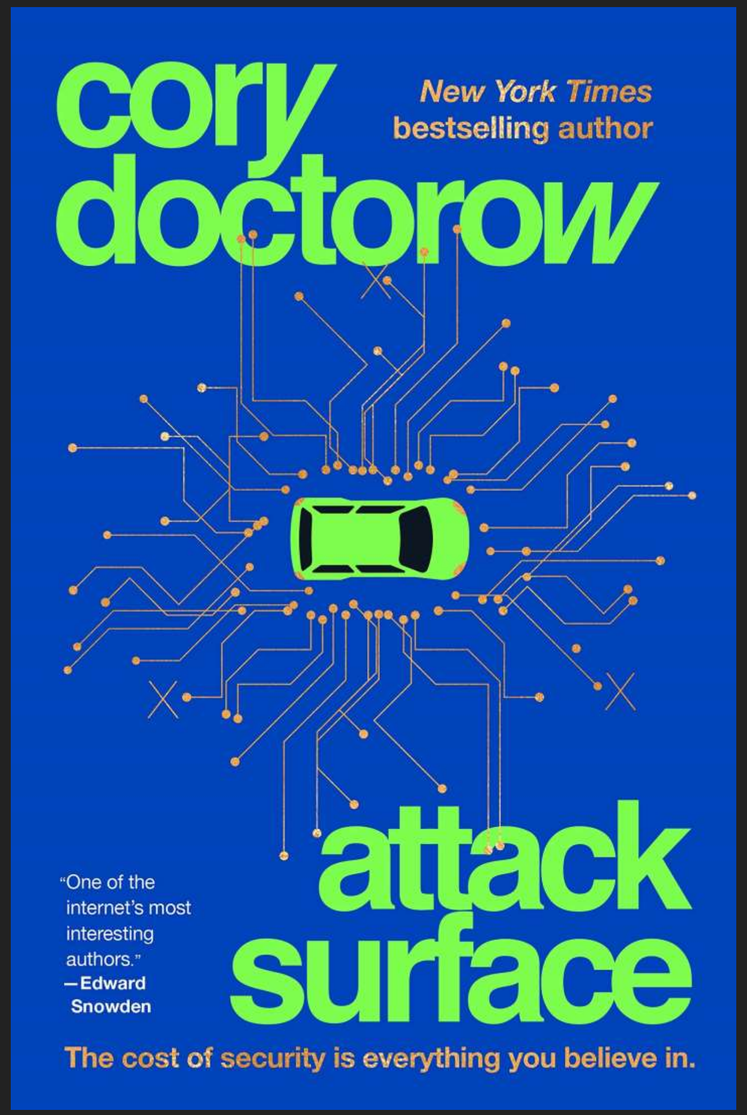
Congress wants answers from Google about Timnit Gebru's firing

The letter, signed by nine members of Congress, sends an important signal about how regulators will scrutinize tech giants.

by **Karen Hao**

December 17, 2020

ANALYSE YOUR ORG


cory doctorow
New York Times bestselling author
attack surface
"One of the internet's most interesting authors."
—Edward Snowden
The cost of security is everything you believe in.