RISK ASSESSMENT / SECURITY & HACKTIVISM

The NSA's SKYNET program may be killing thousands of innocent people

"Ridiculously optimistic" machine learning algorithm is "completely bullshit," says expert.

by Christian Grothoff & J.M. Porup - Feb 16, 2016 3:35am EST

49

An MQ-9 Reaper sits on the tarmac.

In 2014, the former director of both the CIA and NSA proclaimed that "we kill people based on metadata." Now, a new examination of previously published Snowden documents suggests that many of those people may have been innocent.

Last year, The Intercept published documents detailing the NSA's SKYNET programme. According to the documents, SKYNET engages in mass surveillance of Pakistan's mobile phone network, and then uses a machine learning algorithm on the cellular network metadata of 55 million people to try and rate each person's likelihood of being a terrorist.

Patrick Ball—a data scientist and the executive director at the Human Rights Data Analysis Group— who has previously given expert testimony before war crimes tribunals, described the NSA's methods as "ridiculously optimistic" and "completely bullshit." A flaw in how the NSA trains SKYNET's machine learning algorithm to analyse cellular metadata, Ball told Ars, makes the results scientifically unsound.

Somewhere between 2,500 and 4,000 people have been killed by drone strikes in Pakistan since 2004, and most of them were classified by the US government as "extremists," the Bureau of Investigative Journalism reported. Based on the classification date of "20070108" on one of the SKYNET slide decks (which themselves appear to date from 2011 and 2012), the machine learning program may have been in development as early as 2007.

In the years that have followed, thousands of innocent people in Pakistan may have been mislabelled as terrorists by that "scientifically unsound" algorithm, possibly resulting in their untimely demise.

The siren song of big data

SKYNET works like a typical modern Big Data business application. The program collects metadata and stores it on NSA cloud servers, extracts relevant information, and then applies machine learning to identify leads for a targeted campaign. Except instead of trying to sell the targets something, this campaign, given the overall business focus of the US government in Pakistan, likely involves another branch of the US government—the CIA or military—that executes their "Find-Fix-Finish" strategy using Predator drones and on-the-ground death squads.

LATEST FEATURE STORY

FEATURE STORY (7 PAGES)
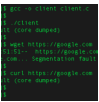
Headshot: A visual history of first-person shooters

Doom, Halo, Goldeneye, Half-Life, Call of Duty... you may recognise a few of these.

STAY IN THE KNOW WITH

LATEST NEWS

TOO BIG TO FAIL

New Monopoly game cracks down on unscrupulous bankers by removing cash

Extremely severe bug leaves dizzying number of apps and devices vulnerable

POISON CLOUDS

Astronomers have detected a Super-Earth atmosphere for the first time

PEW PEW PEW

Space combat is cheaper than ever with *Elite Dangerous: Arena*

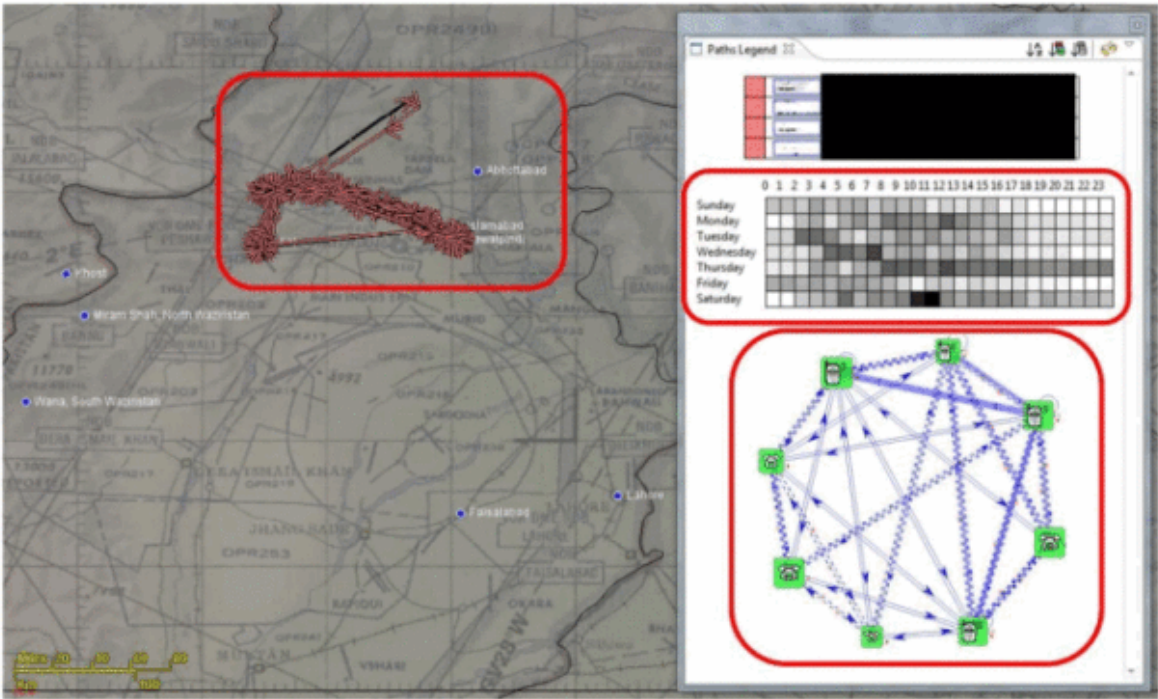**Vulkan now official, with 1.0 API release and Nvidia, AMD drivers [Updated]**

Pressure grows to rethink Snooper's Charter as Labour winds back its support

TOP SECRET//COMINT//REL TO USA, FVEY

From GSM metadata, we can measure aspects of each selector's **pattern-of-life,** **social network,** and **travel behavior**

TOP SECRET//COMINT//REL TO USA, FVEY

Enlarge / From GSM metadata, we can measure aspects of each selector's pattern-of-life, social network, and travel behaviour

In addition to processing logged cellular phone call data (so-called "DNR" or Dialled Number Recognition data, such as time, duration, who called whom, etc.), SKYNET also collects user location, allowing for the creation of detailed travel profiles. Turning off a mobile phone gets flagged as an attempt to evade mass surveillance. Users who swap SIM cards, naively believing this will prevent tracking, also get flagged (the ESN/MEID/IMEI burned into the handset makes the phone trackable across multiple SIM cards).

Enlarge / Travel patterns, behaviour-based analytics, and other "enrichments" are used to analyse the bulk metadata for terroristiness.

Even handset swapping gets detected and flagged, the slides boast. Such detection, we can only speculate (since the slides do not go into detail on this point), is probably based on the fact that other metadata, such as user location in the real world and social network, remain unchanged.

Given the complete set of metadata, SKYNET pieces together people's typical daily routines—who travels together, have shared contacts, stay overnight with friends, visit other countries, or move permanently. Overall, the slides indicate, the NSA machine learning algorithm uses more than 80 different properties to rate people on their terroristiness.

The program, the slides tell us, is based on the assumption that the behaviour of terrorists differs significantly from that of ordinary citizens with respect to some of these properties. However, as The Intercept's exposé last year made clear, the highest rated target according to this machine learning program was Ahmad Zaidan, Al-Jazeera's long-time bureau chief in Islamabad.
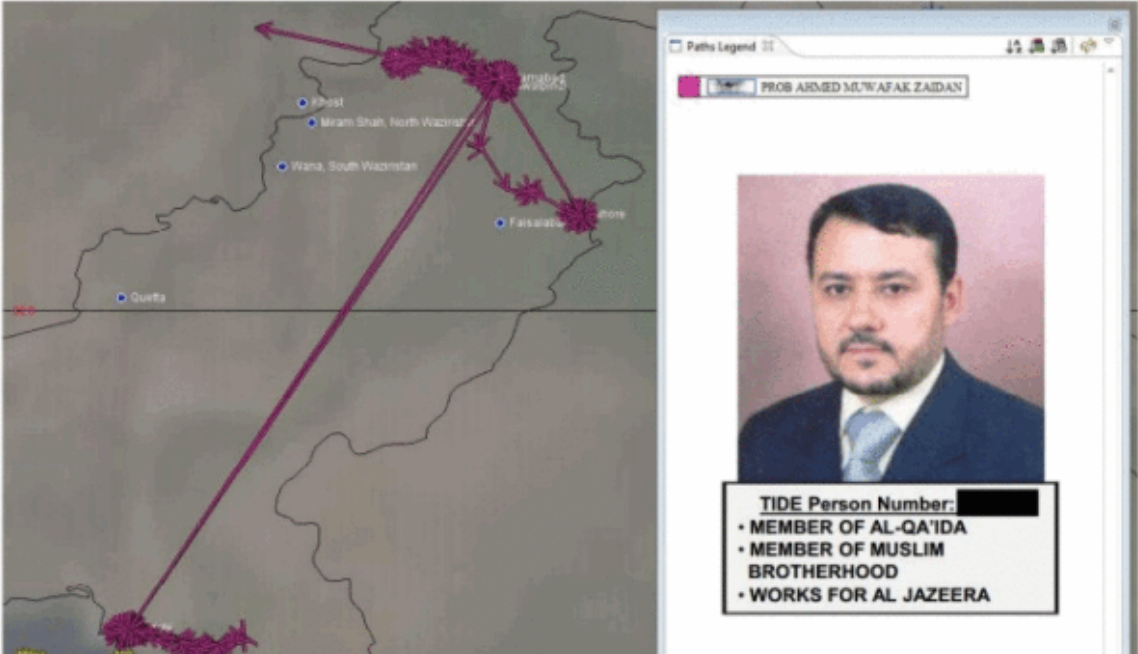
The highest scoring selector that traveled to Peshawar and Lahore is PROB AHMED ZAIDAN

Paths Legend

PROB AHMED MUWAFAK ZAIDAN

TIDE Person Number:
- MEMBER OF AL-QA'IDA
- MEMBER OF MUSLIM BROTHERHOOD
- WORKS FOR AL JAZEERA

As The Intercept reported, Zaidan frequently travels to regions with known terrorist activity in order to interview insurgents and report the news. But rather than questioning the machine learning that produced such a bizarre result, the NSA engineers behind the algorithm instead trumpeted Zaidan as an example of a SKYNET success in their in-house presentation, including a slide that labelled Zaidan as a "MEMBER OF AL-QA'IDA."

Feeding the machine

Training a machine learning algorithm is like training a Bayesian spam filter: you feed it known spam and known non-spam. From these "ground truths" the algorithm learns how to filter spam correctly.

In the same way, a critical part of the SKYNET program is feeding the machine learning algorithm "known terrorists" in order to teach the algorithm to spot similar profiles.

The problem is that there are relatively few "known terrorists" to feed the algorithm, and real terrorists are unlikely to answer a hypothetical NSA survey into the matter. The internal NSA documents suggest that SKYNET uses a set of "known couriers" as ground truths, and assumes by default the rest of the population is innocent.

Pakistan has a population of around 192 million people, with about 120 million cellular handsets in use at the end of 2012, when the SKYNET presentation was made. The NSA analysed 55 million of those mobile phone records. Given 80 variables on 55 million Pakistani mobile phone users, there is obviously far too much data to make sense of manually. So like any Big Data application, the NSA uses machine learning as an aid—or perhaps a substitute, the slides do not say—for human reason and judgement.

SKYNET's classification algorithm analyses the metadata and ground truths, and then produces a score for each individual based on their metadata. The objective is to assign high scores to real terrorists and low scores to the rest of the innocent population.

Sample Travel Report: Haqqani Network

IMSI	seed-contacts	tasked-contact-count	selector_swapping_num	associated_selectors	visits_regularly	other_countries	phrase
▓	▓	3	3	■	lashkargah_city		helmand kandahar AF PK
							farah AF bala_bulk farah masow farah masow nowbahar
▓	▓	14			nowbahar	IR	masow
▓	▓	5	3	■		BA	ghazni AF sharan urgon
▓		1				AE	AF khost_airport kajir_kalay

Enlarge / A sample travel report produced by SKYNET

To do this, the SKYNET algorithm uses the random forest algorithm, commonly used for this kind of Big Data application. Indeed, the UK's GCHQ also appears to use similar machine learning methods, as new Snowden docs published last week indicate. "It seems the technique of choice when it comes to machine learning is Random Decision Forests," George Danezis, associate professor of Security and Privacy Engineering at University College London, wrote in a blog post analysing the released documents.

The random forest method uses random subsets of the training data to create a "forest" of decision "trees," and then combines those by averaging the predictions from the individual trees. SKYNET's algorithm takes the 80 properties of each cellphone user and assigns them a numerical score—just like a spam filter.

SKYNET then selects a threshold value above which a cellphone user is classified as a "terrorist." The slides present the evaluation results when the threshold is set to a 50 percent false negative rate. At this rate, half of the people who would be classified as "terrorists" are instead classified as innocent, in order to keep the number of false positives—innocents falsely classified as "terrorists"—as low as possible.

False positives

We can't be sure, of course, that the 50 percent false negative rate chosen for this presentation is the same threshold used to generate the final kill list. Regardless, the problem of what to do with innocent false positives remains.
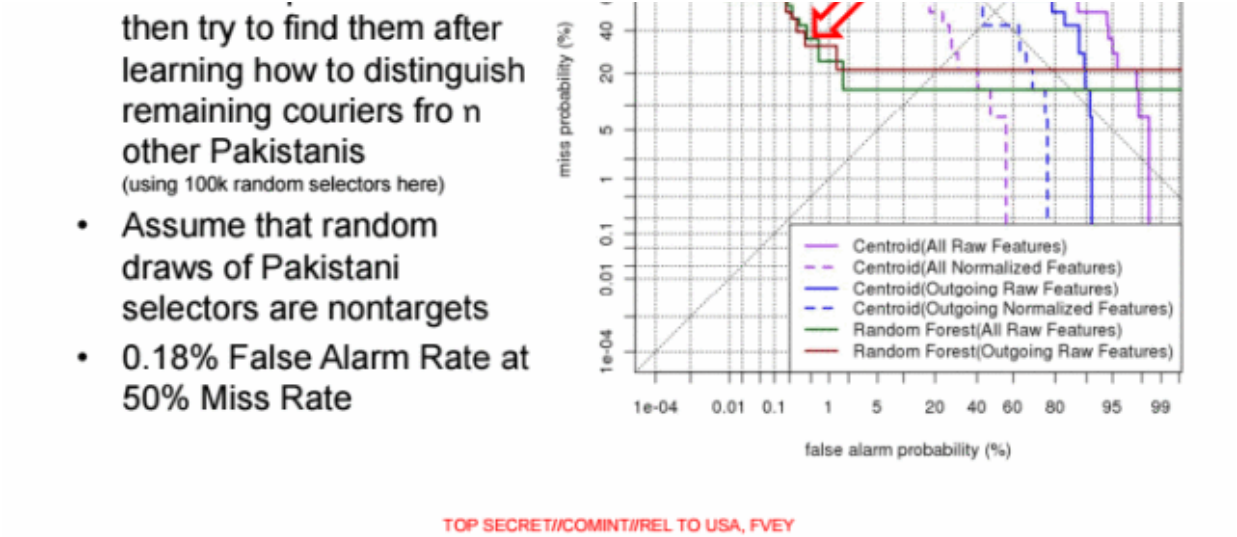
"The reason they're doing this," Ball explained, "is because the fewer false negatives they have, the *more* false positives they're certain to have. It's not symmetric: there are *so many* true negatives that lowering the threshold in order to reduce the false negatives by 1 will mean accepting many thousands of additional false positives. Hence this decision."

Statistical algorithms are able to find the couriers at very low false alarm rates. if we're allowed to miss half of them

then try to find them after learning how to distinguish remaining couriers from other Pakistanis
(using 100k random selectors here)

- Assume that random draws of Pakistani selectors are nontargets
- 0.18% False Alarm Rate at 50% Miss Rate

Centroid(All Raw Features)
Centroid(All Normalized Features)
Centroid(Outgoing Raw Features)
Centroid(Outgoing Normalized Features)
Random Forest(All Raw Features)
Random Forest(Outgoing Raw Features)

miss probability (%)

false alarm probability (%)

One NSA slide brags, "Statistical algorithms are able to find the couriers at very low false alarm rates, if we're allowed to miss half of them."

But just how low is the NSA's idea of "very low"?

"Completely bullshit"

The problem, Ball told Ars, is how the NSA trains the algorithm with ground truths.

The NSA evaluates the SKYNET program using a subset of 100,000 randomly selected people (identified by their MSIDN/MSI pairs of their mobile phones), and a a known group of seven terrorists. The NSA then trained the learning algorithm by feeding it six of the terrorists and tasking SKYNET to find the seventh. This data provides the percentages for false positives in the slide above.

"First, there are *very few* 'known terrorists' to use to train *and test* the model," Ball said. "If they are using the same records to train the model as they are using to test the model, their assessment of the fit is completely bullshit. The usual practice is to hold some of the data out of the training process so that the test includes records the model has never seen before. Without this step, their classification fit assessment is ridiculously optimistic."

The reason is that the 100,000 citizens were selected at random, while the seven terrorists are from a known cluster. Under the random selection of a tiny subset of less than 0.1 percent of the total population, the density of the social graph of the citizens is massively reduced, while the "terrorist" cluster remains strongly interconnected. Scientifically-sound statistical analysis would have required the NSA to mix the terrorists into the population set *before* random selection of a subset—but this is not practical due to their tiny number.

This may sound like a mere academic problem, but, Ball said, is in fact highly damaging to the quality of the results, and thus ultimately to the accuracy of the classification and assassination of people as "terrorists." A quality evaluation is especially important in this case, as the random forest method is known to overfit its training sets, producing results that are overly optimistic. The NSA's analysis thus does not provide a good indicator of the quality of the method.

TOP SECRET//COMINT//REL TO USA, FVEY

We've been experimenting with several error metrics on both small and large test sets

Training Data	Classifier	Features	100k Test Selectors		55M Test Selectors	
			False Alarm Rate at 50% Miss Rate	Mean Reciprocal Rank	Tasked Selectors in Top 500	Tasked Selectors in Top 100
None	Random	None	50%	1/23k (simulated)	0.64 (active/Pak)	0.13 (active/Pak)
Known Couriers	Centroid	All	20%	1/18k		
		Outgoing	43%	1/27k		
+ Anchory Selectors	Random Forest		0.18%	1/9.9	5	1

Random Forest:
- 0.18% false alarm rate at 50% miss rate
- 7x improvement over random performance when evaluating its tasked precision at 100

TOP SECRET//COMINT//REL TO USA, FVEY

If 50 percent of the false negatives (actual "terrorists") are allowed to survive, the NSA's false positive rate of 0.18 percent would still mean thousands of innocents misclassified as "terrorists" and potentially killed. Even the NSA's most optimistic result, the 0.008 percent false positive rate, would still result in many innocent people dying.

"On the slide with the false positive rates, note the final line that says '+ Anchory Selectors,'" Danezis told Ars. "This is key, and the figures are unreported... if you apply a classifier with a false-positive rate of 0.18 percent to a population of 55 million you are indeed likely to kill thousands of innocent people. *[0.18 percent of 55 million = 99,000]*. If however you apply it to a population where you already expect a very high prevalence of 'terrorism'—because for example they are in the two-hop neighbourhood of a number of people of interest—then the prior goes up and you will kill fewer innocent people."

Besides the obvious objection of how many innocent people it is ever acceptable to kill, this also assumes there are a lot of terrorists to identify. "We know that the 'true terrorist' proportion of the full population is very small," Ball pointed out. "As Cory [Doctorow] says, if this were not true, we would all be dead already. Therefore a small false positive rate will lead to misidentification of lots of people as terrorists."

"The larger point," Ball added, "is that the model will totally overlook 'true terrorists' who are statistically different from the 'true terrorists' used to train the model."

In most cases, a failure rate of 0.008% would be great

The 0.008 percent false positive rate would be remarkably low for traditional business applications. This kind of rate is acceptable where the consequences are displaying an ad to the wrong person, or charging someone a premium price by accident. However, even 0.008 percent of the Pakistani population still corresponds to 15,000 people potentially being misclassified as "terrorists" and targeted by the military—not to mention innocent bystanders or first responders who happen to get in the way.

Security guru Bruce Schneier agreed. "Government uses of big data are inherently different from corporate uses," he told Ars. "The accuracy requirements mean that the same technology doesn't work. If Google makes a mistake, people see an ad for a car they don't want to buy. If the government makes a mistake, they kill innocents."

Killing civilians is forbidden by the Geneva Convention, to which the United States is a signatory. Many facts about the SKYNET program remain unknown, however. For instance, is SKYNET a closed loop system, or do analysts review each mobile phone user's profile before condemning them to death based on metadata? Are efforts made to capture these suspected "terrorists" and put them on trial? How can the US government be sure it is not killing innocent people, given the apparent flaws in the machine learning algorithm on which that kill list is based?

"On whether the use of SKYNET is a war crime, I defer to lawyers," Ball said. "It's bad science, that's for damn sure, because classification is inherently probabilistic. If you're going to condemn someone to death, usually we have a 'beyond a reasonable doubt' standard, which is not at all the case when you're talking about people with 'probable terrorist' scores anywhere near the threshold. And that's assuming that the classifier works in the first place, which I doubt because there simply aren't enough positive cases of known terrorists for the random forest to get a good model of them."

The leaked NSA slide decks offer strong evidence that thousands of innocent people are being labelled as terrorists; what happens after that, we don't know. We don't have the full picture, nor is the NSA likely to fill in the gaps for us. (We repeatedly sought comment from the NSA for this story, but at the time of publishing it had not responded.)

Algorithms increasingly rule our lives. It's a small step from applying SKYNET logic to look for "terrorists" in Pakistan to applying the same logic domestically to look for "drug dealers" or "protesters" or just people who disagree with the state. Killing people "based on metadata," as Hayden said, is easy to ignore when it happens far away in a foreign land. But what happens when SKYNET gets turned on us—assuming it hasn't been already?

* * *

Christian Grothoff leads the Décentralisé research team at Inria, a French institute for applied computer science and mathematics research. He earned his PhD in computer science from UCLA, an MS in computer science from Purdue University, and a diploma in mathematics from the University of Wuppertal. He is also a freelance journalist reporting on technology and national security.

J.M. Porup is a freelance cybersecurity reporter who lives in Toronto. When he dies his epitaph will simply read "assume breach." You can find him on Twitter at @toholdaquill.

READER COMMENTS 49
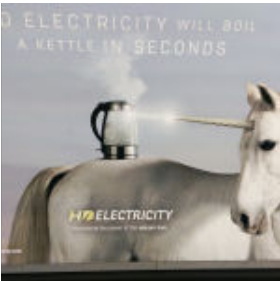
5617   149   6854

YOU MAY ALSO LIKE ◢

It's official: NSA did keep its e-mail metadata program

O2 customers will have their Underground

Headshot: A visual history of first-person shooters

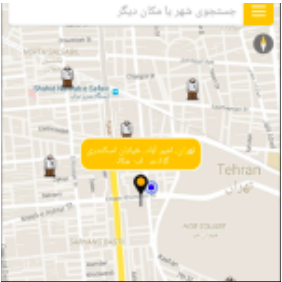**Particle's Electron is a "cellular Arduino" with a global data**

after it "ended" in
2011

journeys tracked,

plan

UK-US deal would allow MI5 to get chat and e-mails directly from US companies

Android app helps Iranians avoid morality police checkpoints

What the US government should've learned about backdoors from the Clipper

Snooper's Charter given thumbs up by UK parliament report, calls for minor tweaks

SITE LINKS

About Us

Advertise with us

Contact Us

MORE READING

RSS Feeds

Newsletters

Visit Ars Technica US

CONDE NAST SITES

Reddit

Wired

GQ

Vanity Fair

Condé Nast Traveller

VIEW MOBILE SITE