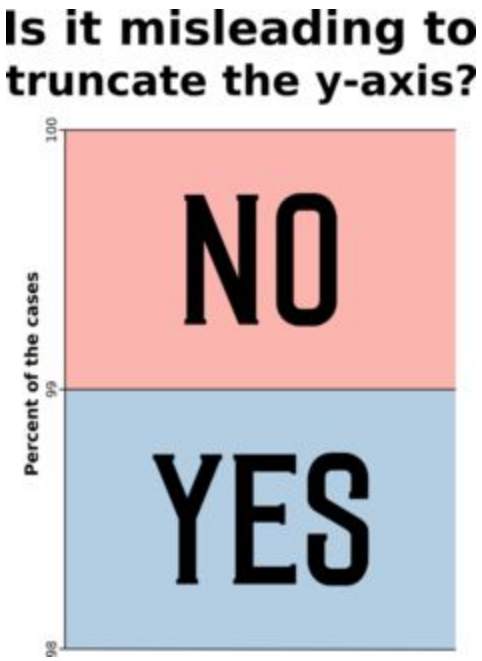
Security Now! #702 - 02-19-19

Authenticity on the Internet

This week on Security Now!

This week we catch up with last week's doozy of a patch Tuesday for both Microsoft and Adobe. We also examine an interesting twist coming to Windows 7 and Server 2008 security updates, eight mining apps pulled from the Windows Store, another positive security initiative from Google, electric scooters being hacked, more shipping away at Tor's privacy guarantees, a year and a half after Equifax and where's the data?, the beginnings of GDPR-like legislation for the US, some closing the loop feedback from our terrific listeners... then we take a look at an extremely concerning new and emerging threat for the Internet.

One of my most enduring pet peeves...



Security News

Last week's Patch Tuesday was a doozy.

These path Tuesdays seem to be getting larger. Remember those quaint old days, a few years ago, when we'd have a BIG patch day of 15 patches? Yeah. Not recently. Last Tuesday, Microsoft's February patches resolved 74 CVEs and three advisories. These encompassed Internet Explorer (IE), Edge, Exchange Server, ChakraCore, Windows itself, Office, Microsoft Office Services and Web Apps, Azure, Team Foundation Services and the .NET Framework. And, of those 74 CVEs, 20 are rated Critical, 54 are rated Important, and three are rated Moderate in severity.

21 of these CVEs were sourced by Trend Micro's Zero-Day Initiative (ZDI). Four of the bugs are public and one is under active attack at the time of release. One of the more interesting things patched was CVE-2019-0626, a Windows DHCP Server Remote Code Execution Vulnerability. Most of us have DHCP servers on our networks. Though most are running an embedded Linux on a fanless consumer router. But many Windows-oriented enterprises will be providing DHCP from a Windows-based server and in that case be glad that you applied last week's patches right away. And if you didn't, consider this: The bug fixed last week allowed attackers to take over your DHCP server just by sending it a specially crafted packet. Whoops! And code execution through a publicly-exposed network service that executes with system privileges places this into the "wormable" category.

The one flaw that was found in the wild isn't very serious by itself. CVE-2019-0676 is an Internet Explorer Information Disclosure Vulnerability. If the user visited a malicious site with IE (who does that much anymore?) an attacker could leverage thee flaw to check for files on a target system. Microsoft hasn't indicated exactly how this bug is being exploited in the wild, but it’s probably restricted to targeted attacks. And considering that Microsoft now lists IE as “a compatibility solution” rather than a browser (ha!), now is a good time to choose another browser.

There are also a pair of critical-rated remote code execution bugs in GDI+ and 15 other "browser and own" vulnerabilities affecting IE, Edge, and ChakraCore. So... our browsers remain the biggest attack surface for our systems.

And Abode?

Weighing in with it's own list of 71 bugs, Adobe didn't quite match Microsoft in total bug count... but considering that 44 of those 71 were rated critical, Adobe did win this month's booby prize. The problems found and fixed included Acrobat, Reader, Flash, ColdFusion, and Creative Cloud.

Among the critical vulnerabilities patched include a zero-day flaw disclosed in January in Acrobat Reader which could lead to the theft of hashed password values... and which one of those little micro patches was published by "0patch" this week.

Other critical bugs resolved in the update include buffer errors, sensitive data leakage, an integer overflow vulnerability which could lead to information disclosure, a double-free bug, security bypass problems, and use-after-free issues leading to arbitrary code execution.

So... if you have any Adobe stuff, check for updates! A whole mess of things were fixed.

Last Friday, February 15th, Microsoft notified of an important forthcoming change:
https://support.microsoft.com/en-us/help/4472027/2019-sha-2-code-signing-support-requirement-for-windows-and-wsus

Under the title: "2019 SHA-2 Code Signing Support requirement for Windows and WSUS" (Windows Server Update Services).

Summary
To protect your security, Windows operating system updates are dual-signed using both the SHA-1 and SHA-2 hash algorithms to authenticate that updates come directly from Microsoft and were not tampered with during delivery. Due to weaknesses in the SHA-1 algorithm and to align to industry standards Microsoft will only sign Windows updates using the more secure SHA-2 algorithm exclusively.

Customers running legacy OS versions (Windows 7 SP1, Windows Server 2008 R2 SP1 and Windows Server 2008 SP2) will be required to have SHA-2 code signing support installed on their devices by July 2019. Any devices without SHA-2 support will not be offered Windows updates after July 2019. To help prepare you for this change, we will release support for SHA-2 signing in 2019. Some older versions of Windows Server Update Services (WSUS) will also receive SHA-2 support to properly deliver SHA-2 signed updates. Refer to the Product Updates section for the migration timeline.

Starting in early 2019, the migration process to SHA-2 support will occur in stages, and support will be delivered in standalone updates.

March 12, 2019 (next Patch Tuesday) for Windows 7 SP1, Windows Server 2008 R2 SP1.
Stand Alone updates that introduce SHA-2 code sign support will be released as security updates.

April 9, 2019 (April's Patch Tuesday) for Windows Server 2008 SP2.
Stand Alone updates that introduce SHA-2 code sign support will be released as security updates.

Then we have May, June, & July to get caught up and ready... Since the AUGUST 2019 updates will be single-signed using SHA-2 only. If systems are not updated by then they will not be able to update.

What's not EXPLICITLY clear is whether user will be required to select this update for installation. Microsoft's documentation states that they will be released as security updates, but in their table for Windows 10 editions they also state:

July 16, 2019 for Windows 10 1507, 1607, 1703
"Windows 10 updates signatures changed from dual signed (SHA1/SHA2) to SHA2 only. No customer action is expected for this milestone.

So... is customer action needed for Win7 and Server 2008 R2?

Eight Monero-mining Win10 apps removed from the Windows Store
https://www.symantec.com/blogs/threat-intelligence/cryptojacking-apps-microsoft-store

Symantec found eight apps on Microsoft's app store that mine Monero without the user's knowledge.

On January 17, we discovered several potentially unwanted applications (PUAs) on the Microsoft Store that surreptitiously use the victim's CPU power to mine cryptocurrency. We reported these apps to Microsoft and they subsequently removed them from their store.

The apps—which included those for computer and battery optimization tutorial, internet search, web browsers, and video viewing and download—came from three developers: DigiDream, 1clean, and Findoo. In total, we discovered eight apps from these developers that shared the same risky behavior. After further investigation, we believe that all these apps were likely developed by the same person or group.

Users may get introduced to these apps through the top free apps lists on the Microsoft Store or through keyword search. The samples we found run on Windows 10, including Windows 10 S Mode.

As soon as the apps are downloaded and launched, they fetch a coin-mining JavaScript library by triggering Google Tag Manager (GTM) in their domain servers. [Google Tag Manager is a tag management system created by Google to manage JavaScript and HTML tags used for tracking and analytics on websites.] The mining script then gets activated and begins using the majority of the computer's CPU cycles to mine Monero for the operators. Although these apps appear to provide privacy policies, there is no mention of coin mining on their descriptions on the app store.

The apps were published between April and December 2018, with most of them published toward the end of the year. Even though the apps were on the app store for a relatively short period of time, a significant number of users may have downloaded them. Although we can't get exact download or installation counts, we can see that there were almost 1,900 ratings posted for these apps. However, app ratings can be fraudulently inflated, so it is difficult to know how many users really downloaded these apps.

When each app is launched, the domain "Fast-search.tk"—the domain for the Fast-search Lite app which is hardcoded into each apps' manifest file—is silently visited in the background and triggers the GTM (Google Tag Manager) with the key GTM-PRFLJPX, which is shared across all eight applications. GTM is a legitimate tool that allows developers to inject JavaScript dynamically into their applications. However, GTM can be abused to conceal malicious or risky behaviors, since the link to the JavaScript stored in GTM is https://www.googletagmanager.com/gtm.js?id={GTM ID} which doesn't indicate the function of the code invoked.

(So... Yet another clever way for bad guys to get bad code loaded into their apps without needing to carry it along. And note, also, that WHAT is injected is also subject to later change. So this is all a complete mess and it makes a joke out of the term "Security Model.")

The script which is downloaded is encrypted JavaScript which Symantec, of course, decrypted, to find a bit of script which invokes the Coinhive library and cryptominer.

The eight apps fall under the category of Progressive Web Applications, which are installed as a Windows 10 app running independently from the browser, in a standalone (WWAHost.exe process) window.

Symantec: We have informed Microsoft and Google about these apps' behaviors. Microsoft has removed the apps from their store. The mining JavaScript has also been removed from Google Tag Manager.

Google continues moving the browser security bar forward
https://developers.google.com/web/updates/2019/02/trusted-types

The next major release of Chrome, and others through 2019, will be offering an experimental new lock-down technology which Google has dubbed "Trusted Types".

Our browser eco-system has become incredibly complex over time. The text-based loosely-typed and interpreted authoring environment results in code which easily does what developers want... but will also obligingly do what developers never intended. In this environment, where pre-code modules are being sucked down with abandon, and developers are rapidly gluing together complex functions built up from code they've never seen, it is incredibly difficult to program defensively.

The result is extreme vulnerability to cross-site scripting (XSS) vulnerabilities which are very difficult to find... even in the too-rare event that someone is looking. Unfortunately, it's often the bad guys who are doing the looking, and the finding.

Without getting too far down into the weeds, the problem arises because elements which make up web pages, are part of the so-called DOM -- the Document Object Model which describes the page's structure, which has been carefully designed -- and the specifications of the sources for these DOM objects are strings. And it turns out that all too often the composition of these strings is subject to malicious manipulation... with the result being that foreign content from some site with typically malicious intent (thus cross-site) can be injected to the innocent page's DOM and made to execute.

Again... The problem is that simple strings can be used as the source specifiers of these objects. As Google terms it... they are insecure by default. Yes, they are easy to use. But also, yes, their abuse is the #1 source of web-based attacks today.

So, what Google's developers are proposing is the addition of a new and optional argument to the existing list of "Content Security Policy" headers. They will be adding "trusted-types *"

If a site (a webserver) emits a page with headers that include

Content-Security-Policy: trusted-types *

...then all of those any places where a string could have been used -- and easily misused -- will refuse to accept a standard string as their argument. Instead, an explicitly specified template and policy will be created to much more tightly control the assignment of values to this page element.

Google says:
In practice, modern web applications need only a small number of policies. The rule of thumb is to create a policy where the client-side code produces HTML or URLs - in script loaders, HTML templating libraries or HTML sanitizers. All the numerous dependencies that do not interact with the DOM, do not need the policies. Trusted Types assures that they can't be the cause of the XSS.

Surprise!!... Xiaomi electric scooters are vulnerable to remote hijacking
Last Tuesday, researcher Rani Idan who is with Zimperium (our somewhat controversial 0-day exploit promoter and reseller) disclosed a vulnerability present in the Xiaomi M365 electric scooter which could potentially permit attackers to remotely control the scooter's velocity including sudden acceleration or braking.

The Xiaomi scooter interface app is running in the phone. It receives and processes the user's password. But, believe it or not, there is no authentication at all between the phone app and the scrooter... thus the scooter will execute any commands it receives without any authentication mechanism. This allows "malicious scooter apps" a user might download to take over and control the scooter without the user's control, overriding the user's control inputs.

You probably don't want to be on a runaway electric scooter.

To demonstrate the vulnerability, Zimperium created proof-of-concept (PoC) code developed as a malicious app that was able to scan for nearby Xiaomi M365 scooters and send crafted payloads to exploit the flaw. Idan says that that vehicles up to 100 meters away can be exploited.

Zimperium says that Xiaomi was made aware of the findings and recently said this was a "known issue internally" caused by "third-party products." However, Zimperium says that the scooters are yet to be patched.

Chipping away at TOR
https://arxiv.org/pdf/1901.04434.pdf
As we've said and covered in the past, the ultimate vulnerability to Internet anonymity is the FACT of their being any traffic flow between endpoints. We can encrypt it so that we cannot read it. We can encrypt the flow's metadata so that we cannot learn anything (other than perhaps size) of the data being moved, and we can introduce camouflage packet padding and short-term aggregation of packets to hide their individual presence. But eventually, the same packet needs to come out of the "cloud" that went in, so any entity that can see enough of the cloud's perimeter can get some sense for which endpoints are communicating. They may not know what they are saying, but the existence of a flow is a form of metadata that reveals something.

Now, in their paper titled: "Peel the onion: Recognition of Android apps behind the TorNetwork" four Italian researchers at the Sapienza University of Rome have chipped away a bit more at the protections offered by TOR. We now have "application deanonymization attacks"...

ABSTRACT
In this work we show that Tor is vulnerable to app deanonymization attacks on Android devices through network traffic analysis. For this purpose, we describe a general methodology for performing an attack that allows us to deanonymize the apps running on a target smartphone using Tor, which is the victim of the attack. Then, we discuss a Proof-of-Concept, implementing the methodology, that shows how the attack can be performed in practice and allows to assess the deanonymization accuracy that it is possible to achieve. While attacks against Tor anonymity have already gained considerable attention in the context of website fingerprinting in desktop environments, to the best of our knowledge this is the firstwork that highlights Tor vulnerability to apps deanonymization attacks on Android devices. In our experiments we achieved an accuracy of 97%.

Their 15-page paper goes into every detail for what they did, what they observed and what they found. The bottom line is that, not surprisingly, mobile applications tend to be quite chatty. They assume and use their connectivity freely... and they are each unique in the "fingerprint" of their chattiness. So even under the multi-layer onion-routing wrapper of a Tor-enabled smartphone, a passive WiFi eavesdropper who only sees packets zipping back and forth, in and out of a smartphone, is able to use an existing app-aware training set to determine WHAT applications the user has and is using on their phone.

CONCLUSION
In this work we have shown that Tor is vulnerable to app deanonymization. We described a general methodology to perform an attack against a target smartphone which allows us to unveil which app the victim is using. The proposed methodology performs network traffic analysis based on a supervised machine learning approach. It leverages the fact that different apps produce different recognizable traffic patterns, even when protected by Tor. We also provided a Proof-of-Concept that implements the methodology, that we employed to assess the accuracy that it can achieve in deanonymizing apps. We performed several experiments achieving an accuracy of 97.3%. We made the software of the Proof-of-Concept, as well as the datasets that we built during the experiments, publicly available, so that it can be used to assess Tor’s vulnerability to this attack, compare alternative methodologies and test possible countermeasures.

This may not seem like a big deal, and all by itself, it isn't. But it is an interesting finding and succeeds in breaking one of the privacy assumptions a TOR user might have. And if they were depending upon that assumption for some reason, this could be important. If nothing else it's a very nice and interesting piece of research that could form the basis for something more.

Where did the stolen Equifax data go? There's been no sign of it since...
https://threatpost.com/equifax-data-nation-state/141929/

The breached data from the massive Equifax incident is nowhere to be found... which suggests that it might have been a higher-end nation-state job.

It's been just shy of a year and a half - 17 months - since the 2017 Equifax data breach was revealed to have compromised the data of about 147.9 million people. This was nearly every adult in the U.S., with more than 45 percent of the US population directly affected by the incident.

But an investigative report by CNBC found that, somewhat surprisingly, none of the data has turned up on the Dark Web. According to CNBC's threat-hunter sources, it's increasingly looking like it was a spy job, carried out by a nation-state; not criminals bent of ID theft or financial gain.

ThreatPost recently asked some security experts what they thought and Troy Hunt, of "HaveIBeenPwned" fame, was quoted:

"Frankly, I think the bullet point under the headline about it being state-sponsored explains a lot. Actors at that level aren't looking to cash data in for a few bitcoin and it wouldn't surprise me in the least if that data never sees the light of day. Just think about how many incidents must be out there already that we may never know about simply because those responsible have no reason to advertise it."

A GDPR for the USofA?
https://www.gao.gov/assets/700/696437.pdf

INTERNET PRIVACY: Additional Federal Authority Could Enhance Consumer Protection and Provide Flexibility.

Two years ago, the US Congress' House Energy and Commerce Committee requested that the preparation of the report by the United States Government Accountability Office (the GAO). The first hearing on the report is scheduled for February 26th, a week from today. During that hearing members of the Committee plan to discuss the GAO's findings and the possibility in drafting the first US federal-level internet privacy law.

If the committee were to follow the GAO's recommendations the US might very well see GDPR-like legislation coming to the US. And in today's wild West "profits over privacy" climate, I'll be surprised if we don't see something. And in their report, the GAO suggest that the Federal Trade Commission (FTC) be put in charge of overseeing enforcement of Internet privacy under the proposed forthcoming legislation. The FTC already has this responsibility, but it has lacked legislation with sufficient teeth allowing to allow it to act. For example, in its entire history, the FTC has been involved in only 101 Internet privacy-related cases despite wide privacy abuse being reported by users and the press. The GAO argues in their report that this new proposed legislation should give the FTC more teeth in dealing with privacy abuse.

To shore up its conclusion that a tough new Internet privacy law is needed, GAO investigators cited the now-famous Facebook/Cambridge Analytica relationship and its own previous reports including:

The dangers to user privacy due to the lack of regulation and oversight in the ever-growing Internet of Things (IoT) sector where devices collect massive amounts of information without users' knowledge.

- Automakers collecting data from smart cars owners.
- The lack of federal oversight over companies that collect and resell user information.
- The lack of protections for mobile users against secret data collection practices.

To build its case, the GAO looked at the FTC's previous 101 user internet privacy investigations and also considered feedback from the private sector, academia, advocacy groups, other government agencies, and nine former FTC and FCC top-ranking officials, including seven former commissioners.

The House Energy and Commerce Chairman Frank Pallone, Jr. who requested the report on behalf of Congress in 2017 said: "This detailed GAO report makes clear now is the time for comprehensive congressional action on privacy that should include ensuring any agency that oversees consumer privacy has the tools to protect consumers. These recommendations and findings will be helpful as we look to develop privacy legislation in the coming months."

It feels as though the time is right for this. The FTC announced last week that it is considering levying a multi-billion dollar fine against Facebook for a series of privacy violations which include the Cambridge Analytica mess. Last year Apple's Tim Cook urged the US to copy the EU's GDPR, and Oregon's Senator Ron Wyden has introduced a bill that would jail company execs for lying or not reporting privacy violations.

I read though the 56-page report and while there was everything you would expect, nothing stood out as surprising. So we're going to need to wait to see how this develops. But as I mentioned, it really does feel as though something's going to happen.

The GAO report concluded, saying:

Recent developments regarding Internet privacy suggest that this is an appropriate time for Congress to consider comprehensive Internet privacy legislation. Although FTC has been addressing Internet privacy through its unfair and deceptive practices authority, among other statutes, and other agencies have been addressing this issue using industry-specific statutes, there is no comprehensive federal privacy statute with specific standards. Debate over such a statute could provide a vehicle for consideration of the Fair Information Practice Principles, which are intended to balance privacy concerns with the need for using consumers’ data. Such a law could also empower a specific agency or agencies to provide oversight through means such as APA (Administrative Procedure Act) section 553 [so-called "informal"] rulemaking, civil penalties for first time violations of a statute, and other enforcement tools. Comprehensive legislation addressing Internet privacy that establishes specific standards and includes APA notice-and-comment rulemaking and first-time violation civil penalty authorities could help

enhance the federal government's ability to protect consumer privacy, provide more certainty in the marketplace as companies innovate and develop new products using consumer data, and provide better assurance to consumers that their privacy will be protected.

Congress should consider developing comprehensive legislation on Internet privacy that would enhance consumer protections and provide flexibility to address a rapidly evolving Internet environment. Issues that should be considered include:

- which agency or agencies should oversee Internet privacy;
- what authorities an agency or agencies should have to oversee Internet privacy, including notice-and-comment rulemaking authority and first-time violation civil penalty authority; and
- how to balance consumers' need for Internet privacy with industry's ability to provide services and innovate

Closing The Loop

Todd Fillingim in Mississippi
Subject: Windows 10 Long Term Service Channel (Branch)
:
Steve,

I've been listening to SN since episode 1 but have never submitted feedback. Love the show and thank you for doing it for so long.

In SN701 you mentioned a version of Win10 that has everything stripped out of it available to enterprise customers. I work for a small utility company and we have just recently started testing this version for use on our control room operator workstations.

Microsoft used to call this version the Long Term Service Branch (LTSB) but recently changed it to Long Term Service Channel (LTSC). It is a version of Win10 that has none of the extra features of the regular version AND every build has a commitment from Microsoft for 10 years of security patches and support.

Interestingly, everything you read on this version from Microsoft tries to talk you out of using this version. They REALLY don't want anyone using this for desktop use.

https://techcommunity.microsoft.com/t5/Windows-IT-Pro-Blog/LTSC-What-is-it-and-when-should-it-be-used/ba-p/293181

Thanks again for the podcast.

Todd.

Owen LeGare in Davis, CA
Subject: How GB/MB/KB is different from GiB/MiB/KiB
:
Hi Steve,

In the last few SN episodes you and Leo were not sure about a size designation when it included the small letter i in it.

I was just looking at the output in dmesg for a flash drive and it listed the size as 8.13 GB /7.57 GiB Apparently the small i indicates a 1024 byte base while GB uses a 1000 byte base. It is nice to know there is a way to indicate this difference when talking about storage sizes.

Dr-Mosfet <anon@grc.com>
Subject: Internet isolationism
:
During the last few podcasts you've discussed various from of Internet isolationism, if Elon Musk Starlink becomes a reality, it could shake things up, in both bad and good ways.

Walter Pereira in Brazil
Subject: Small correction on episode...
:
Hi Steve,

I'm a more than decade long user of SpinRite. The best piece of software I have, period. Not once failed me. As far as I remember, the only software I own with that record. Thanks so much!

But I'm writing to suggest a very small correction on what was affirmed by Leo in episode 698, when you were talking about Apple and Facebook quarrel. Leo said: "... if Apple really cared about this, it seems to me, they would at least give you the option to use DuckDuckGo in Safari. Right?"

I'm a long time user of DuckDuckGo, and it is my default search engine for iPhone since day one. Is a very trivial setting, (Safari > Search).

Best regards,
Walter

Authenticity on the Internet:
Textual AI loose on the Internet

A blog post published last Thursday, summarizing the recent work of six AI researchers at Elon Musk's OpenAI project brought me up short... and immediately became the topic for this week's podcast:

https://blog.openai.com/better-language-models/

I suppose we should have seen this coming. For quite some time we've had 'Bots' roaming the Internet indexing its pages and also warring with each other.

Then, eight years ago we sat silent, witnessing the stunning performance and success of IBM's Watson on Jeopardy. it was somewhat intimidating since I couldn't have answered most of those questions.

But it's one thing for a stylish black cube to sit idly by waiting to be asked a question -- or diagnose cancer -- and another thing entirely for text-generating AI bots to endlessly roam the Internet reading what they find and pouring out non-human generated content under the guise of being human.

For quite a while we've been living with photos being "photoshopped", and more recently we've been growing aware of the possibility that someone's clearly recognized voice might be saying something they never uttered. But in the jargon used on this podcast we would say that those attacks are "targeted". So, while disturbing, their reach is inherently limited. But not so if text-generating AI bots are let loose on the Internet.

So let's back up a bit... OpenAI's Wikipedia page describes "OpenAI" as a "non-profit artificial intelligence research organization that aims to promote and develop friendly AI in such a way as to benefit humanity as a whole." But that description may need to be rewritten.

Last Thursday's blog posting was titled: "Better Language Models and Their Implications"

The descriptive blurb reads: "We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization — all without task-specific training."

I'm going to share a sample of what this thing spits out, but let me first create a deeper context by sharing the description of what they have built:

Our model, called GPT-2 (a successor to GPT), was trained simply to predict the next word in 40GB of Internet text. Due to our concerns about malicious applications of the technology, we are not releasing the trained model. As an experiment in responsible disclosure, we are instead releasing a much smaller model for researchers to experiment with, as well as a technical paper.

GPT-2 is a large transformer-based language model with 1.5 billion parameters, trained on a dataset of 8 million web pages. GPT-2 is trained with a simple objective: predict the next word, given all of the previous words within some text. The diversity of the dataset causes this simple goal to contain naturally occurring demonstrations of many tasks across diverse domains. GPT-2 is a direct scale-up of GPT, with more than 10X the parameters and trained on more than 10X the amount of data.

GPT-2 displays a broad set of capabilities, including the ability to generate conditional synthetic text samples of unprecedented quality, where we prime the model with an input and have it generate a lengthy continuation. In addition, GPT-2 outperforms other language models trained on specific domains (like Wikipedia, news, or books) without needing to use these domain-specific training datasets. On language tasks like question answering, reading comprehension, summarization, and translation, GPT-2 begins to learn these tasks from the raw text, using no task-specific training data. While scores on these downstream tasks are far from state-of-the-art, they suggest that the tasks can benefit from unsupervised techniques, given sufficient (unlabeled) data and compute.
Samples

GPT-2 generates synthetic text samples in response to the model being primed with an arbitrary input. The model is chameleon-like — it adapts to the style and content of the conditioning text. This allows the user to generate realistic and coherent continuations about a topic of their choosing, as seen by the following select samples.

System Prompt (human-written):
In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Model Completion (machine-written, 10 tries):
The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

As the above samples show, our model is capable of generating samples from a variety of prompts that feel close to human quality and show coherence over a page or more of text.

Policy Implications

Large, general language models could have significant societal impacts, and also have many near-term applications. We can anticipate how systems like GPT-2 could be used to create:

- AI writing assistants
- More capable dialogue agents
- Unsupervised translation between languages
- Better speech recognition systems

We can also imagine the application of these models for malicious purposes, including the following (or other applications we can't yet anticipate):

- Generate misleading news articles
- Impersonate others online
- Automate the production of abusive or faked content to post on social media
- Automate the production of spam/phishing content

These findings, combined with earlier results on synthetic imagery, audio, and video, imply that technologies are reducing the cost of generating fake content and waging disinformation campaigns. The public at large will need to become more skeptical of text they find online, just as the "deep fakes" phenomenon calls for more skepticism about images[3].

Today, malicious actors — some of which are political in nature — have already begun to target the shared online commons, using things like "robotic tools, fake accounts and dedicated teams to troll individuals with hateful commentary or smears that make them afraid to speak, or

difficult to be heard or believed". We should consider how research into the generation of synthetic images, videos, audio, and text may further combine to unlock new as-yet-unanticipated capabilities for these actors, and should seek to create better technical and non-technical countermeasures. Furthermore, the underlying technical innovations inherent to these systems are core to fundamental artificial intelligence research, so it is not possible to control research in these domains without slowing down the progress of AI as a whole.
Release Strategy

Due to concerns about large language models being used to generate deceptive, biased, or abusive language at scale, we are only releasing a much smaller version of GPT-2 along with sampling code. We are not releasing the dataset, training code, or GPT-2 model weights. Nearly a year ago we wrote in the OpenAI Charter: "we expect that safety and security concerns will reduce our traditional publishing in the future, while increasing the importance of sharing safety, policy, and standards research," and we see this current work as potentially representing the early beginnings of such concerns, which we expect may grow over time. This decision, as well as our discussion of it, is an experiment: while we are not sure that it is the right decision today, we believe that the AI community will eventually need to tackle the issue of publication norms in a thoughtful way in certain research areas. Other disciplines such as biotechnology and cybersecurity have long had active debates about responsible publication in cases with clear misuse potential, and we hope that our experiment will serve as a case study for more nuanced discussions of model and code release decisions in the AI community.

We are aware that some researchers have the technical capacity to reproduce and open source our results. We believe our release strategy limits the initial set of organizations who may choose to do this, and gives the AI community more time to have a discussion about the implications of such systems.

We also think governments should consider expanding or commencing initiatives to more systematically monitor the societal impact and diffusion of AI technologies, and to measure the progression in the capabilities of such systems. If pursued, these efforts could yield a better evidence base for decisions by AI labs and governments regarding publication decisions and AI policy more broadly.

Tomorrow is here. :-(

~30~