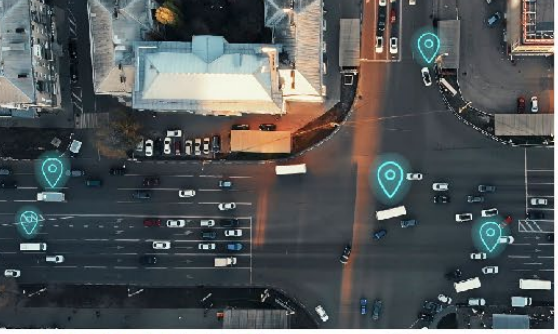
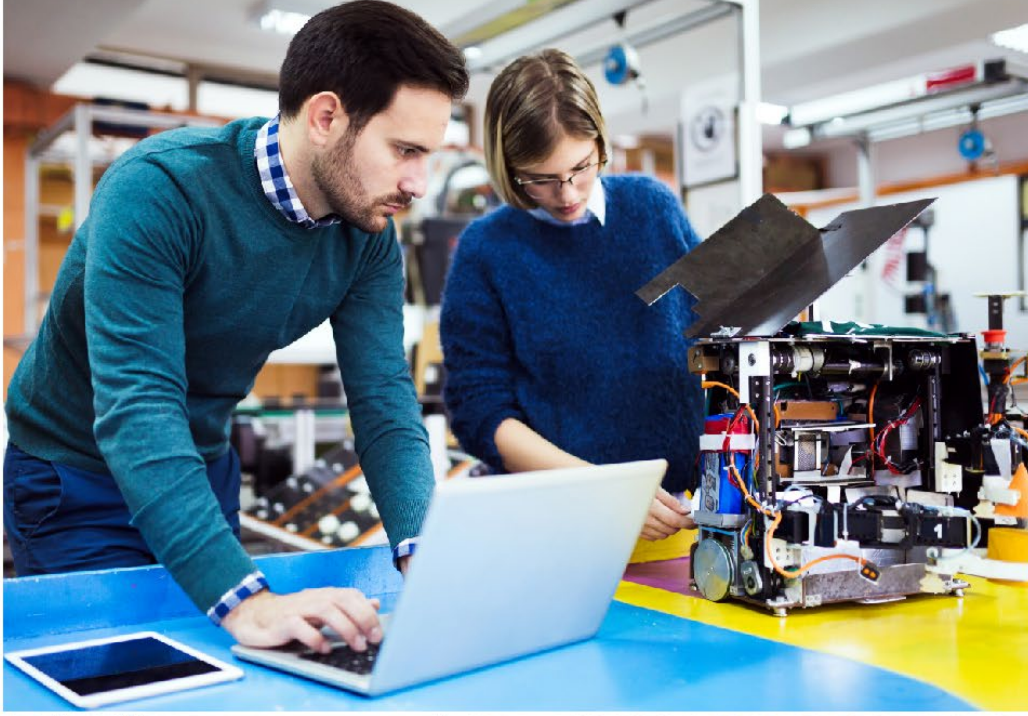
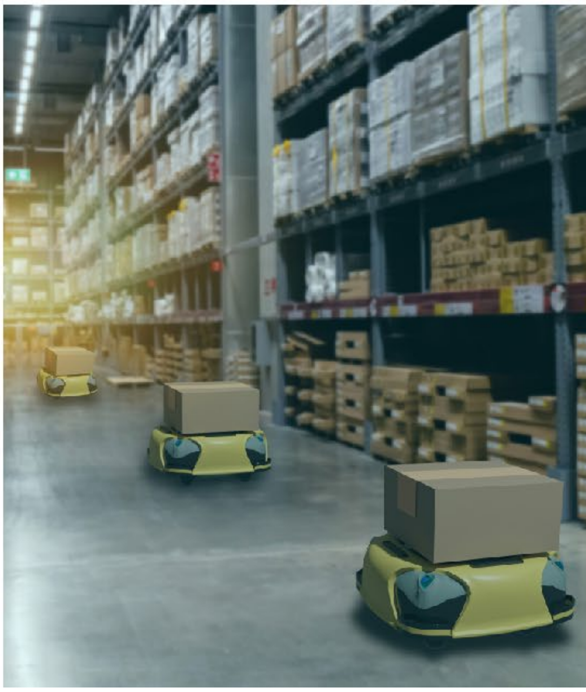






Malicious Uses and Abuses of Artificial Intelligence

Trend Micro Research

United Nations Interregional Crime and Justice Research Institute (UNICRI)

Europol's European Cybercrime Centre (EC3)

TREND MICRO

unicri
United Nations Interregional Crime and Justice Research Institute

CENTRE FOR ARTIFICIAL INTELLIGENCE AND ROBOTICS

EUROPOL
EC3 European Cybercrime Centre

TREND MICRO LEGAL DISCLAIMER

The information provided herein is for general information and educational purposes only. It is not intended and should not be construed to constitute legal advice. The information contained herein may not be applicable to all situations and may not reflect the most current situation. Nothing contained herein should be relied on or acted upon without the benefit of legal advice based on the particular facts and circumstances presented and nothing herein should be construed otherwise. Trend Micro reserves the right to modify the contents of this document at any time without prior notice.

Translations of any material into other languages are intended solely as a convenience. Translation accuracy is not guaranteed nor implied. If any questions arise related to the accuracy of a translation, please refer to the original language official version of the document. Any discrepancies or differences created in the translation are not binding and have no legal effect for compliance or enforcement purposes.

Although Trend Micro uses reasonable efforts to include accurate and up-to-date information herein, Trend Micro makes no warranties or representations of any kind as to its accuracy, currency, or completeness. You agree that access to and use of and reliance on this document and the content thereof is at your own risk. Trend Micro disclaims all warranties of any kind, express or implied. Neither Trend Micro nor any party involved in creating, producing, or delivering this document shall be liable for any consequence, loss, or damage, including direct, indirect, special, consequential, loss of business profits, or special damages, whatsoever arising out of access to, use of, or inability to use, or in connection with the use of this document, or any errors or omissions in the content thereof. Use of this information constitutes acceptance for use in an “as is” condition.

UNICRI DISCLAIMER

The opinions, findings, conclusions and recommendations expressed herein do not necessarily reflect the views of UNICRI, or any other the national, regional or international entity involved.

The designation employed and material presented in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area of its authorities, or concerning the delimitation of its frontiers or boundaries.

Contents of this publication may be quoted or reproduced, provided that the source of information is acknowledged. UNICRI would like to receive a copy of the document in which this publication is used or quoted.

EUROPOL DISCLAIMER

This publication and more information on Europol and the European Cybercrime Centre are available on the internet.

Published by
Trend Micro Research

Written by
Vincenzo Ciancaglini, Craig Gibson, and David Sancho
Trend Micro Research

Odhran McCarthy and Maria Eira
UNICRI

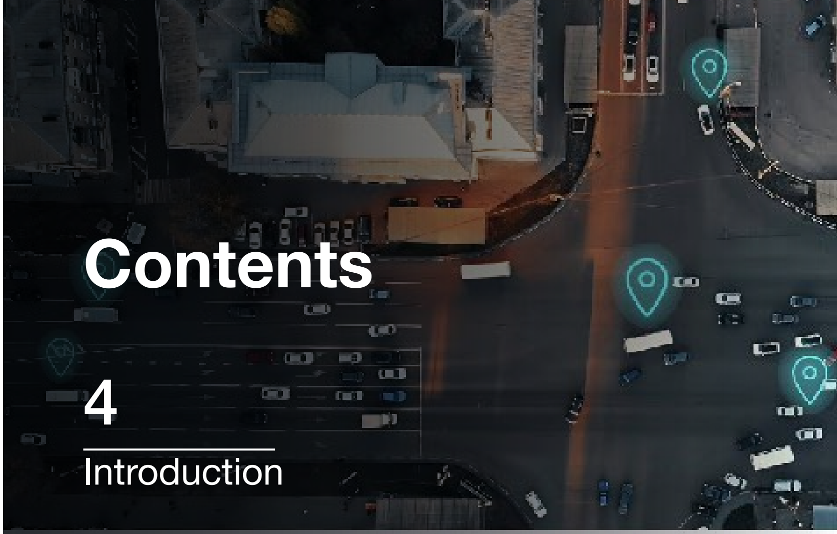
Philipp Amann and Aglika Klayn
Europol’s European Cybercrime Centre (EC3)

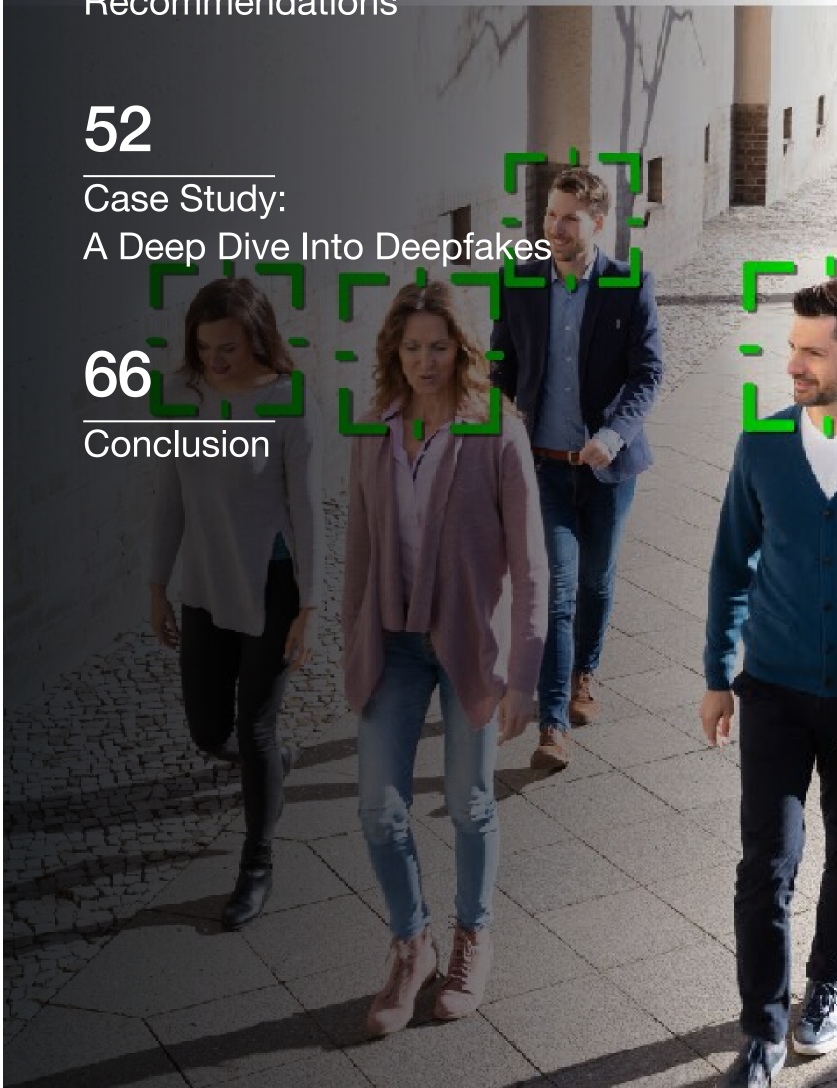
Management
Robert McArdle
Trend Micro Research

Irakli Beridze
UNICRI

Philipp Amann
Europol’s European Cybercrime Centre (EC3)

Stock images used under license from Shutterstock.com





Contents

The emergence of new technologies is shaping the world in an increasing range of sectors. Though there are many ways to define these technologies and their varying functions, it is possible to sum them up briefly.[1, 2, 3] Artificial intelligence, (AI) in particular, is ubiquitous in its applications and holds great promise to address a number of complex global challenges. By capitalizing on the unprecedented quantities of data, AI has shown potential from navigation and content recommendations to detecting cancer more accurately than human radiologists.

AI is a subfield of computer science (with many cross relationships to other disciplines) dedicated to the theory and development of computer systems that can perform tasks normally requiring human intelligence, such as visual perception, speech recognition, translation between languages, decision-making, and problem-solving. Machine learning (ML), itself a subfield of AI, consists of algorithms that use statistical techniques to give computer systems the ability to “learn” from data — that is to say, to progressively improve performance on a specific task. Compared to other computer software, ML algorithms do not require explicit instructions from humans but rather extract patterns and learn implicit rules from a considerable number of examples.

While AI and ML algorithms can bring enormous benefits to society, these technologies can also enable a range of digital, physical, and political threats. Just as the World Wide Web brought a plethora of new types of crime to the fore and facilitated a range of more non-traditional ones, AI stands poised to do the same.[4, 5] In the continuous shift from analogue to digital, the potential for the malicious use of new technologies is also exposed.

Hence, while this report looks at the present state of both AI and ML technologies, it also seeks to predict the possible ways that criminals will exploit these technologies in the future — a task that though seemingly daunting, is paramount for the cybersecurity industry and law enforcement to undertake in the never-ending challenge to always stay one step ahead of criminals.

Introduction

In the dynamic world of technology and computer science, AI continues to offer a wide range of possible applications for enterprises and individuals. Unfortunately, the promise of more efficient automation and autonomy is inseparable from the different schemes that malicious actors are capable of.

For instance, criminals can use AI to facilitate and improve their attacks by maximizing opportunities for profit in a shorter time, exploiting new victims, and creating more innovative criminal business models while reducing the chances of being caught. Additionally, as AI-as-a-Service becomes more widespread, it will lower the barrier to entry by reducing the skills and technical expertise needed to employ AI.

Criminals and organized crime groups (OCGs) have been swiftly integrating new technologies into their modi operandi, thus creating not only constant shifts in the criminal landscape worldwide, but also creating significant challenges for law enforcement and cybersecurity in general.[7] The Crime-as-a-Service (CaaS) business model, which allows non-technologically savvy criminals to procure technical tools and services in the digital underground that allow them to extend their attack capacity and sophistication,[8] further increases the potential for new technologies such as AI to be abused by criminals and become a driver of crime.[9, 10, 11]

Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behavior by analyzing how the environment is affected by their previous actions.[6]

(The British spelling "analysing" is used in the reference for this definition.)

Building knowledge about the potential use of AI by criminals will improve the ability of the cybersecurity industry in general and law enforcement agencies in particular to anticipate possible malicious and criminal activities, as well as to prevent, respond to, or mitigate the effects of such attacks in a proactive manner. An understanding of the capabilities, scenarios, and attack vectors is key to enhancing preparedness and increasing resilience.

In line with the goal to contribute to the body of knowledge on AI, this report, a joint effort among Trend Micro, the United Nations Interregional Crime and Justice Research Institute (UNICRI), and Europol, seeks to provide a thorough and in-depth look into the present and possible future malicious uses and abuses of AI and related technologies.

The intended audience of this report includes cybersecurity experts and practitioners, law enforcement, innovation hubs, policy-makers, and researchers.

The report can also be leveraged by members of this audience as a thinking space for ideas and a call for greater attention to the possible malicious uses or abuses of AI.

The findings contained in the report are based on contributions from the three entities. They have been combined with input collected during a focused workshop in March 2020 that was organized by Europol, Trend Micro, and UNICRI. Workshop participants included members from the Joint Cybercrime Action Taskforce (J-CAT),[12] the International Criminal Court,[13] and several members of Europol's European Cybercrime Centre (EC3) Advisory Groups.[14]

This report uses an important distinction that should be noted from the outset: namely, the distinction between malicious uses and abuses of AI. The "uses" in "malicious AI uses" refers to instances whereby criminals might employ AI systems to further their attack objectives — for example, by using ML to automate cyberattacks and improve malware. On the other hand, the "abuses" in "malicious AI abuses" refers to instances where criminals might try to attack and exploit existing AI systems to break or circumvent them — for example, by hacking smart home assistants.

Machine learning (ML) is a subset of AI, where algorithms are trained to infer certain patterns based on a set of data in order to determine the actions needed to achieve a given goal.[15]

With this distinction in mind, we framed the report around two main components: *present* AI malicious use or abuse, for which there are documented cases and research, and *future* malicious use or abuse for which there is no evidence or literature yet. Nonetheless, we believe that on the basis of technological trends and developments, future uses or abuses could become present realities in the not-too-distant future. To prepare for these uses or abuses, speculative scenarios that are likely to happen must be conceptualized. To ground the report with regard to this, we examine current trends in underground forums, as these can provide insights on what the malicious abuse of the AI threat landscape might look like in the near future.

With respect to present malicious uses or abuses, possible countermeasures to combat the malicious uses of AI are also identified here. It should be emphasized, however, that these countermeasures are not exhaustive and are only suggestive of one or more possible avenues for combatting the specific use or abuse identified.

Finally, through a case study at the end of this report, we take a closer look at deepfakes, a specific use of AI that has received widespread media attention and presents considerable potential for a range of malicious and criminal purposes. We also include an overview of the technological aspects of some specific use cases that are identified in the report, alongside further technical content, in this case study.

The Present State of Malicious Uses and Abuses of AI

A significant leap forward in AI has been observed in recent years. While this paper outlines possible future developments of malicious uses and abuses of AI technology, a prediction like this has value only when it is properly grounded in the present.

To this end, this first section presents the current developments of AI in the scope of possible malicious uses and abuses by covering research outcomes, proofs of concept, or discussions among criminals, provided that they are documented and verifiable. The section includes both cases of criminals exploiting AI technology for their own gain and criminal attempts to abuse legitimate AI systems that are readily available for malicious AI abuses.

AI Malware

The use of AI to improve the effectiveness of malware is still in its infancy. Research is still carried out at the academic level and attacks are mostly theoretical and crafted as proofs of concept by security researchers. Nonetheless, the AI-supported or AI-enhanced cyberattack techniques that have been studied are proof that criminals are already taking steps to broaden the use of AI. Such attempts therefore warrant further observation to stop these current attempts and prepare for future attacks as early as possible before these become mainstream.

Currently, malware developers can use AI in more obfuscated ways without being detected by researchers and analysts. As a consequence, it is only possible to search for observable signs that might be expected from AI malware activity. In fact, one type of malware-related AI exploit involves AI-based techniques aimed at improving the efficacy of "traditional" cyberattacks.

For example, in 2015, a demonstration on how to craft email messages in order to bypass spam filters was made.[16] This system, as demonstrated, uses generative grammar capable of creating a large dataset of email texts with a high degree of semantic quality. These texts are then used to fuzz the antispam system and adapt to different spam filters in order to identify content that would no longer be detected by the spam filters.

In 2017, during Black Hat USA, an information security conference,[17] researchers demonstrated how to use ML techniques to analyze years' worth of data related to business email compromise (BEC) attacks, a form of cybercrime that uses email fraud to scam organizations in order to identify potential attack targets. This system exploits both data leaks and openly available social media information. Notably, based on its history, the system can accurately predict if an attack will be successful or not.

At the same security conference, researchers introduced AVPASS,[18] a tool designed to infer, for any given antivirus engine, its detection features and detection rule chain. The tool then uses this inference to disguise Android malware as a benign application. It should be emphasized that AVPASS achieved a 0% detection rate on the online malware analysis service VirusTotal with more than 5,000 Android malware samples. In other words, AVPASS created operationally undetectable malware.

At present, antivirus vendors also look at ML as their tool of choice for improving their malware detection techniques, thanks to its ability to generalize new types of malware that have never been seen before. However, it has been proven that ML-based detection systems can be tricked by an AI agent designed to probe and find weak spots.[19] Researchers, for instance, have been able to craft malware with features that allow it to remain undetected even by ML-based antivirus engines. The system uses reinforcement learning to develop a competitive, game-based technique between itself and the antivirus detector. It also selects functionality-preserving features of a malicious Windows file and introduces variants that increase the chances of a malware sample passing undetected.

Reinforcement learning refers to the training that an agent undergoes to learn a particular behavior in a dynamic environment. Since the agent is not oriented beforehand on the different outcomes for every action that it can take, it must rely on both the knowledge that it gains through experience and the rewards that it receives for every action that it takes in order to learn the behavior.[20]

Finally, AI can also enhance traditional hacking techniques by introducing new ways of performing attacks that would be difficult for humans to predict. At DEF CON 2017, one of the largest underground hacking conventions, participants Dan Petro and Ben Morris presented DeepHack,[21] an open-source AI tool aimed at performing web penetration testing without having to rely on any prior knowledge of the target system. DeepHack implements a neural network capable of crafting SQL injection strings with no information other than the target server responses, thereby automating the process of hacking web-based databases.

Following a similar approach, DeepExploit[22] is a system that is capable of fully automating penetration testing by using ML. The system interfaces directly with Metasploit, a penetration testing platform, for all the usual tasks of information gathering and crafting and testing an exploit. However, it leverages a reinforcement learning algorithm named Asynchronous Actor-Critic Agents (AC3)[23] in order to learn first (from openly exploitable services such as Metasploitable) which exploit should be used under specific conditions, before testing such conditions on the target server.

In the previously described attacks, AI helps to improve and optimize techniques traditionally used by criminals. Conversely, the second kind of attacks focuses on subverting existing AI systems in order to alter their capabilities or behavior.

For instance, IBM researchers recently presented a novel approach to exploiting AI for malicious purposes. Their system, DeepLocker,[24] embeds AI capabilities *within* the malware itself in order to improve its evasion techniques. This is accomplished by employing a general weakness of most of AI algorithms — that is to say, their *lack of explicability* — to the advantage of the system.

By their own nature, ML algorithms turn out to be, from an implementation point of view, nothing more than multiplications of factors and matrices, whereas an array of factors constituting the input is multiplied by one or several matrices. In turn, the value of these matrices is determined in the training phase of the algorithm, thus leading to the output value of the said algorithm. This applies to almost all supervised machine learning algorithms, with very few exceptions. Because of this, it is generally very hard to explain, in simple language, why a machine-learning algorithm trained with a given dataset would take a specific decision. To illustrate, if we take the example of a spam filter relying on deep learning to identify spam content, it would be very difficult to explain why it flags some content as spam. Likewise, it would be challenging to debug and fix a false positive in which a legitimate message is erroneously marked as spam.

Instead, one normally assumes that the algorithm is a “black box” and that the system design revolves around picking the right black box, tuning its functioning parameters, and finding the proper training set for the given problem.

DeepLocker exploits this inner inscrutability to its own advantage: Malware generally employs multiple techniques to evade detection, from payload obfuscation and encryption, to special checks to frustrate analysis. For example, in targeted attacks, malware often performs specific checks in order to activate only when running on the target machine. These checks, however, are the very last clue that can give away that a piece of software is a targeted malware and can serve as a signature for the malware itself.

DeepLocker manages to obfuscate this last piece of malicious behavior by implementing the said checks through a deep neural network (DNN), thus exploiting the aforementioned “inexplicable” nature of the algorithm to mask those checks as simple mathematical operations between matrices.

As a result, it becomes harder to reverse-engineer the malware behavior or even to trigger it in a test environment, since the trigger conditions are never explicitly encoded in the malware. Consequently, analysis also becomes that much more difficult. DeepLocker, for example, uses many attributes to identify the target machine, including software environment, geolocation, and even audio and video, thus adding to the difficulty in predicting and identifying the trigger conditions for the malware to execute.

DeepLocker brings the obfuscation one step further: Its neural network does not output an explicit activation value. Neither does it answer the question, “Am I on the target machine?” with “Yes” or “No.” Rather, it outputs a decryption key that can be used, when the input conditions are matched, to decrypt the remaining malware payload.

In this way, any explicit information about the malware target is obfuscated in the form of matrix factors, making it next to impossible to reverse-engineer and identify the target. Until the very moment that the malware sits on the target, the rest of its code is fully encrypted and inscrutable.

It is worth noting that attacks such as DeepLocker are difficult to stop as a defender would not know what to look for until the malware is completely decrypted. Similarly, on the target machine, DeepLocker can be stopped by detecting the attack only once it is decrypted and about to run. An antivirus signature in-memory, a behavioral rule, or a firewall rule, would thus be able to see the attack as it happens. Nonetheless, the problem could be much worse if the malware were to be analyzed on a test machine, or if an engineer were to decide whether a dataset is malicious. In such cases, DeepLocker would not decrypt correctly and as a result, the engineer would not be able to see the malicious functions for a proper analysis. If the malware is never correctly analyzed, those in-memory or behavioral rules would thus never be made.

Meanwhile, attacks such as DeepHack and DeepExploit are similarly difficult to detect. Regardless, a well-configured application firewall can help to defend a network against these attacks and even disallow further connections from the AI-enabled attack.

AI Malware at Large

In order to see if malware writers are already using currently available ML algorithms, we used the following approach:

1. We typified popular ML engines and created detection rules using YARA,[25] the industry standard for rule-based behavioral detection of files such as malware.
2. We ran RetroHunt searches with those rulesets on VirusTotal to see if any of those engines had been somehow modified or tampered with.
3. We examined results for those files with traces of malicious code. To avoid false positives, we needed at least two detections from different antivirus engines.

The most popular and easily accessible ML engines are the following:

- TensorFlow + Keras
- RapidMiner
- PyTorch

Both TensorFlow and PyTorch are open-source projects written in Python. Since they are so easily modifiable, the defined YARA rules are very generic and look specifically for recognizable strings in text files or executables. RapidMiner, on the other hand, is a Java-based engine; therefore, the corresponding YARA rule looks for certain relevant strings within Java Archive (JAR) files.

Our search for malicious TensorFlow-, RapidMiner-, or PyTorch-related files did not yield any files. The first attempts resulted in several files that were later proven to be false positives. After refining the initial rules and looking for certain strings related to each of these frameworks, we did not obtain any additional results.

In the case of TensorFlow and PyTorch, the rules are very similar. We looked for Python files (with extension .py) or for portable executables (PE) to account for compiled Python binary files on Windows platforms. The rules, as further detailed in Appendix A, are designed to match certain strings that were identified to be in the open source code of their respective projects.

Although it is impossible to prove for certain that malware using ML does not exist, the search for malicious files that resemble popular ML frameworks did not provide any results at the point of writing.

Abusing AI Cloud Services

It is important to note that ML algorithms do not necessarily need to run on the same host machine where the malware runs. Another approach to detect if malicious actors have already started leveraging ML when developing malware is to check, therefore, whether a particular malware connects to cloud-based AI services.

To follow this approach, we compiled a list of all AI services offered by major cloud providers, including Amazon AWS, Microsoft Azure, Google Cloud, Alibaba Cloud, Yandex, and others. All of these have started offering AI services in recent years: For example, Amazon, through its AWS platform, offers image recognition services (Amazon Rekognition), unstructured text analytics (Amazon Comprehend), or named entity extraction (Amazon Textract), with competitors offering similar services of their own. Each of these services can be accessed via an HTTPS connection to a specific endpoint URL, which uses a specific pattern. For example, all services from Amazon's AWS platform contain "aws.amazon.com" in the hostname and the name of the service in the path.

Using those patterns, we performed an extensive search within Trend Micro's databases, focusing on two datasets in particular:

- The first dataset contains behavioral information of malware samples that have been analyzed in the past.
- The second dataset contains anonymized reports of connections performed by malicious software as detected by Trend Micro's antivirus software.

By using these patterns and focusing on the two datasets, it was possible to check if any novel malware samples found on the internet connects to AI cloud services and why, or if any malicious connection to said services that come from a malicious software on a victim's machine had been detected previously.

As of writing this report, our search did not identify any traces of malicious software exploiting cloud services. There are two possible reasons for this lack of identification. First, scaling matters (from the point of view of the malicious actor) might result in higher costs. Second, without proper measures, the move to exploit cloud services might increase chances of the malicious actor being revealed.

Nevertheless, cloud services should be monitored for this kind of connection, since the next iteration of AI malware targeting them might still emerge.

Abusing Smart Assistants

An alternative approach to attacking AI algorithms is to target AI assistants by either exploiting their presence in households or abusing their development model. In particular, an AI assistant could be targeted by developing adversarial systems or polluting datasets.

An example of an adversarial system is elaborated on in the Trend Micro paper "The Sound of a Targeted Attack,"[26] where an attacker could exploit exposed smart speakers to issue audio commands to a nearby smart assistant, such as Amazon Alexa or Google Home. If a speaker is connected to the internet, its vulnerabilities could be exploited, causing it to play an audio file hosted on an arbitrary web address set up by the attacker. It is also possible for the file to contain speech that issues a specific command to the nearby smart assistants. Additionally, a stealth attack might use an issued command that is not perceivable by the human ear.[27] Exposed devices can be easily found using services such as Shodan, a search engine for internet-connected devices and systems.

The possibility of such attacks is further exacerbated by the fact that smart assistants are often in control of home automation systems. In "Cybersecurity Risks in Complex IoT Environments: Threats to Smart Homes, Buildings and Other Structures,"[28] Trend Micro proves that hijacking a smart assistant through exposed audio devices is only one of the many attacks that can be carried out by a malicious actor interested in breaking into a smart home.

AI-Supported Password Guessing

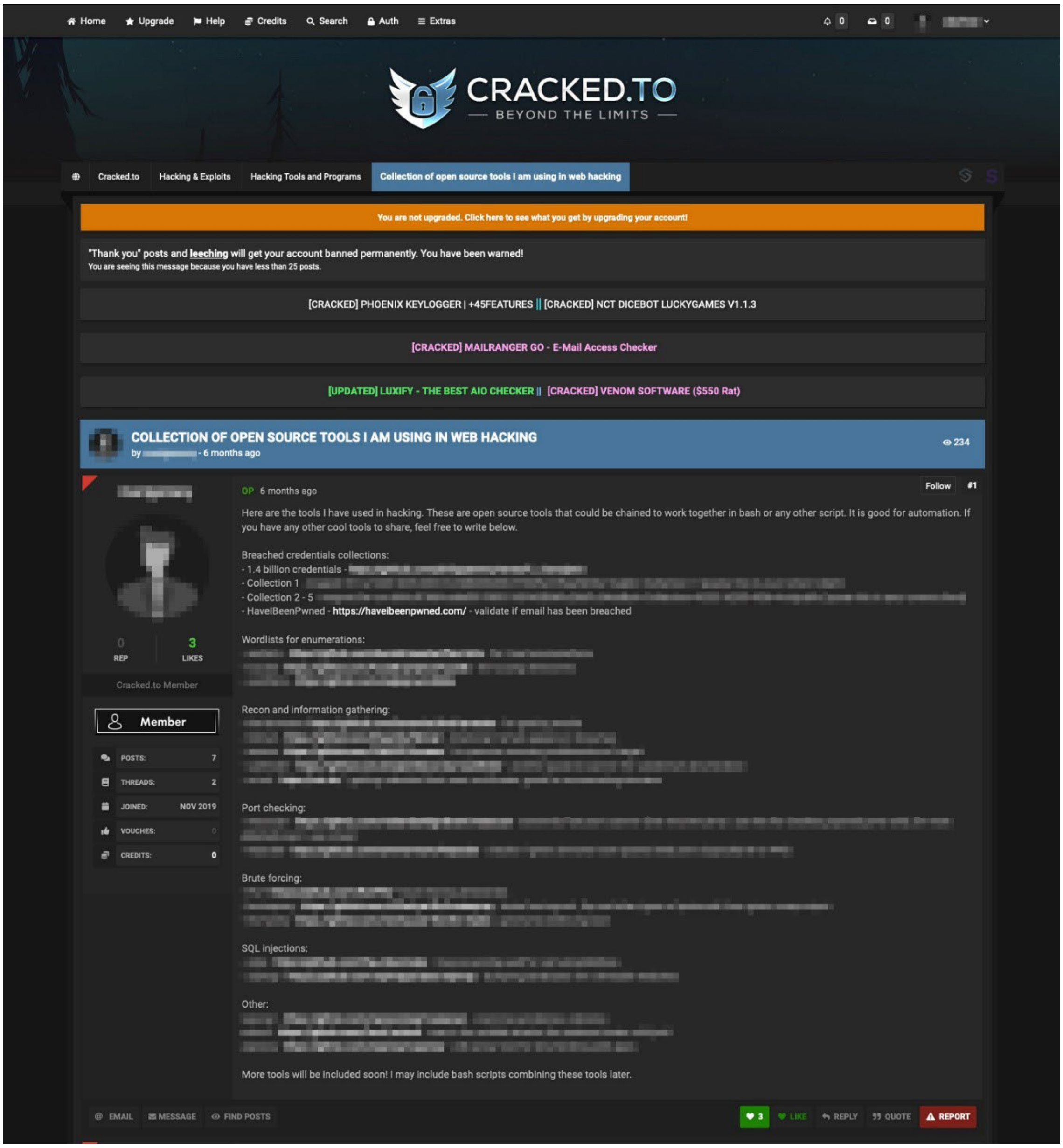
Another application involves employing ML to improve password-guessing algorithms. Despite the fact that this application is being researched with moderate success, it is important to note that this same application has already proven to be more efficient than more traditional approaches. It is also quite hard to detect on its own. As a result, it is not far-fetched to presume that AI-supported algorithms and tools are in constant conceptualization and development by individuals or groups who might abuse these.

Traditional password-guessing tools, such as HashCat[29] and John the Ripper,[30] usually work by comparing many different variations to the password hash in order to identify the password that corresponds to the hash. Attempts are generated from a dictionary of frequently used passwords; after, variations are made based on the composition of the password. For example, variations of "password" might be "password12345" or "p4ssw0rd."

Through the use of neural networks and generative adversarial networks (GANs) in particular, it is possible to analyze a large dataset of passwords and generate variations that fit the statistical distribution, such as for password leaks. This leads to more targeted and more effective password guesses.

A **generative adversarial** network is a type of neural network that consists of two neural network models: a generator and a discriminator, which the GAN uses to distinguish samples that are real from those samples that have been generated.[31]

An early attempt at this is already evident in a post on an underground forum from February 2020. The post mentions a GitHub repository from the same year, in which a software is able to parse through 1.4 billion credentials and generate password variation rules based on its findings.



You are not upgraded. Click here to see what you get by upgrading your account!

"Thank you" posts and leeching will get your account banned permanently. You have been warned!
You are seeing this message because you have less than 25 posts.

[CRACKED] PHOENIX KEYLOGGER | +45FEATURES || [CRACKED] NCT DICEBOT LUCKYGAMES V1.1.3

[CRACKED] MAILRANGER GO - E-Mail Access Checker

[UPDATED] LUXIFY - THE BEST AIO CHECKER || [CRACKED] VENOM SOFTWARE ($550 Rat)

COLLECTION OF OPEN SOURCE TOOLS I AM USING IN WEB HACKING
by [redacted] - 6 months ago
234

[redacted]

0 REP
3 LIKES
Cracked.to Member
Member

POSTS:	7
THREADS:	2
JOINED:	NOV 2019
VOUCHES:	0
CREDITS:	0

OP 6 months ago
Follow #1

Here are the tools I have used in hacking. These are open source tools that could be chained to work together in bash or any other script. It is good for automation. If you have any other cool tools to share, feel free to write below.

Breached credentials collections:
- 1.4 billion credentials - [redacted]
- Collection 1 [redacted]
- Collection 2 - 5 [redacted]
- HaveIBeenPwned - https://haveibeenpwned.com/ - validate if email has been breached

Wordlists for enumerations:
[redacted]

Recon and information gathering:
[redacted]

Port checking:
[redacted]

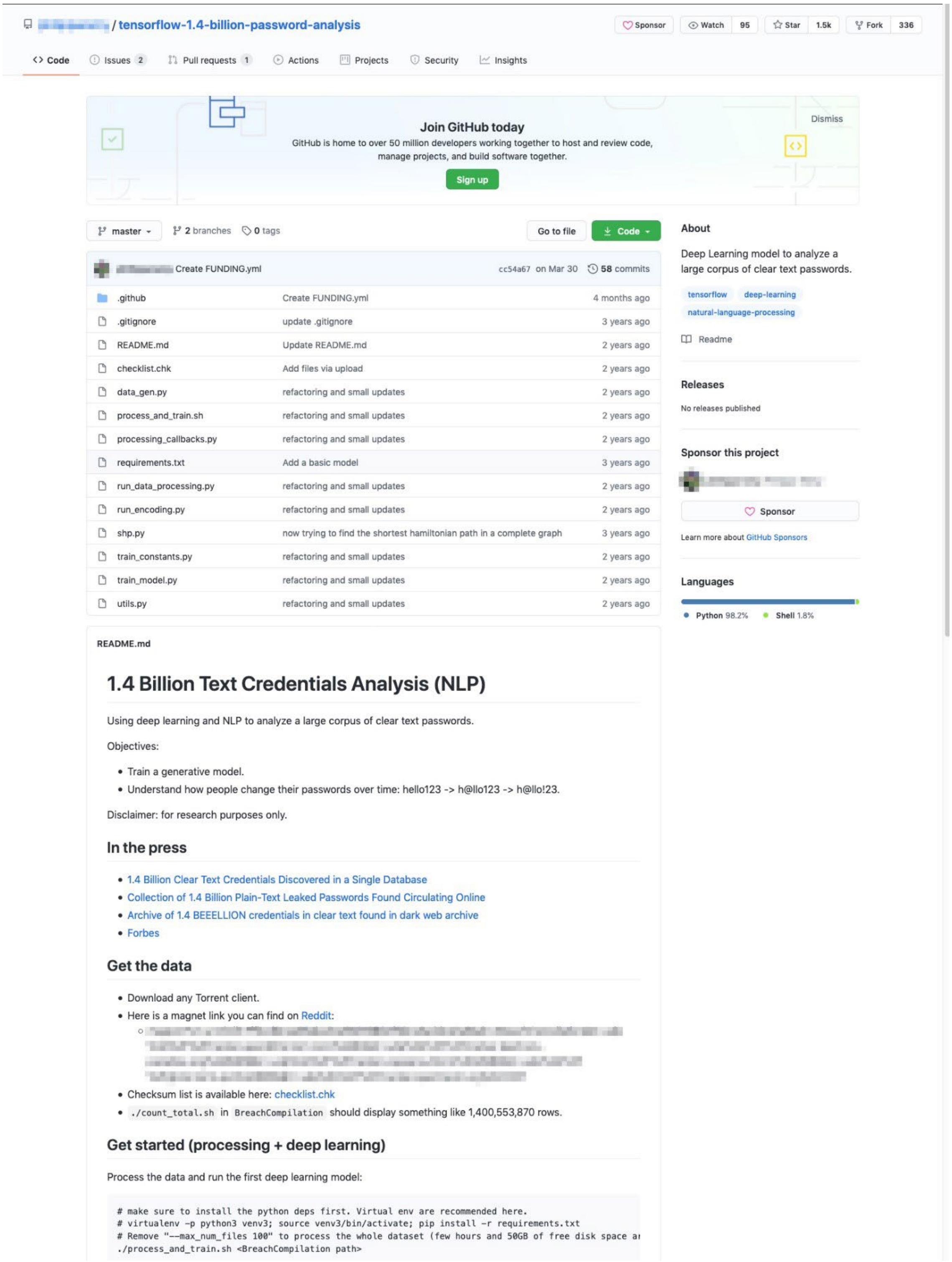
Brute forcing:
[redacted]

SQL injections:
[redacted]

Other:
[redacted]

More tools will be included soon! I may include bash scripts combining these tools later.

Figure 1. An underground post that mentions a password analysis tool


[redacted] / tensorflow-1.4-billion-password-analysis
Sponsor
Watch 95
Star 1.5k
Fork 336
Code
Issues 2
Pull requests 1
Actions
Projects
Security
Insights
Join GitHub today
GitHub is home to over 50 million developers working together to host and review code, manage projects, and build software together.
Sign up
Dismiss
master
2 branches
0 tags
Go to file
Code

[redacted] Create FUNDING.yml cc54a67 on Mar 30 58 commits

.github	Create FUNDING.yml	4 months ago
.gitignore	update .gitignore	3 years ago
README.md	Update README.md	2 years ago
checklist.chk	Add files via upload	2 years ago
data_gen.py	refactoring and small updates	2 years ago
process_and_train.sh	refactoring and small updates	2 years ago
processing_callbacks.py	refactoring and small updates	2 years ago
requirements.txt	Add a basic model	3 years ago
run_data_processing.py	refactoring and small updates	2 years ago
run_encoding.py	refactoring and small updates	2 years ago
shp.py	now trying to find the shortest hamiltonian path in a complete graph	3 years ago
train_constants.py	refactoring and small updates	2 years ago
train_model.py	refactoring and small updates	2 years ago
utils.py	refactoring and small updates	2 years ago

About

Deep Learning model to analyze a large corpus of clear text passwords.

tensorflow deep-learning natural-language-processing

Readme

Releases

No releases published

Sponsor this project

[redacted]

Sponsor

Learn more about GitHub Sponsors

Languages

- Python 98.2%
- Shell 1.8%

README.md

1.4 Billion Text Credentials Analysis (NLP)

Using deep learning and NLP to analyze a large corpus of clear text passwords.

Objectives:

- Train a generative model.
- Understand how people change their passwords over time: hello123 -> h@llo123 -> h@llo!23.

Disclaimer: for research purposes only.

In the press

- 1.4 Billion Clear Text Credentials Discovered in a Single Database
- Collection of 1.4 Billion Plain-Text Leaked Passwords Found Circulating Online
- Archive of 1.4 BEEELLION credentials in clear text found in dark web archive
- Forbes

Get the data

- Download any Torrent client.
- Here is a magnet link you can find on Reddit:
  - [redacted]
- Checksum list is available here: checklist.chk
- `./count_total.sh` in `BreachCompilation` should display something like 1,400,553,870 rows.

Get started (processing + deep learning)

Process the data and run the first deep learning model:

```
# make sure to install the python deps first. Virtual env are recommended here.
# virtualenv -p python3 venv3; source venv3/bin/activate; pip install -r requirements.txt
# Remove "--max_num_files 100" to process the whole dataset (few hours and 50GB of free disk space ar
./process_and_train.sh <BreachCompilation path>
```

Figure 2. A GitHub repository of password analysis tools

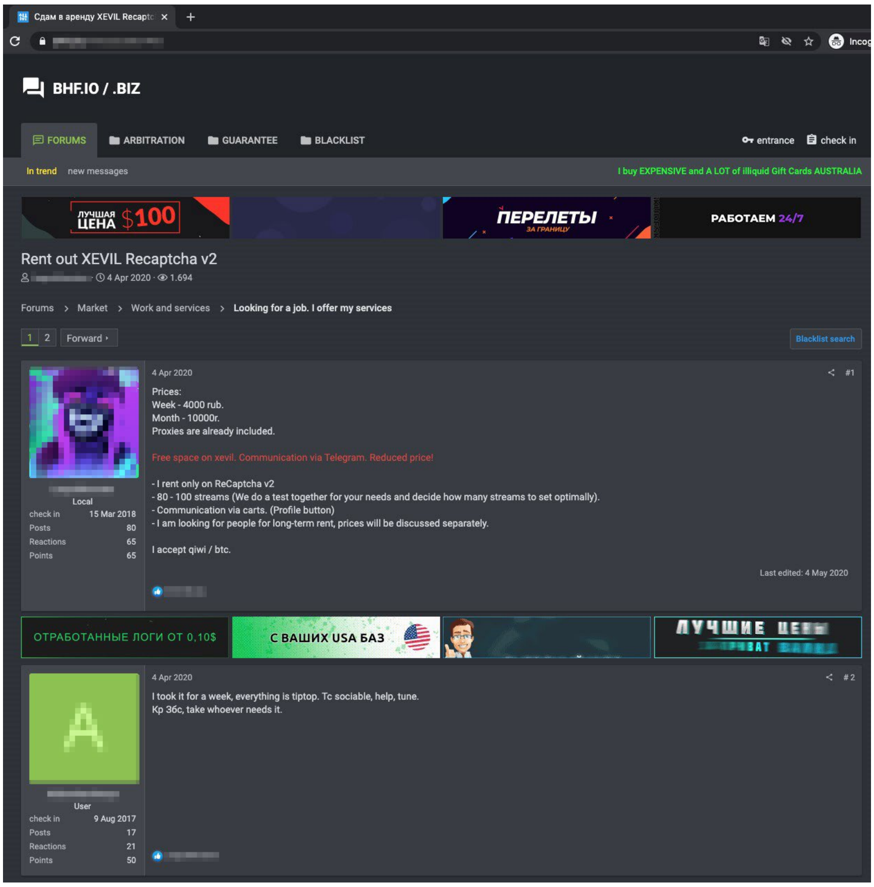
The PassGAN system, published in 2019,[32] presents a similar approach. It uses a GAN to learn the statistical distribution of passwords from password leaks and generates high-quality password guesses. This system was able to match 51% to 73% more passwords compared with HashCat alone.

The case study on deepfakes at the end of this paper contains more information about GANs.

AI-Supported CAPTCHA Breaking

The application of ML for breaking CAPTCHA security systems is frequently addressed on criminal forums. CAPTCHA images are commonly used on websites to thwart criminals when they attempt to abuse web services — particularly when they try to automate attacks (some attempts involve creating new accounts or adding new comments or replies on forums, among others). Developing systems that try to break CAPTCHA images to automate the abuse of those services would thus be a natural progression for cybercriminals.

Software that implements neural networks to solve CAPTCHAs, such as XEvil 4.0,[33] is currently being tested on criminal forums. Moreover, it has been claimed that this neural network can break human-recognition systems using CAPTCHA on Yandex pages. The tool utilizes 80 to 100 CPU threads (with the CPUs running in parallel) to speed up CAPTCHA solving. It is also advertised on Russian underground forums and rented out to users for 4,000 rubles weekly (approximately US$54 as of writing) or 10,000 rubles monthly (approximately US$136 as of writing).



Rent out XEVIL Recaptcha v2

[redacted] · 4 Apr 2020 · 1.694

Forums > Market > Work and services > **Looking for a job. I offer my services**

[redacted]
Local
check in 15 Mar 2018
Posts 80
Reactions 65
Points 65

4 Apr 2020 #1

Prices:
Week - 4000 rub.
Month - 10000r.
Proxies are already included.

Free space on xevil. Communication via Telegram. Reduced price!

- I rent only on ReCaptcha v2
- 80 - 100 streams (We do a test together for your needs and decide how many streams to set optimally).
- Communication via carts. (Profile button)
- I am looking for people for long-term rent, prices will be discussed separately.

I accept qiwi / btc.

Last edited: 4 May 2020

[redacted]

[redacted]
User
check in 9 Aug 2017
Posts 17
Reactions 21
Points 50

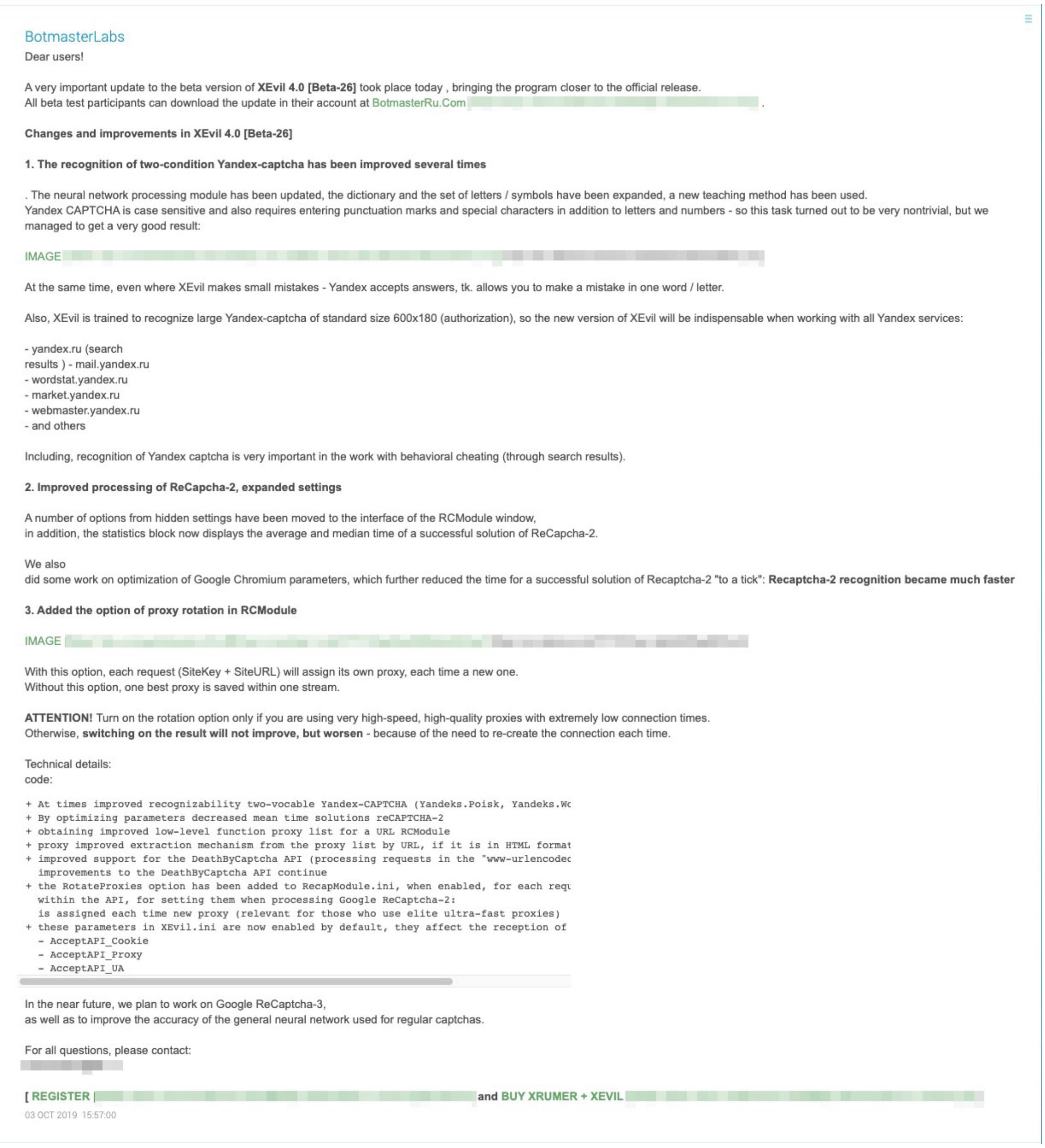
4 Apr 2020 #2

I took it for a week, everything is tiptop. Tc sociable, help, tune.
Kp 36c, take whoever needs it.

[redacted]

Figure 3. An underground post of a CAPTCHA-breaking tool (translated from Russian)

In order to make it more difficult to break CAPTCHA and diminish the efficiency of AI algorithms, CAPTCHA developers should increase the variety of color patterns and shapes. However, given the way that the anti-CAPTCHA defense engines work, it is possible that doing so would cause only a very slight delay in the time it takes for criminals to break the CAPTCHAs. Barring that, perhaps it would be worth developing some sort of software filter to detect scripts or programs that repeatedly try to solve CAPTCHAs too quickly. This software filter could also be developed to send completely fake, unsolvable images to these scripts or programs in order to try and poison their datasets.



BotmasterLabs

Dear users!

A very important update to the beta version of **XEvil 4.0 [Beta-26]** took place today , bringing the program closer to the official release.
All beta test participants can download the update in their account at BotmasterRu.Com [redacted] .

Changes and improvements in XEvil 4.0 [Beta-26]

1. The recognition of two-condition Yandex-captcha has been improved several times

. The neural network processing module has been updated, the dictionary and the set of letters / symbols have been expanded, a new teaching method has been used.
Yandex CAPTCHA is case sensitive and also requires entering punctuation marks and special characters in addition to letters and numbers - so this task turned out to be very nontrivial, but we managed to get a very good result:

IMAGE [redacted]

At the same time, even where XEvil makes small mistakes - Yandex accepts answers, tk. allows you to make a mistake in one word / letter.

Also, XEvil is trained to recognize large Yandex-captcha of standard size 600x180 (authorization), so the new version of XEvil will be indispensable when working with all Yandex services:

- yandex.ru (search
results) - mail.yandex.ru
- wordstat.yandex.ru
- market.yandex.ru
- webmaster.yandex.ru
- and others

Including, recognition of Yandex captcha is very important in the work with behavioral cheating (through search results).

2. Improved processing of ReCapcha-2, expanded settings

A number of options from hidden settings have been moved to the interface of the RCModule window,
in addition, the statistics block now displays the average and median time of a successful solution of ReCapcha-2.

We also
did some work on optimization of Google Chromium parameters, which further reduced the time for a successful solution of Recaptcha-2 "to a tick": **Recaptcha-2 recognition became much faster**

3. Added the option of proxy rotation in RCModule

IMAGE [redacted]

With this option, each request (SiteKey + SiteURL) will assign its own proxy, each time a new one.
Without this option, one best proxy is saved within one stream.

ATTENTION! Turn on the rotation option only if you are using very high-speed, high-quality proxies with extremely low connection times.
Otherwise, **switching on the result will not improve, but worsen** - because of the need to re-create the connection each time.

Technical details:
code:

```
+ At times improved recognizability two-vocable Yandex-CAPTCHA (Yandeks.Poisk, Yandeks.Wc
+ By optimizing parameters decreased mean time solutions reCAPTCHA-2
+ obtaining improved low-level function proxy list for a URL RCModule
+ proxy improved extraction mechanism from the proxy list by URL, if it is in HTML format
+ improved support for the DeathByCaptcha API (processing requests in the "www-urlencodec
  improvements to the DeathByCaptcha API continue
+ the RotateProxies option has been added to RecapModule.ini, when enabled, for each requ
  within the API, for setting them when processing Google ReCaptcha-2:
  is assigned each time new proxy (relevant for those who use elite ultra-fast proxies)
+ these parameters in XEvil.ini are now enabled by default, they affect the reception of
  - AcceptAPI_Cookie
  - AcceptAPI_Proxy
  - AcceptAPI_UA
```

In the near future, we plan to work on Google ReCaptcha-3,
as well as to improve the accuracy of the general neural network used for regular captchas.

For all questions, please contact:
[redacted]

[REGISTER | [redacted] and BUY XRUMER + XEVIL [redacted]

03 OCT 2019 15:57:00

Figure 4. The underground CAPTCHA-breaking tool called XEvil

AI-Aided Encryption

According to Europol's 2019 report titled, "First Report of the Observatory Function on Encryption,"[34] AI applications related to improving or breaking encryption are still in their infancy. However, it is imperative to learn more about AI-aided encryption since there is already considerable potential in this field. An experiment conducted by Google in 2016[35] is strong proof of this potential.

In this experiment, two neural networks named Alice and Bob were trained to communicate with each other without allowing a third neural network, named Eve, to eavesdrop on their communication.

As detailed in Figure 5, neural network Alice received a plaintext *P* as input that it had to encrypt into a cypher-text *C*, which was then received by neural network Bob and spoofed by neural network Eve. Alice and Bob shared a common decryption key *K*, which Bob used to decrypt the cypher-text to plaintext P_{Bob}, while Eve had to reconstruct the plaintext P_{Eve} without the help of the decryption key *K*. No specific encryption algorithm was specified; instead, the networks had to figure out how to securely communicate on their own. Results showed that Alice and Bob were able to learn how to perform forms of encryption and decryption autonomously. On top of this, they learned how to apply the encryption selectively only to the parts of data that were necessary to meet the confidentiality goals set.

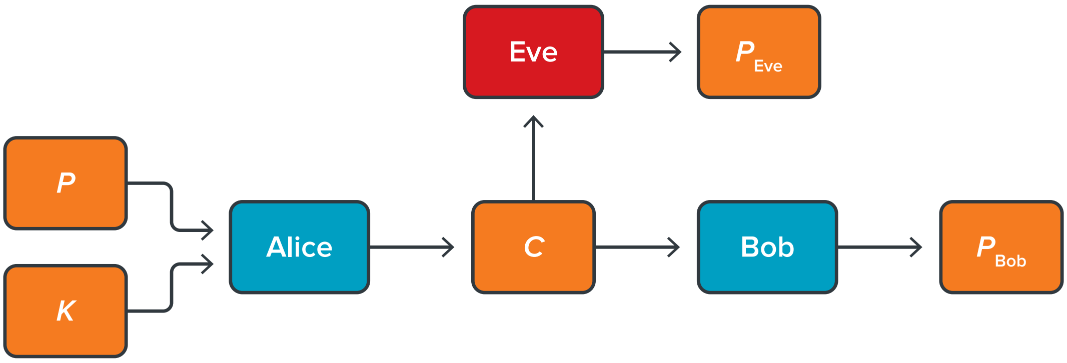


Figure 5. The communication model in Google's research paper

This experiment by Google demonstrates another use of AI that is difficult not only to detect, but also to defend from. From an external perspective, it is not obvious that AI is being used, and Eve (or any eavesdropper) would therefore find it difficult — if not impossible — to piece together the original message.

AI-aided decryption, therefore, has high potential for inclusion in the arsenal of attackers, to protect their communication and C2 infrastructure.

In 2017, researchers from Stanford and Google demonstrated how neural networks can use steganography. In their study, the researchers discovered that a neural network called CycleGAN learned how to cheat when conducting an image-to-image translation.[36] In the process of translating a satellite image into a map, researchers presumed that the aerial reconstruction would not resemble the original image too closely, as they expected that some features would probably be lost in CycleGAN's translation. However, researchers still found these features in the aerial reconstruction produced by the neural network. At that point, they realized that CycleGAN had learned to replicate aspects of the primary image into the map's noise patterns. In doing so, the neural network could hide some information for itself — the same information that it would need to transform the map into a reconstructed satellite image — so that it could produce a reconstruction that was uncannily similar to the original.

It is worth noting, as the researchers did in this case, that neural networks' capability to embed information is a vulnerability that could be useful for malicious actors. This possibility has also been observed by the Criminal Use of Information Hiding or CUING Initiative under Europol.[37] By implementing changes on the primary image, malicious actors could create their own image to suit their schemes. At the same time, developers who do not exercise caution when working with neural networks could be unaware that their personal information is being embedded by the same neural networks.[38]

Trends Found on Underground Forums

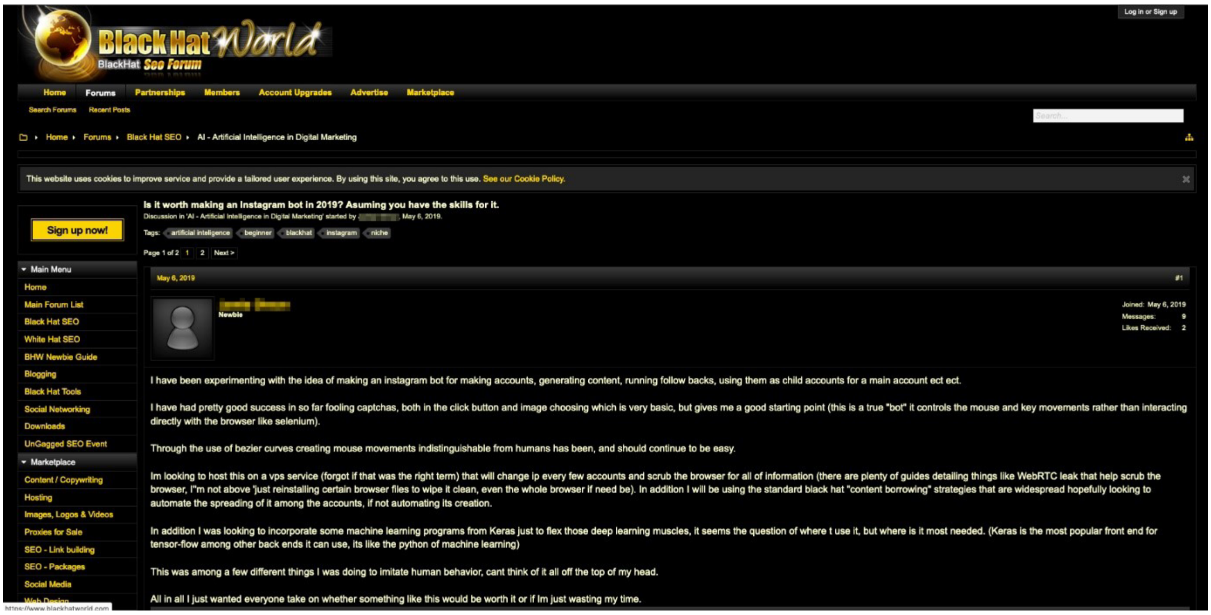
Trends on underground forums are revelatory of the possible landscape of the malicious abuse of AI in the near future.

To observe the relevant discussion trends, we analyzed the most popular topics involving AI, ML, and neural networks on underground forums. While the number of topics is not significant yet, there are some peculiar applications that deserve to be mentioned.

Human Impersonation on Social Networking Platforms

One application type involves intelligent systems that are able to impersonate human behavior to fool bot detection systems on social media platforms. One of the main purposes of these intelligent systems is to monetize, through fraud, songs in services such as Spotify. The systems have an army of bot users consume the specified songs, all the while maintaining a human-like usage pattern and thus generating traffic for a specific artist.

On a similar note, a discussion on the forum *blackhatworld[.]com* talks about the opportunity of developing an Instagram bot to create accounts of both fake followers and children, generate likes, run follow-backs, and so on as seen in the following screenshot. In this case, it is presumed that AI can be used to overcome CAPTCHAs, mimic device movements like selecting or dragging, and control browsers in a human-like behavior.



This website uses cookies to improve service and provide a tailored user experience. By using this site, you agree to this use. See our Cookie Policy.

Is it worth making an Instagram bot in 2019? Asuming you have the skills for it.

Discussion in 'AI - Artificial Intelligence in Digital Marketing' started by [redacted], May 6, 2019.

Tags: artificial inteligence, beginner, blackhat, instagram, niche

Page 1 of 2 1 2 Next >

May 6, 2019 #1

[redacted]
Newbie

Joined: May 6, 2019
Messages: 9
Likes Received: 2

I have been experimenting with the idea of making an instagram bot for making accounts, generating content, running follow backs, using them as child accounts for a main account ect ect.

I have had pretty good success in so far fooling captchas, both in the click button and image choosing which is very basic, but gives me a good starting point (this is a true "bot" it controls the mouse and key movements rather than interacting directly with the browser like selenium).

Through the use of bezier curves creating mouse movements indistinguishable from humans has been, and should continue to be easy.

Im looking to host this on a vps service (forgot if that was the right term) that will change ip every few accounts and scrub the browser for all of information (there are plenty of guides detailing things like WebRTC leak that help scrub the browser, I"m not above 'just reinstalling certain browser files to wipe it clean, even the whole browser if need be). In addition I will be using the standard black hat "content borrowing" strategies that are widespread hopefully looking to automate the spreading of it among the accounts, if not automating its creation.

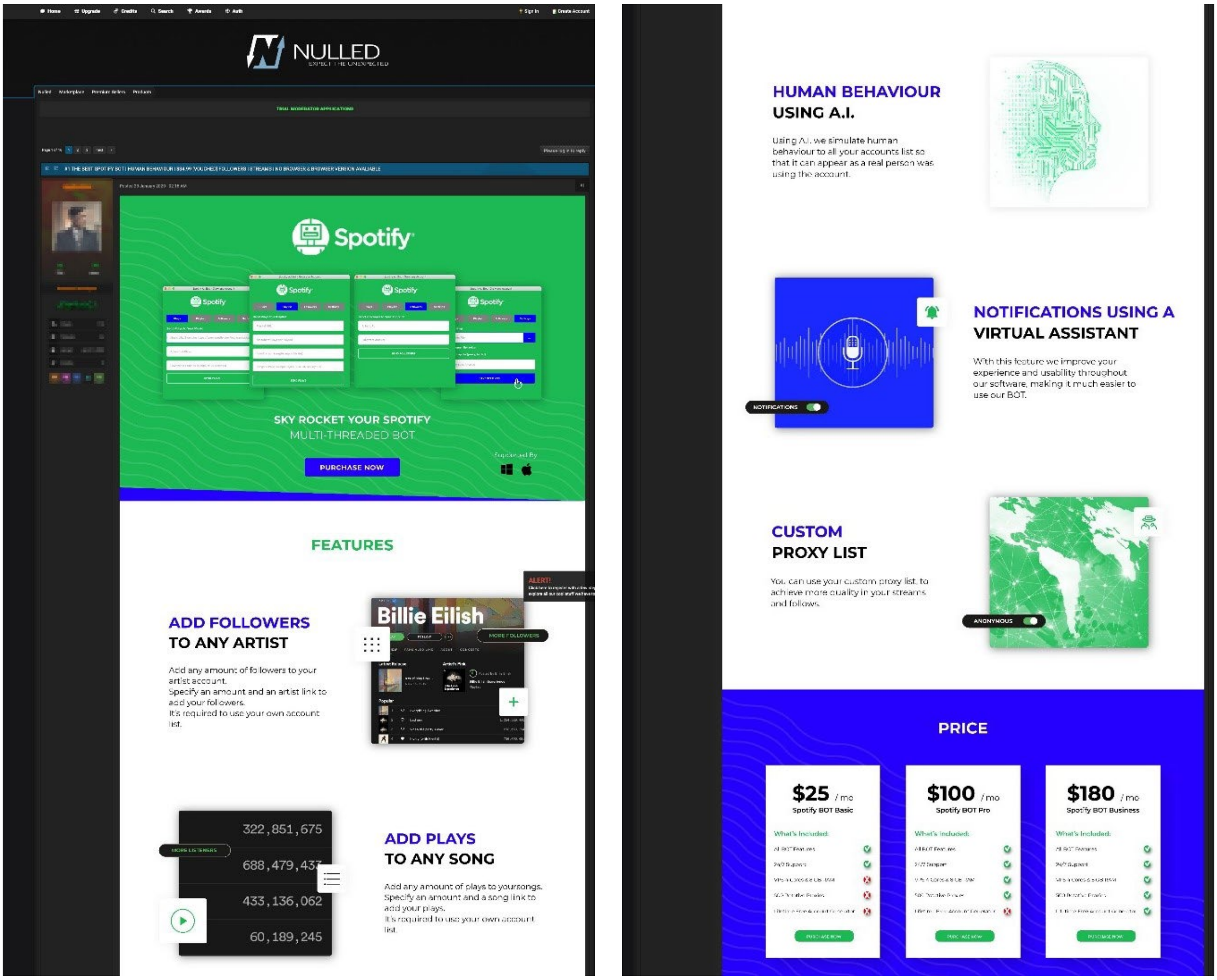
In addition I was looking to incorporate some machine learning programs from Keras just to flex those deep learning muscles, it seems the question of where t use it, but where is it most needed. (Keras is the most popular front end for tensor-flow among other back ends it can use, its like the python of machine learning)

This was among a few different things I was doing to imitate human behavior, cant think of it all off the top of my head.

All in all I just wanted everyone take on whether something like this would be worth it or if Im just wasting my time.

Figure 6. A discussion about Instagram bots on an underground forum

On the forum *nulled[.]to*, an AI-supported Spotify bot (*https://spotify[.]private-cdn[.]live*) is advertised as having the capability to impersonate several Spotify users in parallel by using multiple proxies to avoid being detected and banned by the system. It can also generate traffic and select targeted songs to increase their play count and improve monetization. To monetize this, the AI system is used to create playlists with different songs to better impersonate a user. The song selection follows some human-like musical tastes, rather than choosing a random selection of songs that could alert the company that the entity behind the selection is a bot.



#1 THE BEST SPOTIFY BOT | HUMAN BEHAVIOUR | $34.99 | VOUCHED | FOLLOWERS | STREAMS | NO BROWSER & BROWSER VERSION AVAILABLE

Posted 23 January 2020 - 02:18 AM

ADD FOLLOWERS
TO ANY ARTIST

Add any amount of followers to your artist account.
Specify an amount and an artist link to add your followers.
It's required to use your own account list.

ADD PLAYS
TO ANY SONG

Add any amount of plays to yoursongs.
Specify an amount and a song link to add your plays.
It's required to use your own account list.

HUMAN BEHAVIOUR
USING A.I.

Using A.I. we simulate human behaviour to all your accounts list so that it can appear as a real person was using the account.

NOTIFICATIONS USING A
VIRTUAL ASSISTANT

With this feature we improve your experience and usability throughout our software, making it much easier to use our BOT.

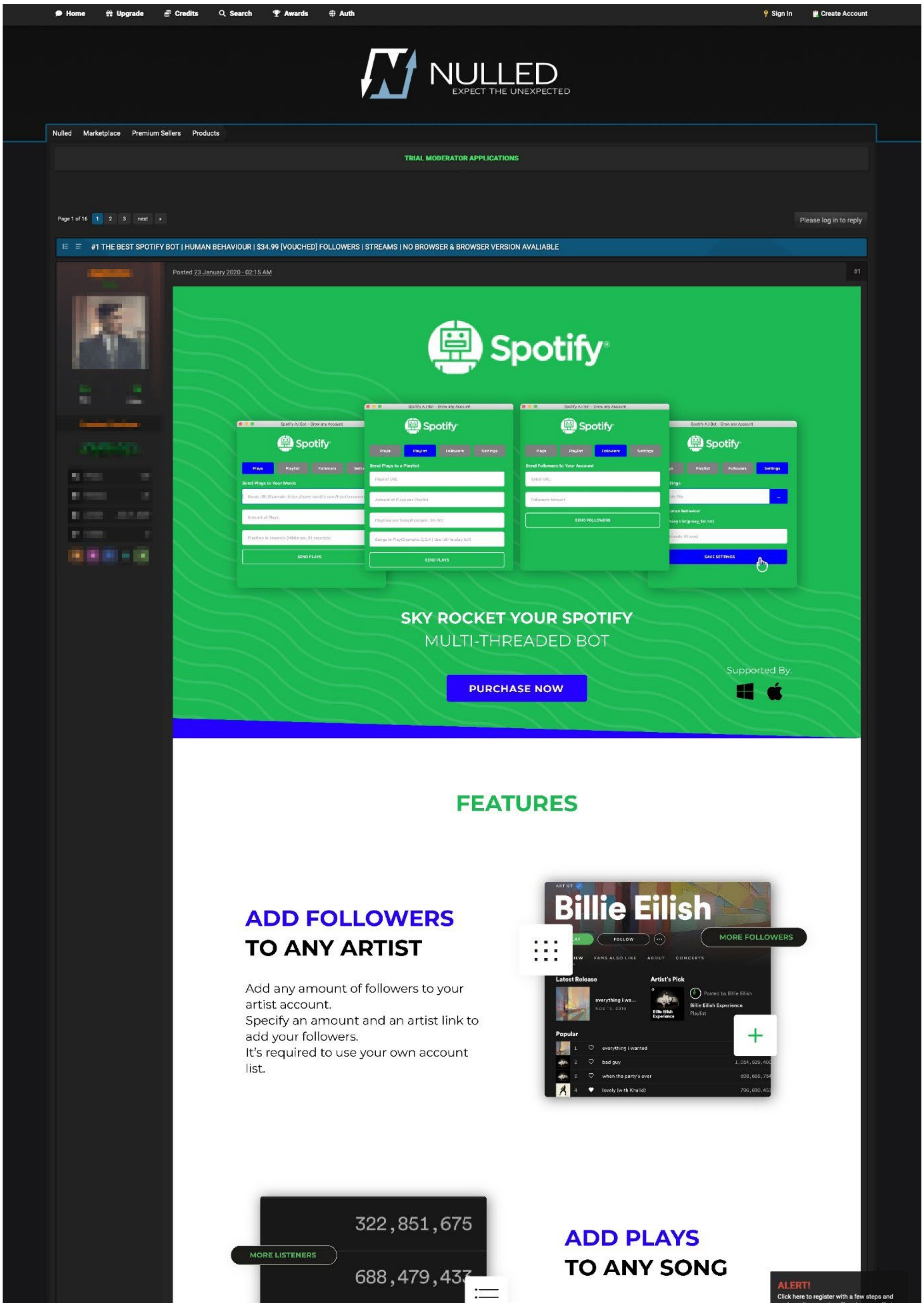
CUSTOM
PROXY LIST

You can use your custom proxy list to achieve more quality in your streams and follows.

PRICE

$25 / mo Spotify BOT Basic		$100 / mo Spotify BOT Pro		$180 / mo Spotify BOT Business	
What's Included:		What's Included:		What's Included:	
All BOT Features	✓	All BOT Features	✓	All BOT Features	✓
24/7 Support	✓	24/7 Support	✓	24/7 Support	✓
VPS 4 Cores & 8 GB RAM	✗	VPS 4 Cores & 8 GB RAM	✓	VPS 4 Cores & 8 GB RAM	✓
500 Rotative Proxies	✗	500 Rotative Proxies	✓	500 Rotative Proxies	✓
Lifetime Free Account Generator	✗	Lifetime Free Account Generator	✗	Lifetime Free Account Generator	✓
PURCHASE NOW		PURCHASE NOW		PURCHASE NOW	

Figure 7. The pricing information for a Spotify bot



#1 THE BEST SPOTIFY BOT | HUMAN BEHAVIOUR | $34.99 [VOUCHED] FOLLOWERS | STREAMS | NO BROWSER & BROWSER VERSION AVALIABLE

[redacted]

Posted 23 January 2020 - 02:15 AM

#1

SKY ROCKET YOUR SPOTIFY
MULTI-THREADED BOT

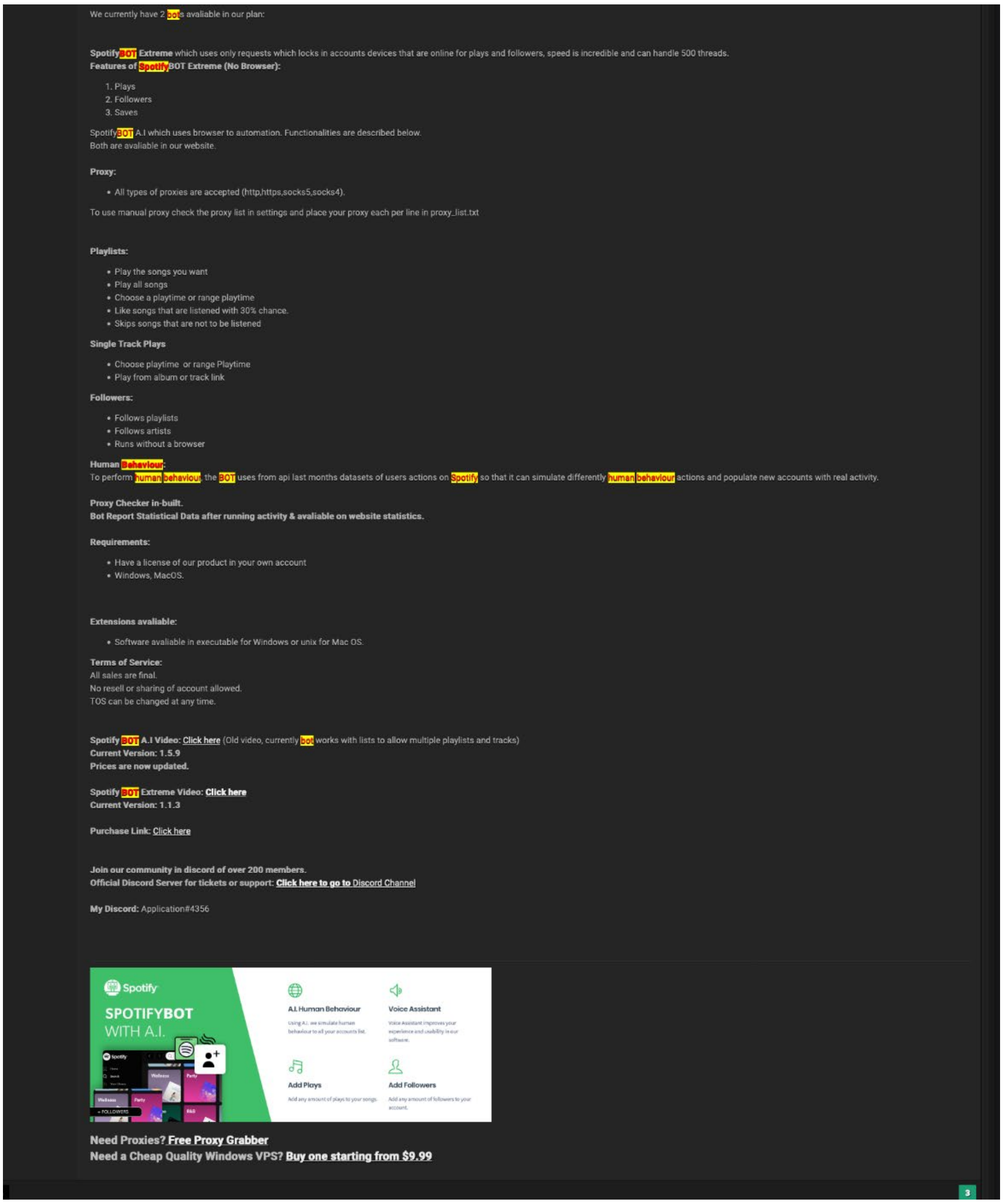
FEATURES

ADD FOLLOWERS
TO ANY ARTIST

Add any amount of followers to your artist account.
Specify an amount and an artist link to add your followers.
It's required to use your own account list.

ADD PLAYS
TO ANY SONG

Figure 8. An underground advertisement for a Spotify bot



We currently have 2 bots avaliable in our plan:

SpotifyBOT Extreme which uses only requests which locks in accounts devices that are online for plays and followers, speed is incredible and can handle 500 threads.
Features of SpotifyBOT Extreme (No Browser):

1. Plays
2. Followers
3. Saves

SpotifyBOT A.I which uses browser to automation. Functionalities are described below.
Both are avaliable in our website.

Proxy:

- All types of proxies are accepted (http,https,socks5,socks4).

To use manual proxy check the proxy list in settings and place your proxy each per line in proxy_list.txt

Playlists:

- Play the songs you want
- Play all songs
- Choose a playtime or range playtime
- Like songs that are listened with 30% chance.
- Skips songs that are not to be listened

Single Track Plays

- Choose playtime or range Playtime
- Play from album or track link

Followers:

- Follows playlists
- Follows artists
- Runs without a browser

Human Behaviour:
To perform human behaviour, the BOT uses from api last months datasets of users actions on Spotify so that it can simulate differently human behaviour actions and populate new accounts with real activity.

Proxy Checker in-built.
Bot Report Statistical Data after running activity & avaliable on website statistics.

Requirements:

- Have a license of our product in your own account
- Windows, MacOS.

Extensions avaliable:

- Software avaliable in executable for Windows or unix for Mac OS.

Terms of Service:
All sales are final.
No resell or sharing of account allowed.
TOS can be changed at any time.

Spotify BOT A.I Video: Click here (Old video, currently bot works with lists to allow multiple playlists and tracks)
Current Version: 1.5.9
Prices are now updated.

Spotify BOT Extreme Video: Click here
Current Version: 1.1.3

Purchase Link: Click here

Join our community in discord of over 200 members.
Official Discord Server for tickets or support: Click here to go to Discord Channel

My Discord: Application#4356

Spotify
SPOTIFYBOT
WITH A.I.
A.I. Human Behaviour
Using A.I. we simulate human behaviour to all your accounts list.
Voice Assistant
Voice Assistant improves your experience and usability in our software.
Add Plays
Add any amount of plays to your songs.
Add Followers
Add any amount of followers to your account.
+ FOLLOWERS
Wellness
Party
R&B

Need Proxies? Free Proxy Grabber
Need a Cheap Quality Windows VPS? Buy one starting from $9.99

3

Figure 9. A detailed advertisement post for a Spotify bot on a forum

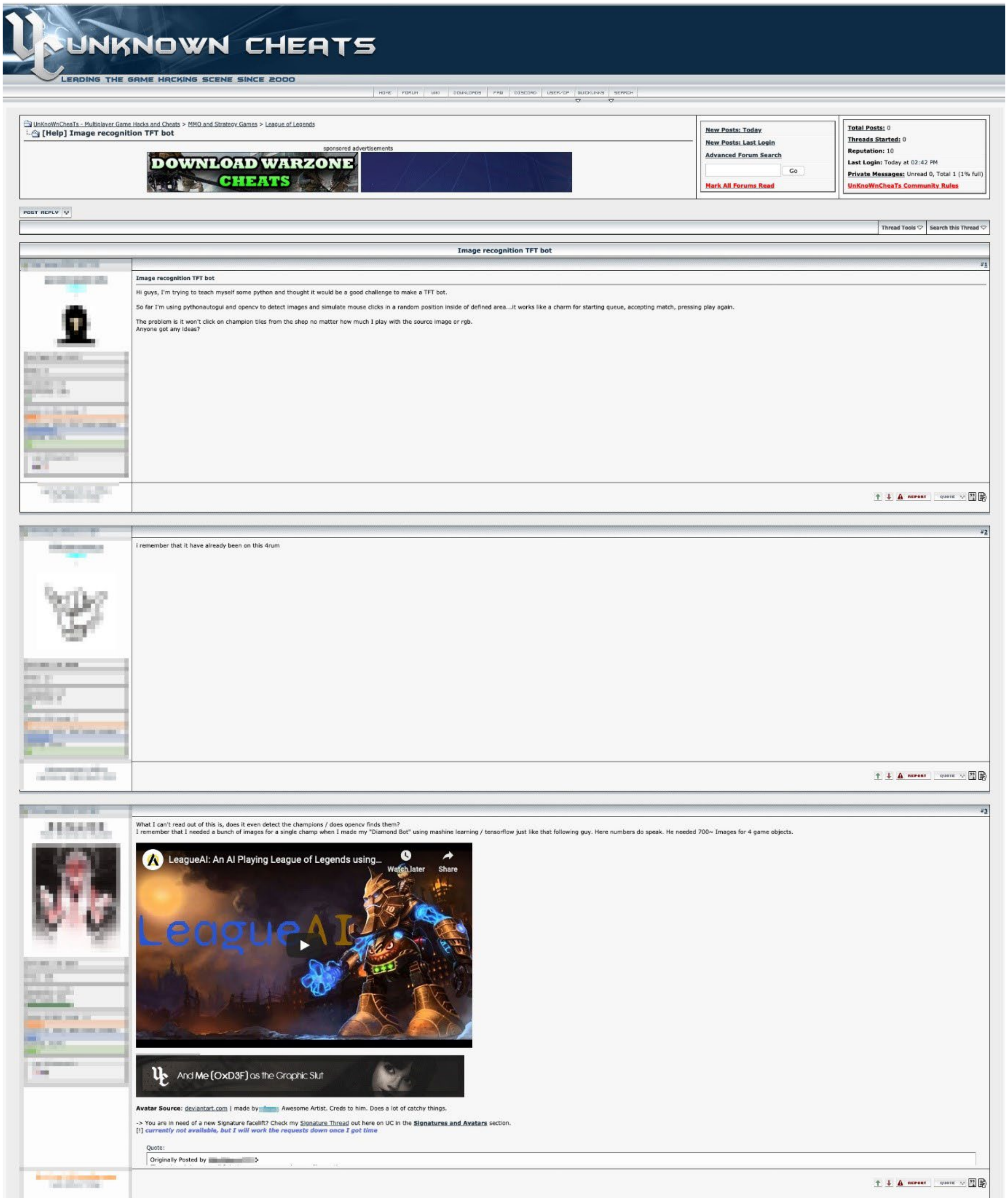
Online Game Cheats

AI is also used in the game cheats domain, especially for games where monetization is substantial. As detailed in the paper by Trend Micro titled, “Cheats, Hacks, and Cyberattacks: Threats to the Esports Industry in 2019 and Beyond,”[39] the growth of the eSport industry means that it is also becoming an increasingly interesting target for possible attacks.

Furthermore, these attacks go beyond targeting mere hobby games. They are, in fact, geared toward professional and lucrative eSports. As online gaming competitions involve substantial sums of prize money, they are a prime target for cheating attacks that use the latest technologies available. In fact, criminals are already using online gaming for money-laundering purposes.[40, 41, 42]

On several different forums, posts that inquire about developing "aimbots" (software that exploits computer vision algorithms to automatically aim at targets) are common. These bots use deep learning algorithms to be capable not just of playing games such as League of Legends,[43] but also developing cheats for those kinds of games.

As evidenced by the following screenshots, the investigation performed on underground forums reveals the potential of using AI algorithms to profit from eSports.


UNKNOWN CHEATS
LEADING THE GAME HACKING SCENE SINCE 2000
HOME | FORUM | WIKI | DOWNLOADS | FAQ | DISCORD | USER/CP | QUICKLINKS | SEARCH

UnKnoWnCheaTs - Multiplayer Game Hacks and Cheats > MMO and Strategy Games > League of Legends
[Help] Image recognition TFT bot

Total Posts: 0
Threads Started: 0
Reputation: 10
Last Login: Today at 02:42 PM
Private Messages: Unread 0, Total 1 (1% full)
UnKnoWnCheaTs Community Rules

Image recognition TFT bot

#1

Image recognition TFT bot

Hi guys, I'm trying to teach myself some python and thought it would be a good challenge to make a TFT bot.

So far I'm using pythonautogui and opencv to detect images and simulate mouse clicks in a random position inside of defined area...it works like a charm for starting queue, accepting match, pressing play again.

The problem is it won't click on champion tiles from the shop no matter how much I play with the source image or rgb.
Anyone got any ideas?

#2

i remember that it have already been on this 4rum

#3

What I can't read out of this is, does it even detect the champions / does opencv finds them?
I remember that I needed a bunch of images for a single champ when I made my "Diamond Bot" using mashine learning / tensorflow just like that following guy. Here numbers do speak. He needed 700~ Images for 4 game objects.

LeagueAI: An AI Playing League of Legends using...
Watch later
Share
LeagueAI
And Me (OxD3F) as the Graphic Slut

Avatar Source: deviantart.com | made by [redacted] Awesome Artist. Creds to him. Does a lot of catchy things.

-> You are in need of a new Signature facelift? Check my Signature Thread out here on UC in the **Signatures and Avatars** section.
[!] ***currently not available, but I will work the requests down once I got time***

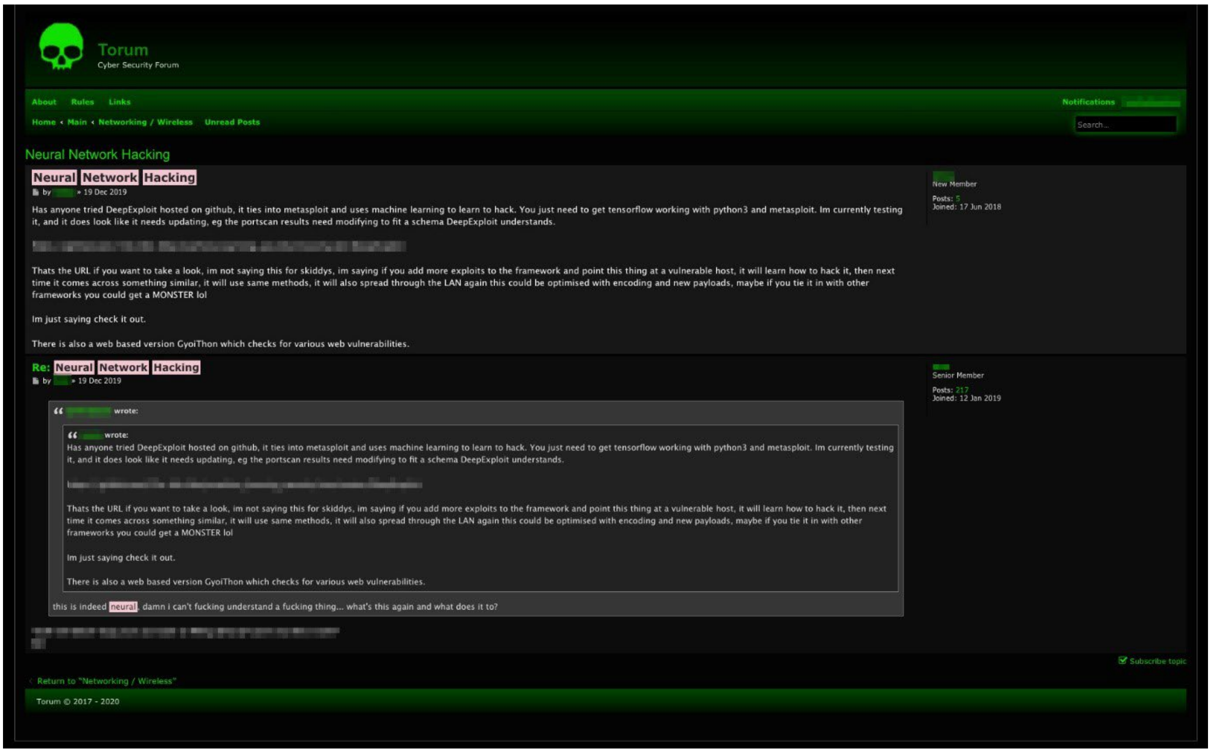
Quote:
Originally Posted by [redacted]

Figure 10. A post on a forum about AI cheating in eSports

AI-Supported Hacking

Several discussions on underground forums provide additional insights on the practical use of frameworks such as DeepHack and DeepExploit.

In Torum, a darknet forum, one user claims to be looking for information on how to successfully use DeepExploit and how to interface it with Metasploit.



Neural Network Hacking

Neural Network Hacking
by [redacted] » 19 Dec 2019

Has anyone tried DeepExploit hosted on github, it ties into metasploit and uses machine learning to learn to hack. You just need to get tensorflow working with python3 and metasploit. Im currently testing it, and it does look like it needs updating, eg the portscan results need modifying to fit a schema DeepExploit understands.

[redacted]

Thats the URL if you want to take a look, im not saying this for skiddys, im saying if you add more exploits to the framework and point this thing at a vulnerable host, it will learn how to hack it, then next time it comes across something similar, it will use same methods, it will also spread through the LAN again this could be optimised with encoding and new payloads, maybe if you tie it in with other frameworks you could get a MONSTER lol

Im just saying check it out.

There is also a web based version GyoiThon which checks for various web vulnerabilities.

[redacted]
New Member
Posts: 5
Joined: 17 Jun 2018

Re: Neural Network Hacking
by [redacted] » 19 Dec 2019

> “ [redacted] wrote:
>
> > “ [redacted] wrote:
> > Has anyone tried DeepExploit hosted on github, it ties into metasploit and uses machine learning to learn to hack. You just need to get tensorflow working with python3 and metasploit. Im currently testing it, and it does look like it needs updating, eg the portscan results need modifying to fit a schema DeepExploit understands.
> >
> > [redacted]
> >
> > Thats the URL if you want to take a look, im not saying this for skiddys, im saying if you add more exploits to the framework and point this thing at a vulnerable host, it will learn how to hack it, then next time it comes across something similar, it will use same methods, it will also spread through the LAN again this could be optimised with encoding and new payloads, maybe if you tie it in with other frameworks you could get a MONSTER lol
> >
> > Im just saying check it out.
> >
> > There is also a web based version GyoiThon which checks for various web vulnerabilities.
>
> this is indeed neural, damn i can't fucking understand a fucking thing... what's this again and what does it to?

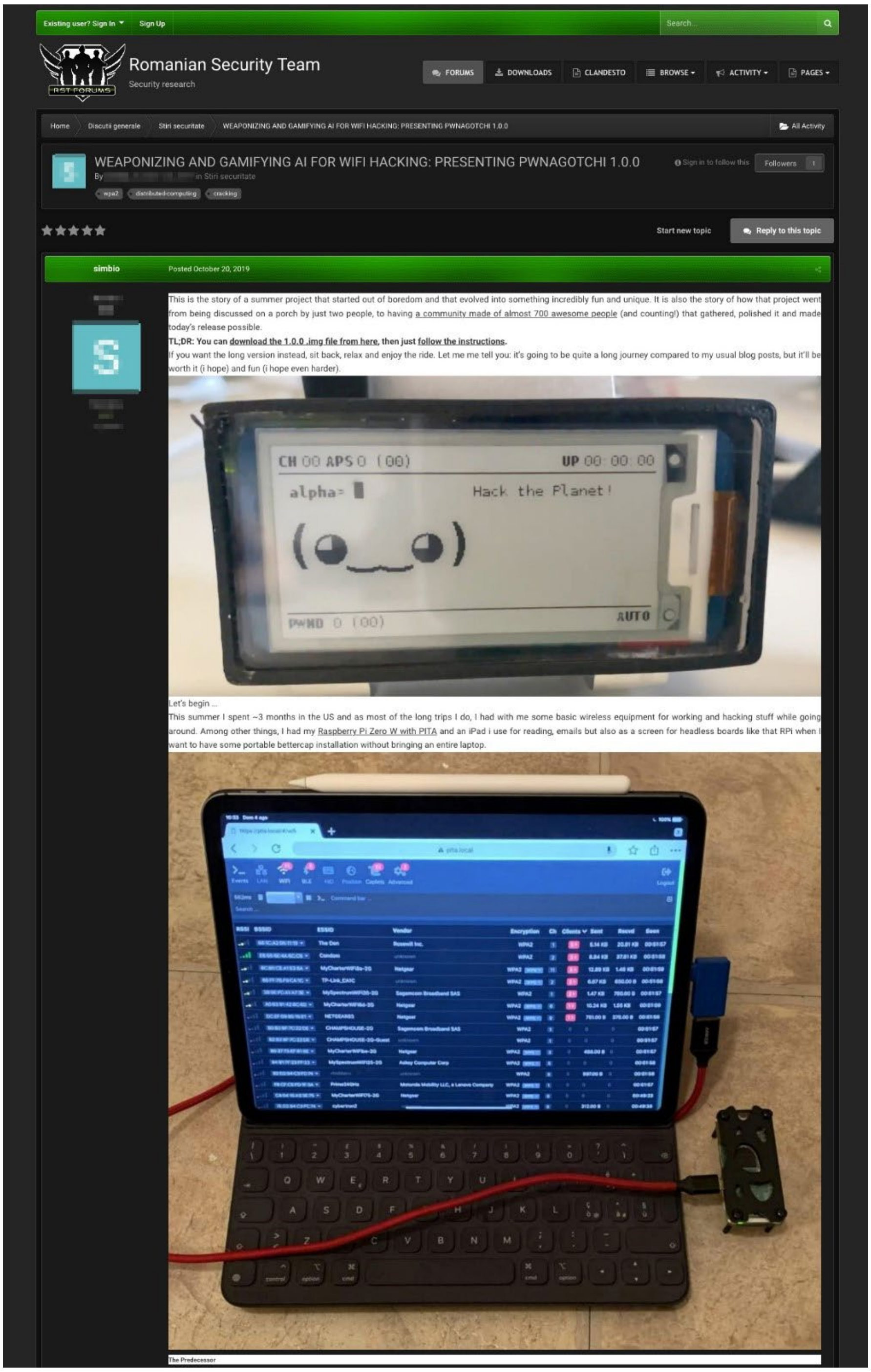
[redacted]

[redacted]
Senior Member
Posts: 217
Joined: 12 Jan 2019

Figure 11. A user on a forum inquiring about the use of DeepExploit

As alluded to previously, AI-enabled password-guessing tools are also already in use by hackers and other criminals. For instance, on the forum *cracked[.]to*, a post listing a collection of open-source hacking tools includes an AI-based software capable of analyzing a large dataset of passwords recovered from public leaks. The tool is able to train a GAN that understands how people tend to modify and update passwords (for example, modifying “hello123” to “h@llo123,” and then to “h@llo!23”), which provides considerable help in optimizing password-guessing tools.

A discussion thread on *rstforums[.]com*, meanwhile, discusses the tool called “PWnagotchi 1.0.0.” The last version of this tool, which had been originally developed to perform Wi-Fi hacking through de-authentication attacks, it is powered by a neural network model to optimize its performance by using a gamification strategy. Through this strategy, the system is rewarded for every successful de-authentication attempt. As a result, the incentive-based system causes the tool to autonomously improve its performances.



WEAPONIZING AND GAMIFYING AI FOR WIFI HACKING: PRESENTING PWNAGOTCHI 1.0.0

By [redacted] in Stiri securitate

simbio Posted October 20, 2019

This is the story of a summer project that started out of boredom and that evolved into something incredibly fun and unique. It is also the story of how that project went from being discussed on a porch by just two people, to having a community made of almost 700 awesome people (and counting!) that gathered, polished it and made today's release possible.

TL;DR: You can download the 1.0.0 .img file from here, then just follow the instructions.

If you want the long version instead, sit back, relax and enjoy the ride. Let me me tell you: it's going to be quite a long journey compared to my usual blog posts, but it'll be worth it (i hope) and fun (i hope even harder).

Let's begin ...

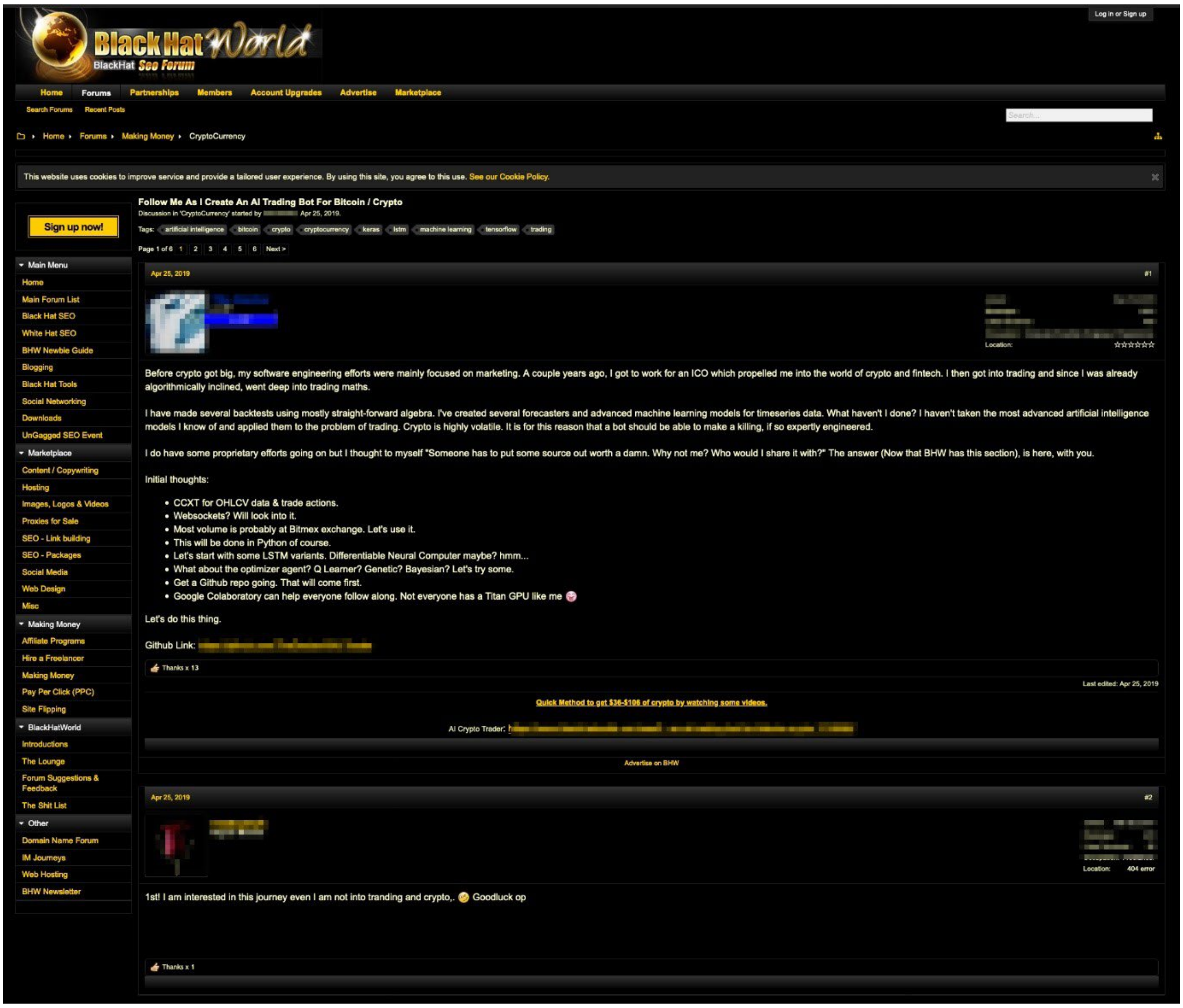
This summer I spent ~3 months in the US and as most of the long trips I do, I had with me some basic wireless equipment for working and hacking stuff while going around. Among other things, I had my Raspberry Pi Zero W with PITA and an iPad i use for reading, emails but also as a screen for headless boards like that RPi when I want to have some portable bettercap installation without bringing an entire laptop.

The Predecessor

Figure 12. A PWnagotchi 1.0.0 description post

AI-Supported Cryptocurrency Trading

AI-based tools are also available at a time that is ripe for developing new monetization schemes for malicious use. To illustrate, although AI-based tools for financial trading have been known for a while, the forum at *blackhatworld[.]com* contains a reference to AI enabled bots that are specially dedicated to cryptocurrency trading. The bots referenced in this forum are able to learn successful trading strategies from historic data and apply them to make better predictions on more profitable trades. Notably, although the post dates back to April 2019 and the corresponding GitHub repository has not seen much activity ever since, it is still indicative that cryptocurrency is a lucrative opportunity for financial gain by criminals.



This website uses cookies to improve service and provide a tailored user experience. By using this site, you agree to this use. See our Cookie Policy.

Follow Me As I Create An AI Trading Bot For Bitcoin / Crypto

Discussion in 'CryptoCurrency' started by [redacted] Apr 25, 2019.

Tags: artificial intelligence, bitcoin, crypto, cryptocurrency, keras, lstm, machine learning, tensorflow, trading

Page 1 of 6 1 2 3 4 5 6 Next >

Apr 25, 2019 #1

[redacted]

Location: ☆☆☆☆☆☆

Before crypto got big, my software engineering efforts were mainly focused on marketing. A couple years ago, I got to work for an ICO which propelled me into the world of crypto and fintech. I then got into trading and since I was already algorithmically inclined, went deep into trading maths.

I have made several backtests using mostly straight-forward algebra. I've created several forecasters and advanced machine learning models for timeseries data. What haven't I done? I haven't taken the most advanced artificial intelligence models I know of and applied them to the problem of trading. Crypto is highly volatile. It is for this reason that a bot should be able to make a killing, if so expertly engineered.

I do have some proprietary efforts going on but I thought to myself "Someone has to put some source out worth a damn. Why not me? Who would I share it with?" The answer (Now that BHW has this section), is here, with you.

Initial thoughts:

- CCXT for OHLCV data & trade actions.
- Websockets? Will look into it.
- Most volume is probably at Bitmex exchange. Let's use it.
- This will be done in Python of course.
- Let's start with some LSTM variants. Differentiable Neural Computer maybe? hmm...
- What about the optimizer agent? Q Learner? Genetic? Bayesian? Let's try some.
- Get a Github repo going. That will come first.
- Google Colaboratory can help everyone follow along. Not everyone has a Titan GPU like me 😛

Let's do this thing.

Github Link: [redacted]

Thanks x 13

Last edited: Apr 25, 2019

Quick Method to get $36-$106 of crypto by watching some videos.

AI Crypto Trader: [redacted]

Advertise on BHW

Apr 25, 2019 #2

[redacted]

Location: 404 error

1st! I am interested in this journey even I am not into tranding and crypto,. 😊 Goodluck op

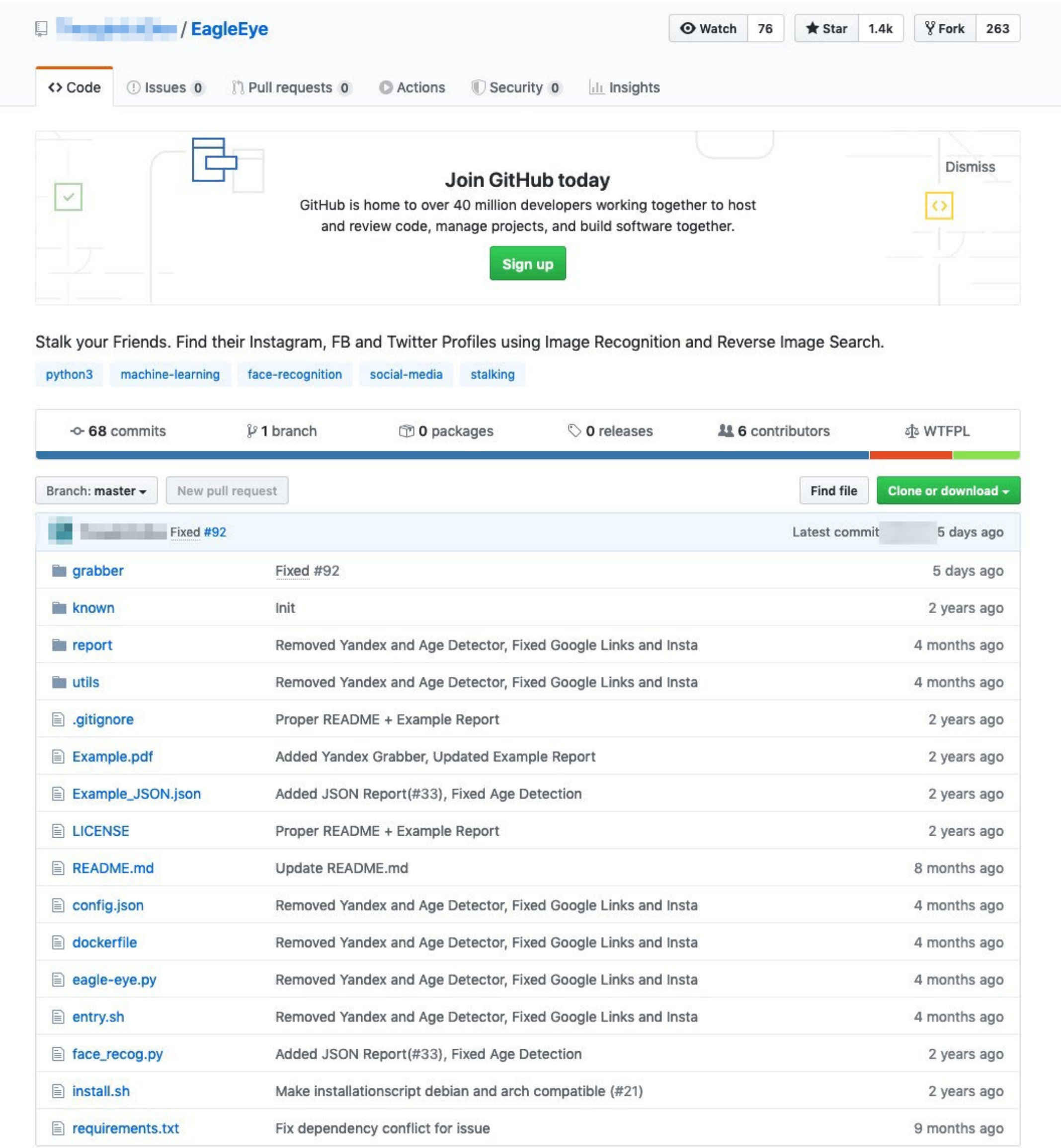
Thanks x 1

Figure 13. A discussion on AI bots for cryptocurrency trading

Social Engineering

As AI technology develops, so do the different schemes for social engineering that might profit from such technology. This is crucial to note, as Europol has noted in 2020 that social engineering still serves as a top threat that is utilized for facilitating other forms of cybercrime.[44] In evidence of these schemes, several interesting discussions related to AI-enabled tools to improve social engineering tasks have been found on different underground forums.

On the forum French Freedom Zone, the reconnaissance tool named “Eagle Eyes”[45] has been claimed as capable of finding all social media accounts associated with a specific profile. Accordingly, the tool can even match profiles with different names by comparing a user’s profile photos through facial recognition algorithms.



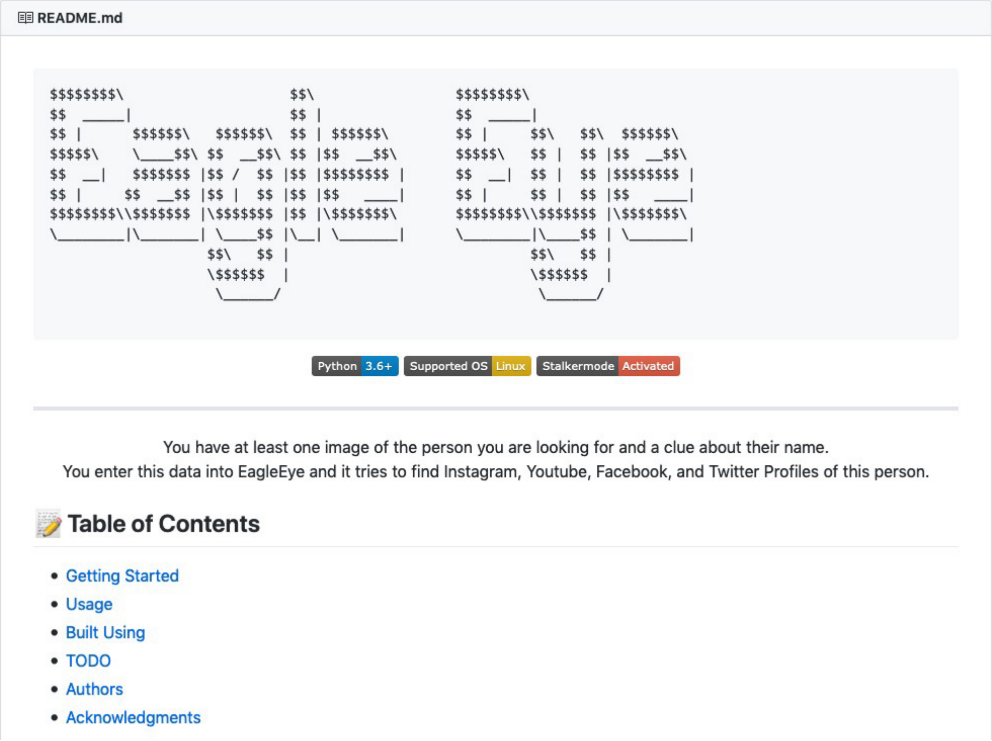
Join GitHub today

GitHub is home to over 40 million developers working together to host and review code, manage projects, and build software together.

Stalk your Friends. Find their Instagram, FB and Twitter Profiles using Image Recognition and Reverse Image Search.

[redacted] Fixed #92 Latest commit [redacted] 5 days ago

grabber	Fixed #92	5 days ago
known	Init	2 years ago
report	Removed Yandex and Age Detector, Fixed Google Links and Insta	4 months ago
utils	Removed Yandex and Age Detector, Fixed Google Links and Insta	4 months ago
.gitignore	Proper README + Example Report	2 years ago
Example.pdf	Added Yandex Grabber, Updated Example Report	2 years ago
Example_JSON.json	Added JSON Report(#33), Fixed Age Detection	2 years ago
LICENSE	Proper README + Example Report	2 years ago
README.md	Update README.md	8 months ago
config.json	Removed Yandex and Age Detector, Fixed Google Links and Insta	4 months ago
dockerfile	Removed Yandex and Age Detector, Fixed Google Links and Insta	4 months ago
eagle-eye.py	Removed Yandex and Age Detector, Fixed Google Links and Insta	4 months ago
entry.sh	Removed Yandex and Age Detector, Fixed Google Links and Insta	4 months ago
face_recog.py	Added JSON Report(#33), Fixed Age Detection	2 years ago
install.sh	Make installationscript debian and arch compatible (#21)	2 years ago
requirements.txt	Fix dependency conflict for issue	9 months ago



README.md

```
$$$$$$$$\                   $$\                 $$$$$$$$\
$$  _____|                  $$ |                $$  _____|
$$ |      $$$$$$\   $$$$$$\  $$ | $$$$$$\        $$ |      $$\   $$\  $$$$$$\
$$$$$\    \____$$\ $$  __$$\ $$ |$$  __$$\       $$$$$\    $$ |  $$ |$$  __$$\
$$  __|   $$$$$$$ |$$ /  $$ |$$ |$$$$$$$$ |      $$  __|   $$ |  $$ |$$$$$$$$ |
$$ |     $$  __$$ |$$ |  $$ |$$ |$$   ____|      $$ |      $$ |  $$ |$$   ____|
$$$$$$$$\\$$$$$$$ |\$$$$$$$ |$$ |\$$$$$$$\       $$$$$$$$\\$$$$$$$ |\$$$$$$$\
\________|\_______| \____$$ |\__| \_______|      \________|\____$$ | \_______|
                    $$\   $$ |                            $$\   $$ |
                    \$$$$$$  |                            \$$$$$$  |
                     \______/                              \______/
```

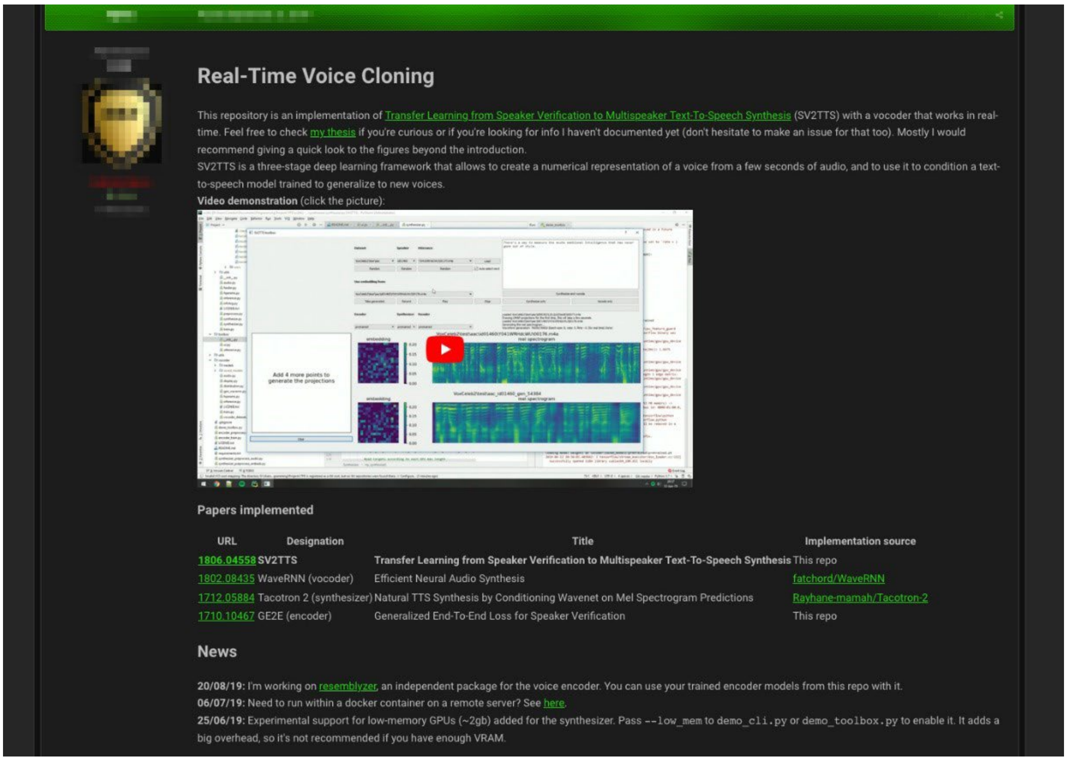
You have at least one image of the person you are looking for and a clue about their name.
You enter this data into EagleEye and it tries to find Instagram, Youtube, Facebook, and Twitter Profiles of this person.

📝 Table of Contents

- Getting Started
- Usage
- Built Using
- TODO
- Authors
- Acknowledgments

Figure 14. The GitHub repository for the Eagle Eye tool

Additionally, a post on *rstforums[.]com* advertises a tool capable of performing real-time voice cloning.[46] With just a five-second voice recording of a target, a malicious actor can already clone their voice. Detailed information about voice cloning can be found in the case study at the end of this paper.



Real-Time Voice Cloning

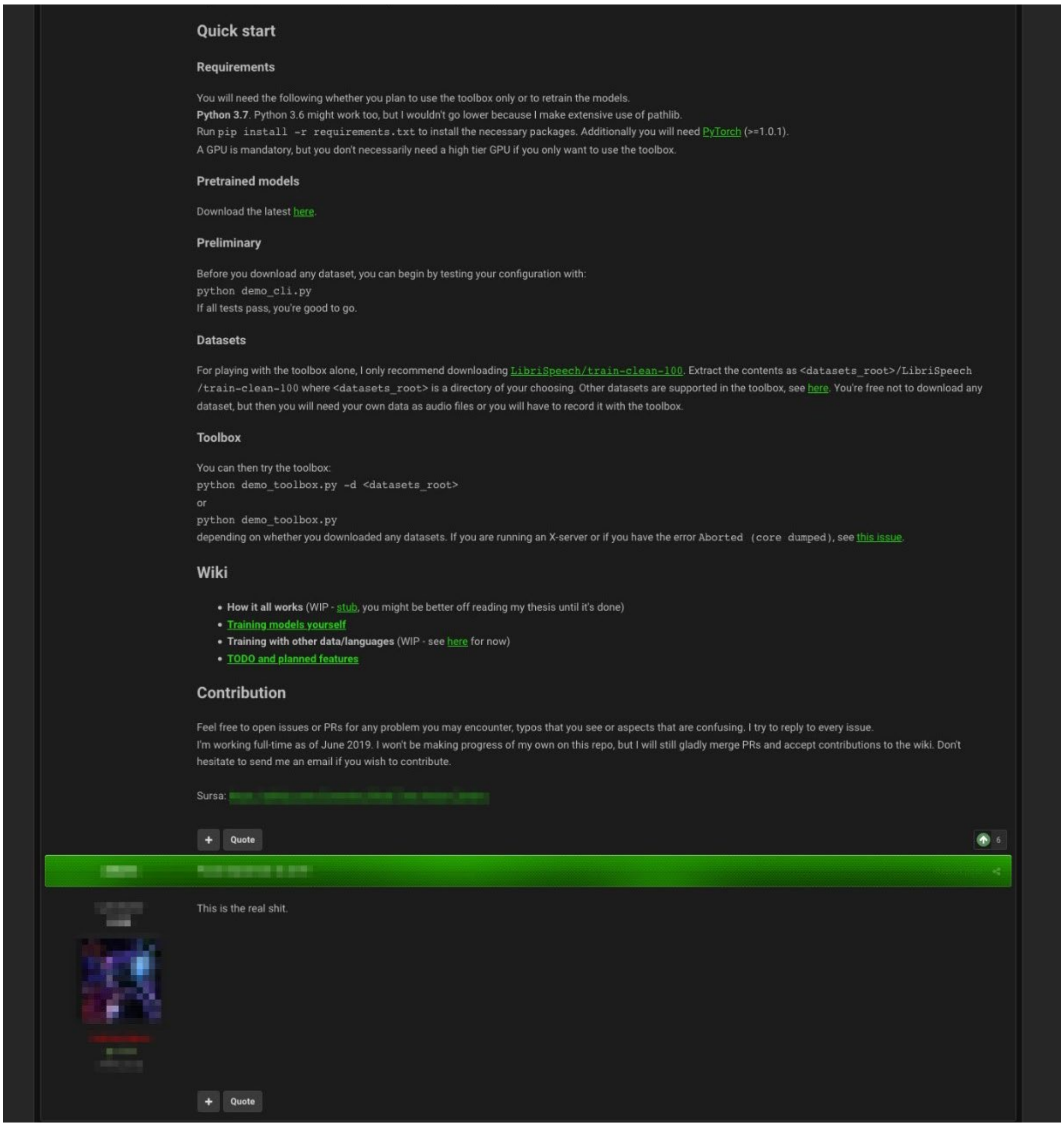
This repository is an implementation of Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis (SV2TTS) with a vocoder that works in real-time. Feel free to check my thesis if you're curious or if you're looking for info I haven't documented yet (don't hesitate to make an issue for that too). Mostly I would recommend giving a quick look to the figures beyond the introduction.
SV2TTS is a three-stage deep learning framework that allows to create a numerical representation of a voice from a few seconds of audio, and to use it to condition a text-to-speech model trained to generalize to new voices.
Video demonstration (click the picture):

Papers implemented

URL	Designation	Title	Implementation source
1806.04558	**SV2TTS**	**Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis**	This repo
1802.08435	WaveRNN (vocoder)	Efficient Neural Audio Synthesis	fatchord/WaveRNN
1712.05884	Tacotron 2 (synthesizer)	Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions	Rayhane-mamah/Tacotron-2
1710.10467	GE2E (encoder)	Generalized End-To-End Loss for Speaker Verification	This repo

News

20/08/19: I'm working on resemblyzer, an independent package for the voice encoder. You can use your trained encoder models from this repo with it.
06/07/19: Need to run within a docker container on a remote server? See here.
25/06/19: Experimental support for low-memory GPUs (~2gb) added for the synthesizer. Pass `--low_mem` to `demo_cli.py` or `demo_toolbox.py` to enable it. It adds a big overhead, so it's not recommended if you have enough VRAM.



Quick start

Requirements

You will need the following whether you plan to use the toolbox only or to retrain the models.
Python 3.7. Python 3.6 might work too, but I wouldn't go lower because I make extensive use of pathlib.
Run `pip install -r requirements.txt` to install the necessary packages. Additionally you will need PyTorch (>=1.0.1).
A GPU is mandatory, but you don't necessarily need a high tier GPU if you only want to use the toolbox.

Pretrained models

Download the latest here.

Preliminary

Before you download any dataset, you can begin by testing your configuration with:

```
python demo_cli.py
```

If all tests pass, you're good to go.

Datasets

For playing with the toolbox alone, I only recommend downloading `LibriSpeech/train-clean-100`. Extract the contents as `<datasets_root>/LibriSpeech/train-clean-100` where `<datasets_root>` is a directory of your choosing. Other datasets are supported in the toolbox, see here. You're free not to download any dataset, but then you will need your own data as audio files or you will have to record it with the toolbox.

Toolbox

You can then try the toolbox:

```
python demo_toolbox.py -d <datasets_root>
```

or

```
python demo_toolbox.py
```

depending on whether you downloaded any datasets. If you are running an X-server or if you have the error `Aborted (core dumped)`, see this issue.

Wiki

- **How it all works** (WIP - stub, you might be better off reading my thesis until it's done)
- **Training models yourself**
- **Training with other data/languages** (WIP - see here for now)
- **TODO and planned features**

Contribution

Feel free to open issues or PRs for any problem you may encounter, typos that you see or aspects that are confusing. I try to reply to every issue.
I'm working full-time as of June 2019. I won't be making progress of my own on this repo, but I will still gladly merge PRs and accept contributions to the wiki. Don't hesitate to send me an email if you wish to contribute.

Sursa: [redacted]

This is the real shit.

Figure 15. An advertisement for a voice-cloning tool

Future Scenarios of Malicious Uses and Abuses of AI

The development of new technologies, including AI systems, contributes to increased automation that can result in the delegation of control in some levels of society.

In addition to progressively integrating AI techniques to enhance the scope and scale of their cyberattacks,[47] cybercriminals can exploit AI both as an attack vector and an attack surface. Thanks to the previously described service-based criminal business model, the development and distribution of AI-enabled toolkits in underground markets might provide less technologically savvy criminals with enhanced capabilities.

This second section offers a comprehensive overview based on present-day experience and explores plausible scenarios in which criminals might abuse AI technologies to facilitate their activities. Alongside this abuse, we expect cybercriminals to continuously employ AI techniques to conceal their illicit activities online and avoid detection.

Other authors have proposed an interesting angle from which to analyze this topic. From this angle, if one considers the different phases in the anatomy of a cyberattack, one could theorize how cybercriminals might use AI techniques to improve any of them. Naturally, there is room to optimize or enhance all the classical attack steps: reconnaissance, penetration, lateral movement, command and control (C&C), and exfiltration. [A complete study of possible AI-enabled attacks can be found in a research by the Swedish Defence Research Agency (FOI).][48]

Social Engineering at Scale

Social engineering scams are not new, but they remain successful because they prey on people's inherent trust in, fear of, and respect for authority. Scammers, on the other hand, try to convince their victims that their scheme is legitimate, and that the targeted victims can safely wire money under whatever pretense the scam calls for. Still, these scams are laborious because the perpetrator needs to spend significant amounts of time trying to convince the victim that the story is true.

An innovative scammer, however, can introduce AI systems to automate and speed up the detection rate at which the victims fall in or out of the scam, which would allow them to focus only on those potential victims who are easy to deceive. Whatever false pretense a scammer chooses to persuade the target to participate in, an ML algorithm would be able to anticipate a target's most common replies to the chosen pretense. The algorithm can then act according to these anticipated replies in order to continue the whole act and lend credibility to the story. More alarmingly, this relatively easy process only requires enough written material to identify the most likely replies for each incoming message.

Unfortunately, criminals already have enough experience and sample texts to build their operations on. Even if scammers were to automatically categorize texts only within the first three email exchanges — regardless of whether the potential victim is likely to pay or not — scammers would be able to save themselves considerable effort and time, as these three emails would suffice for the ML algorithm to continue the operation.

In fact, this use of AI systems would be the equivalent of automating the first-level tech support in a corporation. Simply focusing on those three email exchanges would give the highest returns on the scammers' investment as they would then be able to automate the rest. Thus, by automating the one real considerable expense in such a criminal endeavor — in this case, people — a scammer could reduce most of the cost and maintain their income, therefore increasing the success rate of their scamming operation.

Another plausible scenario would involve the online verification systems that are used in the banking sector to validate a creator of a new account. These verification systems are based on Know Your Customer principles (also known as KYC). Anecdotally, there have been attacks where video streams were manipulated in real time to insert fake IDs to fool the verification system. Combined with the possibility of synthesizing voices and faces, criminals can abuse these systems at scale, for example, to create new banking accounts or authorize payments. This is important to note, as video-based authentication is currently one of the recommended identification mechanisms for the banking industry in several countries.

When trying to defend verification systems against these kinds of attack, there are two approaches: The first is to try detecting whether the message has been fabricated (in the case of KYC, this would involve verifying the video). The second approach is to ascertain if the source is trustworthy, in order to learn if the message is genuine or not.

Depending on the nature of the message, detecting the attack might be difficult. For instance, in the case of the aforementioned email scam, a simple spam filter would suffice. On the other hand, a manipulated photo or video might be impossible to detect as fake.

Nonetheless, combining both approaches against these kinds of attack would likely be the most successful strategy. It would be essential not only to detect whether the message has been automatically generated, but also to determine whether the origin of the message is untrustworthy. Imagine, for example, how the same IP address sending KYC authentication files with different images would likely raise a red flag in the receiving organization.

Content Generation

Content generation refers to the ability of an algorithm to generate arbitrary content that would look human-made. This algorithm would also make it possible to set constraints in content generation and have a system imitating or cloning certain aspects of an existing piece of content. There are claims that AI-based content generators are becoming so powerful that their release to the public is considered a risk.[49]

Such claims refer to the Generative Pretrained Transformer 3 (GPT-3), a text synthesis technology released by OpenAI[50] in June 2020 that uses deep learning to produce human-like text and is able to adapt and fine-tune its behavior with the addition of a few domain-specific examples. Bearing a capacity of over 175 billion machine learning parameters (10 times more than its closest competitor, Microsoft Turing NLG), this technology is capable of synthesizing not only English text, but also code in several programming languages and even guitar tablatures, allowing for applications such as:

- Generating a full, human-sounding text of a simple title
- Turning the textual description of an application into working code
- Changing the writing style of a text while maintaining the content
- Passing the Turing test for a human-sounding chatbot

Criminals could thus employ ML to generate and distribute new content such as (semi) automatically created and high-quality (spear-)phishing and spam emails in less popular languages.[51] In effect, this would further automate and amplify the scope and scale of malware distribution worldwide.

Moreover, such content-generation techniques could significantly ameliorate disinformation campaigns by automatically combining legitimate with false information while also learning which kinds of content work best and which are the most widely shared.[52]

The ability to generate working code from a mere textual description lowers the knowledge barrier required to become a programmer and could foster a new generation of “script kiddies” — that is, people with low technical knowledge but malicious intentions who exploit ready-made tools to perform malicious actions.

Text content synthesis can also be employed to generate semantically sound content for fake websites and more importantly, to reproduce a specific text style. In particular, style-preserving text synthesis is a technique that employs an AI system to generate a text that imitates the writing style of an individual. Notably, the AI system does not need a large volume but rather only a few samples of an individual’s writing style for its training. As a result, the technique holds some implications specifically for BEC, as it allows a malicious actor the opportunity, for instance, to imitate the writing style of a company's CEO in order to trick target recipients inside the company into complying with any of their fraudulent requests

Indeed, the capabilities of technologies such as GPT-3 lead us to believe that it could truly be, in terms of impact, the next deepfake.

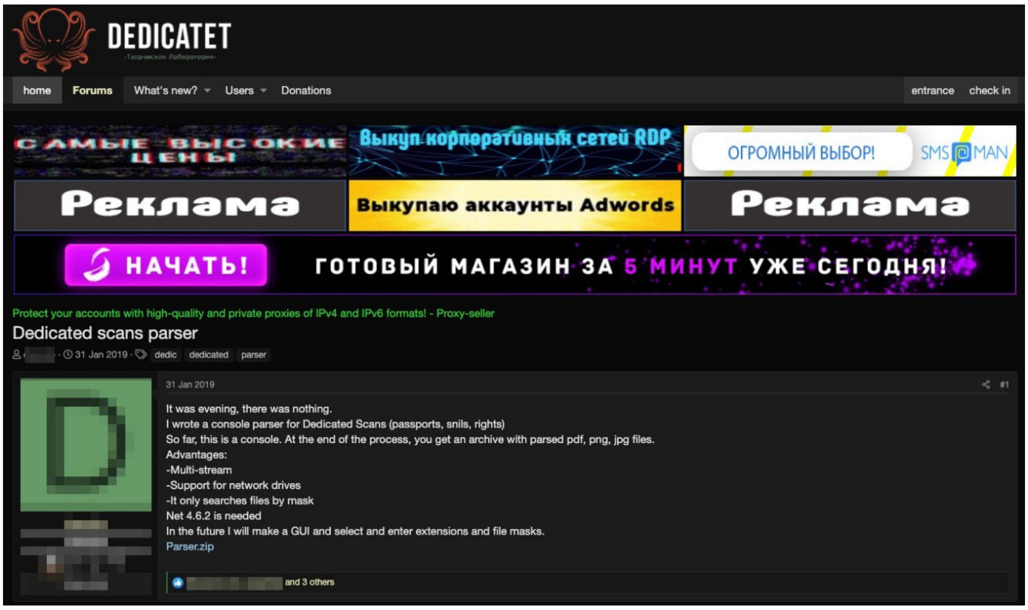
Content Parsing

Content parsing is an activity generally associated with the need for identifying and extracting structured information from unstructured documents. A broader definition might include object recognition and extraction from static images or videos, which fall de facto under facial recognition, covered in the following sections of this document.

More specifically, there is an application linked to text parsing called Named Entity Recognition (NER). In brief, NER tries to identify the semantically significant parts and pieces of individual information, such as names, addresses, and numbers in an arbitrary text. Furthermore, it has the capability to distinguish between credit cards, phone numbers, and currencies, among others. A considerable amount of research has been produced that aim at improving the technology for identifying semantic pieces of text in a more refined way. To illustrate, this kind of identification involves the capability to distinguish, using only context, if the word "bush" refers to vegetation or to "George Bush," the former US president, and in particular, which George Bush the text refers to.

In a similar stroke, malicious actors have been working at developing document-scraping malware that, once installed on a target machine, would go through all of the documents in that machine and either exfiltrate the entire document (which is risky and very inefficient with the growing number of documents), or look for specific bits of information. For example, the installed document-scraping malware could look for all personal employee data on the server of a human resources department.

So far, there are still limitations in the malicious use of document-scraping malware. For instance, its ability to parse content is still rudimentary and based on hard-coded rules and regular expressions. Secondly, one cannot exfiltrate information such as "phone numbers" as yet. Instead, one can look for "sets of digits that resemble a phone number," albeit without the certainty of finding a perfect match. Lastly, the scope of data scraping is still broad, able to filter only by file names or extension.



Dedicated scans parser

[redacted] · 31 Jan 2019 · dedic dedicated parser

[redacted]

31 Jan 2019 #1

It was evening, there was nothing.
I wrote a console parser for Dedicated Scans (passports, snils, rights)
So far, this is a console. At the end of the process, you get an archive with parsed pdf, png, jpg files.
Advantages:
-Multi-stream
-Support for network drives
-It only searches files by mask
Net 4.6.2 is needed
In the future I will make a GUI and select and enter extensions and file masks.
Parser.zip

[redacted] and 3 others

Figure 16. A document scanning tool for compromised servers

However, with the development of more efficient NER techniques in the future, it is likely that more sophisticated scraping malware will be able to better identify relevant content and even perform targeted searches.

For example, malicious actors might be able to evolve from scraping a company server for "everything that resembles a phone number" to scraping "all the emergency contact phone numbers for all the top tier managers."

Aside from text synthesis, there have been several advancements in image synthesis in the last few years, to the point where an AI system is now able to synthetize a photorealistic image from a text description.[53] Needless to say, the ability to generate arbitrary images that are indistinguishable to the naked eye makes for a powerful tool for social engineers. This is another field where the use of GAN models can already produce high-quality results.[54] GANs for image synthesis are discussed in further detail in the case study that follows this paper.

Improved Social Profile Aging for Forums and Botnets

Another intriguing possibility for the malicious use of AI systems involves user behavior emulation. One immediate application would be to use it for evading a security system based on detecting biometrics and specific human behavior. For example, TypingDNA[55] offers a commercial product with the capability to lock one's computer whenever it detects a typing pattern that is different from the usual user patterns. Reports also suggest that AI detection systems are already in use by financial institutions to detect abnormal spending or money transfer patterns. At the same time, there are reports that companies are already starting to see criminals attempting to circumvent these systems by trying to stick to “normal” usage parameters, such as working hours and fake IPs within trusted ranges when operating hacked bank accounts or stolen credit cards.

Another notable application involves the use of these techniques to create false behavior in stolen accounts to avoid their closure. This concept is used on Spotify accounts to automate fraudulent monetization while avoiding detection by the system, as previously described in the subsection titled “Human Impersonation on Social Networking Platforms."

The same could be achieved to maintain activity on stolen Twitter or Facebook accounts that the criminals can then monetize for longer. The quality of these services — which sell likes or followers — can be measured by how “old” these “bot” accounts are. Naturally, any technique that helps criminals keep fake accounts look legitimate for longer is bound to be a precious tool for them.

In order to improve a botnet’s C&C communication, one proposal suggests[56] a swarm-based method that is grounded on AI principles. In this method, upon joining the botnet, the infected node sends a multicast heartbeat. Older members learn the presence of this new bot and update their tables to reflect this. The commands are sent to random nodes which, upon decryption, can be forwarded to neighboring nodes to repeat the process. This is a novel way of managing a botnet. However, at the moment, it is just a theoretical possibility.

AI-based techniques that try to emulate real user behavior are not easy to identify. Conceivably, by using AI techniques specifically aimed at detecting them, it would be possible to increase the likelihood of unmasking those human-like bots. On the other hand, the described serverless C&C techniques[57] appear much more difficult to figure out. It is important to stress that from an outsider’s perspective, network traces would not be enough to piece together what is really happening. Unfortunately, network traces are often the only thing available to a defender. As in all cases, the more data the defender has, the higher the likelihood of them finding out what is really happening. Like a puzzle, the more pieces one holds, the clearer the picture becomes.

Robocalling v2.0

A so-called robocalling operation can be optimized in much the similar way. Robocalling has become a way of performing a phishing scam through a regular telephone. In this kind of scam, an automated caller delivers the voice phishing message and tries to entice the victim to visit a malicious website. By adding smart automation to such a scamming system, the perpetrators can monitor whether the scam is successful or not and what kind of arguments and logic are the most convincing to potential victims. By tweaking those parameters, they can refine the scam to get better results over time. With enough data points, the attackers can leverage these as features to train progressively better ML models to amplify their attacks.

Another intriguing possibility for robocaller improvement might involve the use of audio deepfakes in order to fool the user into thinking that they are dealing with a person whom they know. For instance, a once-popular scam involved receiving some form of communication from an acquaintance who is in distress. Allegedly, they are stranded somewhere in the world without money and are in need of help in order to reach their hotel. This kind of fraudulent distress call can be made much more convincing, however, if it used a voice message from the acquaintance to better convey their concern and panic.

Detecting fabricated sound files is as challenging as attempting to verify any other kind of deepfake. The added layer here is that robocalling scams are perpetrated through a regular telephone. This means that as long as they sound convincing to the user, they could be undetectable because there is no automated layer that can flag this as a scam. In contrast, any other scam where the sound file has a chance to pass a detection filter could be a chance for a defender to find out that the file is a fabrication, not a real sound file.

More information and use cases about audio deepfakes can be found in the deepfakes section.

Criminal Business Intelligence

The same approach in using AI to improve and optimize the effectiveness of criminal operations can be applied to any other scam as well, such as regular email phishing. ML, in particular, is already being applied to improve the success rates of any corporate endeavor from sales to marketing. Presumably, therefore, ML could be just as effective for malicious uses.

Imagine, for instance, that a regular phishing operation targeted at banks adds a small tag on emails or embedded phishing links. Whenever the potential victim receives the email, the scammer would know whether the receiver has seen it (otherwise it must be assumed that it went to the spam folder), or if the link has been clicked on. Additionally, the scammer would learn whether any personal information has been entered on the phishing page, along with the quality of that information.

By co-relating all this data, the scammer can have a good picture of what kind of emails are more successful for each bank. Moreover, they would learn which email databases are more likely to elicit good success rates versus those databases that have been reused repeatedly and would thus no longer yield good results.

By automating all this, the whole scamming operation can be self-sustained, and the scammer can remove email databases that are unlikely to deliver phishing emails, craft new emails that are more likely to succeed, and send them to those addresses that belong to more susceptible people. These engines can also compile new email databases that list potential victims who are more likely to fall into more scams or that identify those people who open fraudulent emails and select links in them.

It is also notable how modern marketing and advertising companies already use similar systems. Therefore, it would not be a leap in imagination to presume that malware-based organizations have started to use the same AI-based methods to improve their own efficiency.

On the same front of empowering criminal business intelligence, AI could also be used to optimize the data flow in Traffic Distribution Systems (TDS).

TDS serve as intermediaries that buy and sell traffic between websites. The main functions of TDS include controlling and filtering web traffic (which involve selecting a link) and collecting related statistics. They can also filter traffic based on the cybercriminal's preference, such as a user's web browser and location (via IP address, for instance).

For example, a cybercriminal might set parameters in a TDS to redirect users in the US to banking trojans but deliver ransomware to other countries and avoid deploying their malware to the Commonwealth of Independent States (CIS). Traffic distribution systems are a staple for distributing malware via exploit kits and drive-by downloads and can act as a service for mass-marketing malware. TDS vendors sell the traffic from when a victim clicks on a link.

While TDS currently rely on hard-coded rules set by malicious actors, it is easy to see how an AI system could be used to optimize the traffic distribution itself in order to maximize the revenue of the malicious actor.

In general, it is likely that in the near future, an evolution of the CaaS concept[58] would include more and more AI-enabled services. This is because AI-related activities generally require a substantial initial investment in terms of human knowledge and training datasets that could be easily recovered by offering AI services to multiple interested parties.

Abusing Image Recognition Systems

Autonomous Cars

Vehicles use a variety of sensors. These include cameras to perform ML-guided image recognition of signs and other elements of their environment such as pedestrians and lane separation lines. These ML models also determine the appropriate behavior for the vehicle to take. In an autonomous car from the 4G era, this would not be a serious issue since, owing to poor data quality, the driver has to be in control at all times on public roadways, and full autonomy by the vehicle is not possible. In a 5G-connected and fully autonomous vehicle, however, the visual sensors might have the ability to both perceive traffic lines and execute navigation based on what they see.

To add, autonomous cars are "trained" to obey the law (to not cross solid traffic lines, for example), but this training could also be exploited to attack an autonomous car's AI models. "Autonomous Trap 001," a piece of performance art by James Bridle,[59] is based precisely on such a concept. The trap, illustrated in the following figure, involves drawing two concentric circles in salt, with the inner circle drawn in solid lines and the outer circle drawn in broken ones. Since an autonomous vehicle is trained to know that it can cross broken traffic lines, it can enter the circle with ease. However, the same vehicle would not be able to leave the circle without crossing the solid lane lines, which would mean breaking the law.

Placing traps (such as this circle) in crime-convenient spots that are covered by a jammer (or in a place without cell coverage) would thus collect jammer-blinded, autonomous cars like flypaper — an opportunity for chaos that criminals could take advantage of.

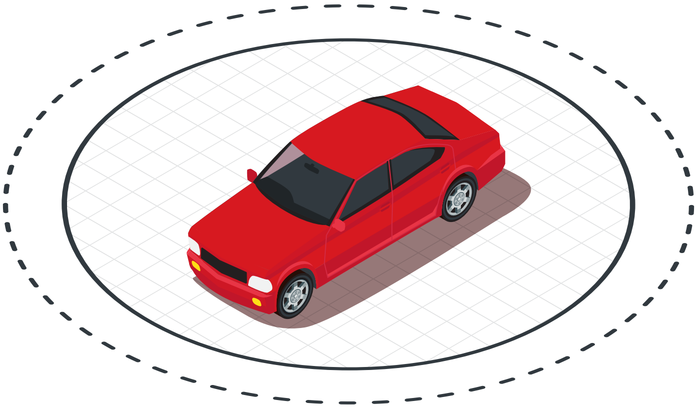


Figure 17. Autonomous cars can be trapped with a ring of "detour left" signs or other similar methods, as illustrated by James Bridle's work that used traffic runes drawn in salt.[60]

Note: The illustration in Figure 17 recreates the photo of the trap in James Bridle's website.

A similar adversarial method that abuses the rules of ML image sensor models can be achieved by making other signs or marks. One example requires only a small piece of black tape on a speed limit sign. In field tests, tape was used to alter the digit "3" in a 35-mile-per-hour speed limit sign. As a result, the image recognition model mistakenly recognized "35 mph" as "85 mph." Consequently, it made this mistake 58% of the time, resulting in the car's acceleration to illegal velocity.[61]

Similar attacks can also be used against AI systems without altering the traffic signs themselves, but merely their placement.

To illustrate, attackers can make a kilometer-wide circle constructed with traffic signs such as "detour left" that would eventually trap all autonomous vehicles that follow the detour sign. The only way to escape would be for their human drivers to prompt an "illegal" way to leave the circle by eventually ignoring — that is, "disobeying" — the "detour left" signs. It is also important to note here that automated traffic signs would allow this attack against autonomous vehicles to be carried out via a hacker making changes to the sign displays.

More importantly, these possible methods for manipulating AI systems that manage autonomous vehicle traffic can be used to manipulate even non-autonomous vehicle traffic by using compliant autonomous vehicles as obstacles. Such methods, when applied to non-autonomous vehicles, could cause major delays in the response time of police and emergency services. Combined with mass malicious ridesharing (the use of many stolen credit cards to swamp an area with rideshare vehicles), this could cause heavy traffic at a downtown core. The resulting traffic could then be increased to amplify the effect of a terrorist event, give more time for a bank robbery to continue, or prolong the duration of a fire by delaying the arrival of firefighters.

The sophistication of and investment in this kind of attack is minimal, and yet the ability to prevent it is complex. These attacks, in fact, only require a small number of autonomous vehicles to become fully effective: Fewer than one in 100 vehicles would suffice, and even fewer during busy times of the day.

AI-facilitated attacks targeting connected vehicle ecosystems, including telematics, infotainment (such as podcasts), and Vehicle-to-Everything (V2X) communication systems, might result in vehicle immobilization, road accidents, financial losses, and disclosure of sensitive or personal data. Even worse, it could endanger road users' safety.[62] Notably, as vehicle automation levels[63] increase and various technologies are progressively integrated into vehicles, the potential attack surface becomes broader as well.

Finally, as the same or similar technology is used in autonomous trucks[64] or autonomous cargo ships,[65] the potential targets for malicious use also increase.

Drones, Connected Skies, and the Internet of Flying Things

High-quality airspace management data, including telemetry from SIM-enabled devices, is used to manage the drone aircraft of the Internet of Flying Things (IoFT). The flight support application "flight control," in particular, is controlled by ML models that react to telemetry received from IoFT on-board sensors[66], including image recognition cameras.

These cameras and other telemetry are fed back their own version of traffic management: in this case a kind of air traffic control that requires 5G-level, high-quality data. For unmanned drones to realize their potential, they must have the kind of real-time feedback that requires ML-optimized 5G.[67]

To operate Beyond Visual Line of Sight (BVLOS), these drones also need several AI-related mechanisms: from computer vision to identify obstacles and references on the way, to a thorough ML-driven optimization of their safe flight (through collision avoidance), safe delivery (including delivery of humans), and safe navigation (by following "no fly" zones) behaviors. Nonetheless, these behaviors can be exploited through the application of gaps in the basic behaviors. Often, these behavioral gaps can be triggered by broadcasting false information via radio, paired with the display of false information such as a large letter "H" in a circle on the ground that might trick a "lost" helicopter into landing. This speculative approach would likely capture ML-optimized drones in a fashion similar to the Iran capture of a US CIA Sentinel drone.[68]

The capture of autonomous drones is of special interest as these tend to contain a variety of interesting payloads such as intellectual property and live encryption keys. These can be used against previously captured outbound radio traffic intercepted from the drone and might then be used to push bad intelligence and false telemetry toward upstream systems. These upstream systems can, in turn, be poisoned with false telemetry, including images prompting tactical or executive action on incorrect targets. A few methods include SIM-jacking or SIM-swapping[69] (using social engineering or other methods to take control of a SIM and all its data), drone fleet management, account hijacking (controlling fleets of SIMs and their data), and others.

In the case of Amazon delivery drones, these could be used to create larger financial incentives, as the target (in this case, the parcel) would be easier to monetize. As a result, this would create the equivalent of "porch pirates" or in this context, "drone pirates."[70]

Additionally, a drone can be caused to drop off narcotics or antibiotics in criminal-controlled locations. Again through SIM-jacking or SIM-swapping, a drone can be hijacked and the payload changed for another to facilitate undetectable smuggling.

A drone with a UDOO single board computer (SBC) can also be used to collect Wi-Fi passwords. As the drone flies, it scans for Wi-Fi signals and once it finds a router, it attempts to hack the router's password. When this proves impossible, the drone's SBC resorts to saving the packet that contains the handshake, to be cracked later on using Aircrack-ng.[71]

As for mitigation, new EU rules for unmanned aircraft are in development.72 One of these involves limiting the size of individual high-damage drone strikes using consumer drones. In effect, the new limit would then also reduce a drone's carrying capacity for explosives, although there is some delay in implementing these regulations due to the Covid-19 pandemic.73

Escaping an Image Recognition System

Image recognition systems such as social scoring and social credit technologies fall under "reputation systems." These are very similar to business intelligence and marketing systems called Social Network Analysis and Propensities (SNAP),[74] which analyze personal social network behaviors to determine the likelihood of a particular action being taken, such as the purchase of a product. These SNAP systems also identify who the "influencer" or leader of a group is. Similar national intelligence systems are used to identify criminal activities such as the purchasing habits of terrorist bomb makers or members of organized groups such as criminals. Without social network discovery tools such as SNAP, it would be difficult to make full use of image recognition tools for discovering who the leader of a criminal group is, or even if such a "group" exists.

Nevertheless, the use of such image recognition systems by a country's government could pose a serious threat to privacy and human rights. For example, during the mid-2010s, NtechLab, a Russian company that specializes in facial recognition, released the FindFace app. With this app, a user can take a picture of any individual and then match their face to their social media accounts. These days, FindFace is no longer a consumer app but an instrument for Moscow's ambitious, multimillion-dollar surveillance project. Using live and real-time recognition, this technology can select individual faces amongst a crowd and instantly match them with a record of criminals in a police database.[75]

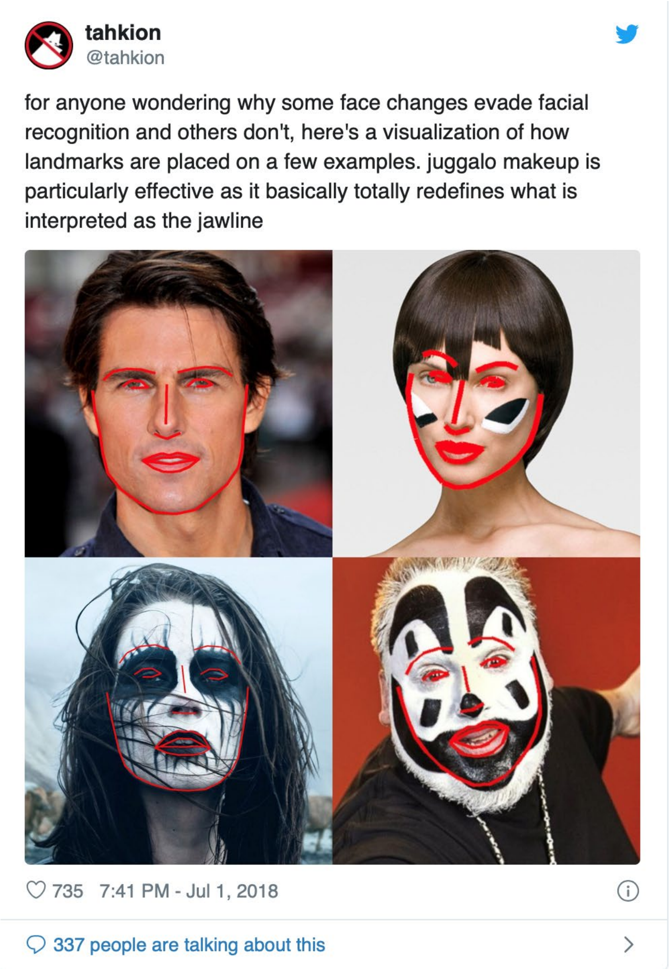
With respect to drones, criminals have already successfully used these — without facial recognition — to deliver contraband into prisons[76] and fire pepper spray at labor protesters.[77] Interestingly, the use of these "para-militarized" functions in the latter example can be speculated as the evolution of black mask counter-protest police activities such as those performed by the Black Bloc during G20 events.[78] Since black mask protest organizers hide in crowds and attempt to provoke more aggressive on-camera police response, they need a way to avoid being hit by their own pepper-spray drones as they attempt to escalate conflict. The need for this extra measure of protection from their own drones is due to the fact that in the fashion of a terrorist cell hierarchy, many members of this group do not know each other. As a result, they are unable to share photos with one another's faces for uploading on a drone safelist.

However, specific designs on face masks (for example, a black-and-red happy face) can be a quick and convenient method for being safelisted. Since it is possible for a criminal group to assign a unique mask logo to each of their members (who might all be part of a safelist), it would be difficult to detect and identify an individual member, even with the use of image recognition for mask logos. Nonetheless, if such a safelist of participant faces belonging to a criminal group did exist, it could serve as a source of evidence were it to be recovered by police. For this reason, participants of organized crime would have to safelist specific person-agnostic facial patterns, which can then be disposed of or wiped off after the event.

It's also worth noting here that techniques for evading facial recognition are used not only by black mask members but also by other protesters. Tribal knowledge on these techniques have been shared online as far back as 2010.

These disposable cosmetic "virtual faces" can protect privacy-conscious individuals. However, they can also be used by malicious groups to avoid eventual police arrests or micro-targeted drone strikes when any of these are based on facial recognition. Similarly, these virtual faces are also used by individuals in countries that have outlawed wearing masks in public.[79] Guises for evading facial recognition will likely be around for a while yet, as the materials for constructing them are cheap, easily hidden, easily disavowed, and easily available. After all, it is unlikely that common consumer items for making such guises (like lipstick or face paint) would be outlawed.

Other disguises, such as the Juggalo clown makeup[80] (used by fans of hip-hop duo The Insane Clown Posse), are promoted as a way to circumvent facial recognition.[81] Through Juggalo makeup, an individual can have parts of their face misclassified, such as their eyes or jawline.[82] In fact, a facial recognition ML classifier system would have to be "tuned" particularly to detect these faces in addition to normal human faces. To do this, federated logging across multiple disharmonized systems could unite the anonymized person with something identifiable such as cellular metadata or a vehicle license plate.



tahkion
@tahkion

for anyone wondering why some face changes evade facial recognition and others don't, here's a visualization of how landmarks are placed on a few examples. juggalo makeup is particularly effective as it basically totally redefines what is interpreted as the jawline

♡ 735 7:41 PM - Jul 1, 2018

337 people are talking about this

Figure 18. A collection of images analyzed against facial recognition: the top left face is correctly classified; the top right face has eyes and right eyebrow misclassified; the bottom left face has eyes and eyebrows misclassified; the bottom right face has jawline, nose, mouth, and eyebrows misclassified.[83]

Image credit: Ian O'Neill/Twitter

Although the aforementioned means are considered as effective for working around facial recognition, some systems (such as Apple's Face ID) use depth perception to identify dimples and other facial features.[84] However, this evolution in the system is less effective when used at a distance and with only a single camera.

Lastly, there is another matter of urgency with regard to facial recognition drones and the danger that they could pose. Indeed, it should be noted that tiny facial recognition drones carrying a gram of explosive are in development today, specifically for micro-targeted, single-person bombings. These drones look like insects or small birds.[85] Given the nature of this technology, devices like these are operated via cellular internet. It is also safe to assume that they will be delivered in the hands of criminals — either through purchase or development — soon.

In sum, addressing the issue of evading facial recognition might be difficult for the entities that use ML-based surveillance systems, a problem that is also being reviewed by the aforementioned CUING Initiative. Particularly, it is paramount for these discussions to discern an ethical process for identifying people who wish to evade facial recognition.

Remote Machine Learning Sets Pollution

Security Algorithms

In the case of security algorithms, when a fraudulent activity is detected, the anti-fraud system typically blocks it and prompts an alert. In an anti-fraud ML model, the activities that provoke an anti-fraud response can be mapped very clearly and effectively. In fact, the anti-fraud alerting thresholds can be built into an attacker's threat model (based on criminal success logs) and represented in their attack infrastructure, architecture, and spending.

Speculatively, these security algorithms would potentially use a business intelligence-based marketing communications analytic tool, such as that of Cloudera,[86] to send specific criminal "marketing campaigns" to a chosen target victim or "customer." When ML is used to perform accurate mapping of the anti-fraud systems of a specific victim group, the "marketing campaign" mentioned previously can be very accurately tuned to always stay just below the anti-fraud enforcement thresholds. In effect, this maximizes the profitability of the criminal enterprise investment.

In summary, by using ML to identify blind spots in detection thresholds (based on what was blocked and what was not) the attacker can stay in that blind spot and remain undetected and persistent until the detection rules are updated. Detection of this could be as simple as using Benford's Law,[87] which states that in naturally occurring collections of numbers, the most frequent leading digit is likely to be the smallest. Therefore, if natural number statistics used in AI do not behave as statistically expected, security rules such as alerting can then be triggered. Looking for "mathematically-consistent anomalies" that operate just below the common alerting thresholds is, as a matter of fact, a statistical rule used in detecting fraud. Meanwhile, more advanced actuarial techniques borrowed from the finance industry would have the capability to detect more complex fraud such as money laundering.

AI-Enabled Stock Market Manipulation

AI-enabled, algorithmic stock trading systems have replaced many human traders in the past years, particularly in the area of high-frequency trading (HFT). However, the increased reliance on such systems has provided opportunities for AI-aided stock market manipulation as well. Indeed, there are already cases where private investors have managed to crack the algorithms of a large electronic marketplace and manipulate share prices in their favor.[88] Experiments with adversarial bots also demonstrate that by using automation, the stock market can be manipulated.[89]

HFT algorithms exaggerate and accelerate market activity. Similarly, stock market "flash crashes" and the consumption of asset value through a series of small returns on larger trading fees can create a sudden unpredicted change in the market with consequent change in investor confidence. This can lead to further systemic risk to the investors if exploited by criminals.

Variations on these security "criminal ML detection" approaches could be used to detect SIM-jacking (through the detection of statistically anomalous behavior), mass wiretap (by giving alerts on unexpected telecommunication metadata such as erratic latency), and other optimized telecommunication abuses. Traditional fraud types like credit card fraud can also be found in nontraditional methods such as SIM card auto-pay and other money-laundering methods.

Business Process Compromise and Injection of Safelisted Telemetry

Phones, laptops, internet of things (IoT) devices, and most other types of internet-connected technology are powerful surveillance devices and can be thought of as sensors. The remote sensor information (called sensor telemetry) is used by corporations that manage the devices. For example, the device can be managed to follow an instruction such as, "When Sensor Telemetry X says that Event Y occurred, execute Event Z." Nevertheless, while most organizations have network and software security, very few have data security. This gap means that attacks relying on the manipulation of data (such as fraud) are relatively undetectable when they also leverage unsecure endpoints such as sensors.

Notably, this telemetry is very important. Very large and complex sensor networks such as carrier telephone networks or national corporate IT networks can make profitable decisions based on detailed insights of the sort generated by AI-based telemetry analysis. When a number of sensors are compromised (through the use of a botnet, SIM-jacking, or other means) an ML model will receive false sensor telemetry, and since it bases its decisions on this telemetry, it unwittingly opens the door to large-scale persistent fraud. Unfortunately, an ML model only has to be tricked a small number of times with false telemetry before it is retrained and updated, thus incorporating this false telemetry into its models. At this point, the attacker can step back and let the AI algorithm, as it were, do their job for them.

AI is normally performed offline. When AI is executed in a real-time production network such as in 5G, the assets previously controlled by the AI models will then be controlled by the attacker. Furthermore, since this data is not filtered or secured at the data level, it can be modified to provoke a disaster by producing responses unexpected by the systems developers, including ones that are subject to abuse.

Example of a Business Process Compromise: Vehicle Navigation Design Gap

In February 2020, the artist Simon Weckert conducted a piece of installation and performance art by filling a small red cart with legitimate, active, fully enabled phones.[90] He then pulled his cart full of 99 active phones down several German streets. Google Maps considered this as 99 cars moving slowly, since in Google Maps, the slow movement of so many phones was interpreted as a high-traffic condition. Although the description of Weckert's installation and performance piece references Moritz Ahlert's "The Power of Virtual Maps," a paper which analyzes power relations in cartography and virtual modifications in the way people interact with maps since Google Maps came to be,[91] the result of his piece has some implication on the design gap in vehicle navigation. Indeed, as a result of Weckert's legitimate activity (perceived as traffic by Google), Google Maps then redirected other real-world cars to other much less effective and more distant paths.

While Mr. Weckert's activity was a legitimate and very original example of performance and installation art, it is therefore easy to see how this might be abused. Using 99 burner phones to send 99 ridesharing cars (or even more) to a certain location at a certain time would achieve the same result, with the added impact of physical cars being actually present and thus forming a real obstacle. One can assume, for example, that enough traffic density would pose a problem to fire, ambulance, and police response.

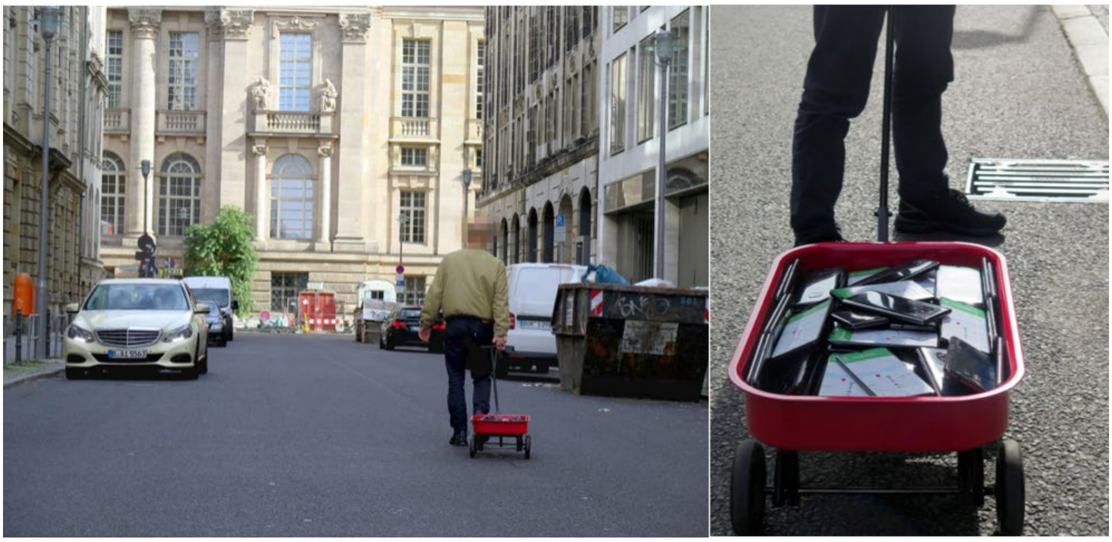


Figure 19. The artist Simon Weckert conducting his performance and installation art with a cart full of 99 active phones in the streets of Germany.[92]

Image credit: Simon Weckert/Simon Weckert

Insider Attacks: Banking and Trading Floor AI

Some bank tasks are heavily automated for the purpose of improving and accelerating their performance. The value proposition of one bank over another might be determined by infinitesimal improvements in the speed of trading (which involves taking advantage of many tiny opportunities),[93] or the tiniest improvement in the accuracy of stock trading (which involves many tiny reductions in loss). The largest banks all have the top trading floor deployments, which results in nearly identical systems between banks and nearly identical financial AI deployments. These AI deployments in financial technology (FinTech) have, in turn, nearly identical mathematical quantitative or "Quant"[94] models, which results in ferocious competition over the most talented staff and the tiniest competitive edge that they can represent.

With regard to this context, AI is important for banks as it can improve their profitability. Meanwhile, human Quants who use the output of AI models are also important since they are the drivers of the global stock market.

There is, however, a deeper level where AI data scientists are responsible for making, tuning, and implementing the models themselves. For competitive reasons, the identities of these individuals are kept secret according to a contract that also contains punitive clauses prohibiting them from working for a competing bank for a long period after they leave their current bank.

Still, it should be noted that these trading algorithm data scientists are in a powerful position because they control the ability of the bank to make both trades and profit. Often, their ML models and code are so obscure that the execution of these models is difficult to reverse-engineer, even for their creator. These scientists are required (as their core job function) to manipulate their bank's presence in stock markets. They are the uncontested secret masters of the single machine that is most responsible for a bank's ability to make money. In effect, nobody knows how they do what they do, except for them.

It is speculated that this opportunity would be too tempting to ignore. For some, the ability to develop AI models that are "self-fulfilling financial prophecies" allows nearly undetectable insider trading since this is a non-alerted behavior. Banking scientists of this sort can dictate the massive buying power of a bank, buy what they know the bank will buy, and then sell after the bank has bought it. Since they sell after the bank has bought, the scientist's stock goes up in value and they can then cash out the difference. This insider trading can be undetectable if performed in the so-called gaps of audit controls under the Anti-Money Laundering/Combating the Financing of Terrorism (AML/CFT) or insider trading controls (which only look for specific indicators).[95] Indeed, none of these controls can detect ML-driven insider trading if certain steps are taken. If the AI rules treat AML/CFT-auditable behavior as off-limits, this would mean that the only insider trades that are performed would be the ones for which there are no AML/CFT or other audit functions. Additionally, it should be noted that this practice is already being followed by criminals today, albeit manually through reverse-audit of global standards.

This practice can also be reverse-engineered to short-trade a stock. Speculatively, an algorithmic AI money-laundering type sequences these trades and splits their results for them to remain below financial detection thresholds.

Local Library Poisoning by Resident Malware

Present-day malware can alter the general behavior of a computer system by injecting external code into resident software libraries, for example, to persist reboots or trigger behaviors under specific conditions. This kind of deterministic mechanism is nowadays relatively easy to detect, as it involves either a code injection with a specific and detectable signature, or even in cases of heavily encrypted code, a behavioral signature that can be detected by an external agent.

However, ML, and deep learning systems in particular, do not follow the philosophy of a classical computer program. As we mentioned in the previous subsection titled “AI Malware,” decisions are not encoded in deterministic and explicit lines of code but are rather encoded in the form of trained models that are represented by matrices of numbers.

This approach raises the known problem of the explicability of machine learning models: Whenever an ML system makes a decision, for many other ML algorithms this decision is taken by a black box of pre-trained matrices without a clear, explicable reason for taking a particular decision. For example, a spam filter that relies on deep learning to identify spam content would be very hard to debug in case of a false positive, as it would lack any specific information explaining why it classified a given message as spam.

Criminals, for their part, might see this particular issue as a source of future opportunities. Whereas a library alteration might be spotted by either code or behavioral signatures — by looking for anomalies in either the code or its execution, for instance — anomalies in an ML model would be harder to identify due to the model’s lack of explicability.

Therefore, future malware might impact host machines by targeting, exfiltrating, or altering an ML model directly, rather than libraries or software APIs. After all, a difference in an ML model that is due to malware would be hardly distinguishable from a difference that is due to a model update.

It will thus become paramount to ensure that integrity check mechanisms for ML models used in production are put in place, in the form of either checksum algorithms or integrity-training sequences with known outcomes for the model.

AI-Supported Ransomware

The 2017 WannaCry and NotPetya cyberattacks spread worldwide with an unprecedented speed due to self-replicating functionalities.[96, 97] In light of the proliferation of ransomware attacks and the continuous efforts of cybercriminals to optimize their effectiveness, AI-supported ransomware attacks might emerge with self-propagation features. In the future, deep neural networks can be abused to enhance the target selection based on predefined attributes or used to disable security measures of the target system[98] that will facilitate lateral movement. Additionally, AI could exacerbate ransomware attacks by intelligent targeting and intelligent evasion.[99, 100] Intelligent targeting would allow for finding new vulnerabilities by using a combination of attack techniques and leveraging the most effective ones in terms of infiltration.[101]

The impact of such an AI-supported ransomware attack on international private companies could be devastating, as their global affiliates could be rapidly infected. Recent experience[102, 103, 104, 105] indicates that cybercriminals progressively use ransomware to target entire municipalities or city-wide services, which could make potential AI-supported ransomware attacks even more damaging. The same possibility applies to critical infrastructure and essential services that are also highly vulnerable to such attacks and could result in serious real-life consequences with a large geographical reach.

Escaping AI Detection Systems

AI provides extensive possibilities for attackers to enhance their operational security and avoid detection. Ransomware families such as Cerber have already employed a multilayered antimalware detection approach, including a separate loader designed to evade ML solutions.[106] Attackers can also use AI for protection of their criminal infrastructure by detecting intruders or suspicious environments (such as sandbox), or they can create a self-destructive mechanism in the malware that is triggered when a suspicious behavior is detected.[107] Such forms of evidence erasure will then significantly obstruct the processes of evidence gathering, investigation, and technical attribution.

Fraud and Voice Recognition in Banks

Banks use voice authentication for a variety of purposes, especially in securing their “softest security perimeter,” the human element of call centers. Call centers, by their nature, employ staff that are subject to some of the most sophisticated fraud practitioners in the world who, in turn, often have very advanced voice-acting skills.

The objective of these attackers is account hijacking. They seek to gain access to account credentials or ask for a password reset. Otherwise, they aim to gain access to the target customer’s account authority. This account authority is used to drain the victim’s account of resources or even to gain credibility for even larger attacks. This secondary target might also include authorizing large payments from the victim’s company.[108]

The adoption of voice biometrics has reduced the incidence of fraud against customers by as much as 90% in account takeover cases.[109] The accuracy of these models is driven by the overlap of two types of AI systems: voice analytics and fraud analytics. When both function properly, these AI systems are combined to reduce call center attacks on banks, as well as online attacks from a variety of sources. These include performing brute-force attacks on customer accounts (since they now need voice matching), compromising the PIN database (in which the PIN is a voice), credential sharing (people cannot share their voice), phishing or vishing (a person cannot easily give up their voice), and PIN reset function (the reset PIN is still the person's voice and would still be difficult to copy).

ML-powered voice authentication ironically finds as its weakness the inherent trust that technologists place in new technologies. When the security represented by voice biometrics is compromised by voice-cloning tools such as LyreBird[110] or others that are less legitimate, the assets protected by voice biometrics are placed at risk. Currently, this situation with ML-powered voice authentication is an arms race in which the accuracy of detection models used in voice security platforms (such as Nuance[111]) is in competition with the accuracy of models generated by non-security, voice-cloning ML.

Recommendations

This section of the report proposes a non-exhaustive list of recommendations on how to enhance preparedness to address the current and potential future evolution of malicious uses and abuses of AI.

- **AI for Good:**
  - Harness the potential of AI technology as a crime-fighting tool to future-proof the cybersecurity industry and policing, building on ongoing efforts to ensure that AI is trustworthy: meaning that it is lawful, ethically adherent, and technically robust.
  - Promote responsible AI innovation and the exchange of best practices in public forums such as the AI for Good Global Summit,[112] the Europol-European Network and Information Security Agency (ENISA) IoT and AI Security Conference,[113] the Europol-INTERPOL Cybercrime Conference,[114] and the UNICRI-INTERPOL Global Meeting on AI for Law Enforcement.[115]
- **Further research:**
  - Enhance cyber resilience to present and future malicious uses or abuses of AI by developing specific, in-depth, forward-looking threat assessments and continuously mapping the AI threat landscape.
  - Adopt risk management approaches to classify the threats stemming from the current and future uses and abuses of AI and prioritize the response measures accordingly.
- **Secure AI design frameworks:**
  - Promote the development of robust AI systems following security-by-design and privacy-by-design principles.
  - Develop specific data protection frameworks to enable continuous development, experimentation, and training of AI systems in a secured and controlled environment (sandbox concept) to ensure maximum accuracy of AI systems.
  - Encourage the adoption of a human-centric approach to AI, such as the European Union (EU) framework for Trustworthy AI.[116]
  - Set forth technical standards to promote the cybersecurity of AI, such as the efforts of the European Telecommunications Standards Institute (ETSI) Industry Specification Group on Securing Artificial Intelligence (SAI).[117]

- **Policy:**
  - De-escalate politically loaded rhetoric on the use of AI for cybersecurity purposes that can obstruct the ability of the cybersecurity industry and law enforcement authorities to respond to and stay ahead of cybercriminal developments.
  - Promote the development of internationally applicable and technology-agnostic policy response measures to prevent the potential malicious use of AI without impeding the innovative and positive applications of AI.
  - Acknowledge that the implementation of AI for cybersecurity and crime-fighting purposes can leave an impact on individual rights; address these concerns by systematically fostering informed public debates in order to generate public consent and develop appropriate measures.
  - Ensure that the use of AI for cybersecurity purposes, including by law enforcement authorities, follows ethical guidelines and is subject to effective oversight to provide for long-term sustainability.
- **Outreach:**
  - Leverage public-private partnerships and foster the establishment of multidisciplinary groups of AI cybersecurity experts, such as the ENISA Ad-Hoc Working Group on AI Cybersecurity.[118]
  - Enhance AI literacy and cyber hygiene by developing innovative prevention and awareness-raising campaigns on AI cyberthreats, such as Trend Micro's Project 2020 web series.[119]
  - Promote capacity building, cross-sectorial collaboration, and knowledge sharing on tools, tactics, and techniques to deter, detect, mitigate, and investigate misuses of AI.

Case Study: A Deep Dive Into Deepfakes

Deepfakes

One of the most visible malicious uses of AI (and perhaps the most immediately tangible, damaging application) is the phenomenon of so-called deepfakes. A portmanteau of "deep learning" and "fake media," deepfakes involve the use of AI techniques to manipulate or generate visual and audio content that is difficult for humans or even technological solutions to immediately distinguish from authentic ones.

The technology has been lauded as a powerful weapon in today's disinformation wars, whereby one can no longer rely on what one sees or hears. Moreover, coupled with the reach and speed of the internet, social media, and messaging applications, deepfakes can quickly reach millions of people in an extremely short period. With regard to such a short timeframe, deepfakes present considerable potential for a range of malicious and criminal purposes which includes:

- Destroying the image and credibility of an individual,
- Harassing or humiliating individuals online,
- Perpetrating extortion and fraud,
- Facilitating document fraud,
- Falsifying online identities and fooling KYC mechanisms,
- Falsifying or manipulating electronic evidence for criminal justice investigations,
- Disrupting financial markets,
- Distributing disinformation and manipulating public opinion,
- Inciting acts of violence toward minority groups,
- Supporting the narratives of extremist or even terrorist groups, and
- Stoking social unrest and political polarization

Other, more far-fetched examples include the possibility of triggering an international diplomatic incident, for instance through the release of a deepfake involving a leader of a country with advanced nuclear technology announcing or calling for a nuclear strike against another country or territory. Although unlikely to result in a kinetic retaliatory response, it could momentarily create mass panic and confusion.

One side effect of the use of deepfakes for disinformation is the diminished trust of citizens in authority and information media.[120] Flooded with increasingly AI-generated spam and fake news that build on bigoted text, fake videos, and a plethora of conspiracy theories, people might feel that a considerable amount of information, including videos, simply cannot be trusted, thereby resulting in a phenomenon termed as "information apocalypse" or "reality apathy."[121] In fact, one of the most damaging aspects of deepfakes might not be disinformation per se, but rather the principle that any information could be fake.[122] This leaves the door open for another damaging aspect of the spread of deepfakes and fake news in general. Ultimately, difficulty in authenticating videos allows any compromising information to be denied because any audio-visual content could be fabricated. In other words, anyone can claim a video to be fake, thus completely shirking any accountability for their actions. This is how deepfakes could be used not only to threaten our already vulnerable information ecosystem but also to undermine the possibility of a reliable shared "reality."

Deepfakes began their rise to prominence around the end of 2017 when an anonymous Reddit user posted videos of popular figures such as Taylor Swift, Scarlett Johansson, Aubrey Plaza, Gal Gadot, and Maisie Williams with their faces superimposed on the bodies of women in pornographic movies.[123] Although these videos were quickly taken down by the hosting platform, the potential of this unexpected facial replacement technique rapidly gained both media and public attention, leading to its propagation across the internet.

While this case study focuses on the malicious use of deepfakes, it is important to note that there are also many positive applications of this technology. For example, video synthesis can allow the automobile industry to simulate autonomous car accidents in order to learn and avoid errors in the future, while the film industry can use this technology for digital resurrection of deceased actors. Voice synthesis can equally be used to "enable" individuals to speak several languages with their own voice, while medical researchers can use it to develop artificial voice replacements for people who have lost the ability to speak.[124]

The Technology Behind Deepfakes

Deepfakes are created by manipulating content to produce a synthetic audio-visual file and make it appear genuine. The technology used to generate deepfake audio-visual content is contingent upon the different types of digital forgeries. In general, deepfake images and videos can be categorized as follows:

- **Face replacement:** swapping the face of one person by stitching it onto that of another person

- **Face re-enactment:** manipulating the features of a target's face to make it look like they are saying something that they are not
- **Face generation:** creating convincing, yet entirely fictional synthetic images of individuals
- **Speech synthesis:** generating audio content by using and training algorithms to create a deepfake voice or a synthetic audio file
- **Shallowfakes:** creating audio-visual forgeries of a less sophisticated nature using rudimentary editing techniques

Independent of the format (sound, image, or video), synthetic media can be generated using GANs, an innovative unsupervised machine learning application invented in 2014 that has revolutionized the deep learning field. It consists of two artificial neural networks: The generative network generates new data with the same characteristics as the training set, while the discriminative network separates generated data from training data.[125] For example, a GAN trained on photographs can generate new ones that look similar to a dataset of stored photos. The discriminator receives randomly-produced photographs by the generator from the real dataset and tries to identify them. The generator's objective is to fool the discriminator network by producing increasingly better novel candidates, while the aim of the discriminator is to become even better at detecting synthesized images. This loop process drives both the discriminator and the generator to improve the creation of real-looking media. A GAN can also be trained on thousands of photos of faces to produce a new portrait that approximates those photos without being an exact copy of any one of them.[126] Although deepfakes usually require a large number of images to create a realistic forgery, in the near future GANs will be trained on less information and be able to swap heads, whole bodies, and voices.[127]

Another technique uses a more primitive type of artificial neural network called autoencoder. This neural network consists of an encoder that compresses an image by reducing it to a lower dimensional latent space, and a decoder that reconstructs the image from the latent representation. Deepfakes use this architecture by having two encoder-decoder pairs, where each pair is used to train on an image set. During the training process, one encoder finds and learns similarities between the two faces and reduces them to their shared common features, such as eyes, nose, and mouth positions. Two networks use the same encoder but different decoders for each image. To illustrate, one decoder recovers Face A and another decoder recovers Face B. To swap faces, the image of Face A is encoded with the common encoder and decoded with the decoder of Face B. The decoder then reconstructs the face of Person B with the expressions and orientation of Face A. To produce a convincing video, this has to be done on every frame.[128]

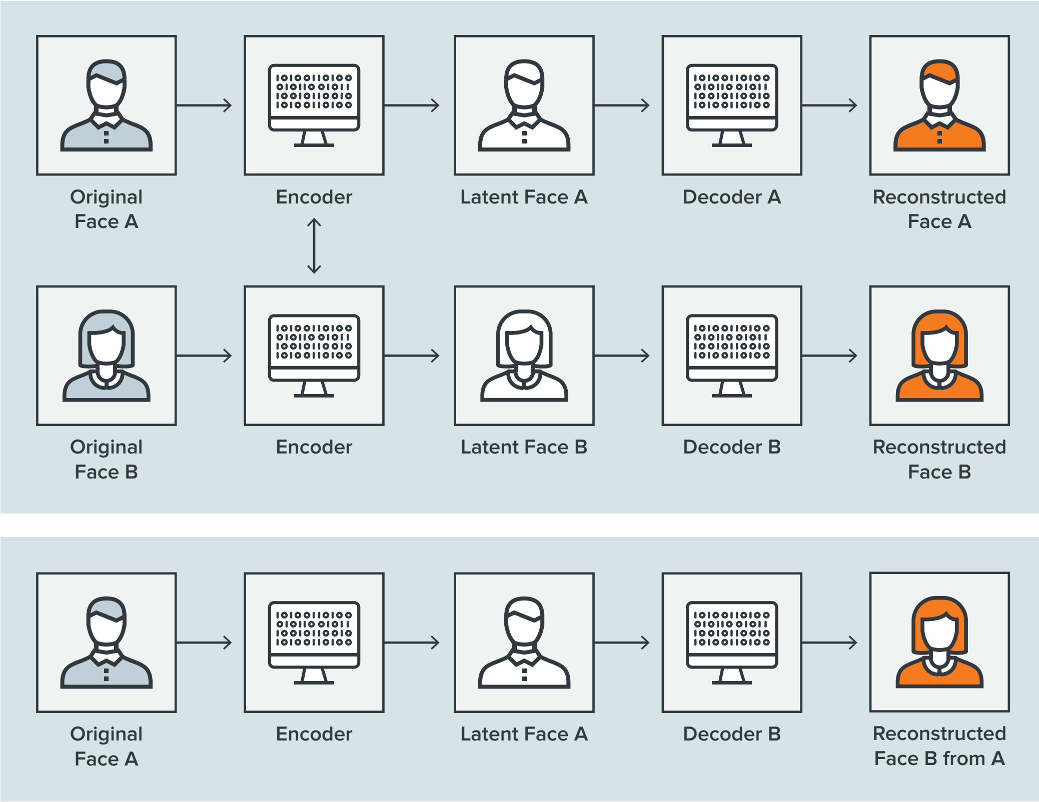


Figure 20. A deepfake creation model using two encoder-decoder pairs. Two networks use the same encoder but different decoders for the training process (top). An image of Face A is encoded with the common encoder and decoded with Decoder B to create a deepfake (bottom).[129]

Alongside AI-powered deepfakes, less sophisticated forms of audio and video manipulation have also come to the fore and therefore create the possibility for criminal or malicious use. As aforementioned, these shallowfakes, also known as cheap fakes, are videos or audio recordings that are either presented out of context or are doctored with simple editing tools, such as deliberately slowing down or speeding up the pacing to portray the subject in a misleading manner. Other possible manipulations consist of selectively splicing together some unaltered clips of a speech, removing or softening the tone of the text, or adding context by changing the start and end times of a video. Such editing does not actually alter the original content but merely reframes how the viewer sees it, thereby ascribing new meaning to a previously innocuous video. Although they differ from deepfakes in that their creation does not involve the use of AI, shallowfakes nevertheless merit acknowledgment as part of the broader deepfake phenomenon.

As indicated previously, audio recordings can also be deepfaked to create audio content, such as "voice skins" or "voice clones" of public figures. For deepfake audio, machine learning engines are trained on conference calls, YouTube, social media updates, and even TED talks to copy the voice patterns of some target individuals and then generate new audio clips with the same voice characteristics. Because this method can use the real voice of the targets and it is possible to splice their words up, algorithmically generated voices alone can sound incredibly real.[130]

Deepfake Creation: Apps and Tools

Deep learning models can be employed by malicious actors who, in turn, can easily use open-source tools that are available on GitHub, distributed on forums, or sold in underground markets. Additionally, they can use functions available for commercial or personal use from major software companies such as NVIDIA and Google. This means that while technical knowledge and understanding of computational parameters are required in order to develop such a technique, much of the software is already available to the public for general use.

A challenge with deepfakes creation, however, is that deepfakes are hungry for data, which is to say that they require large amounts of audio-visual data to train their models. However, service portals that require a training dataset as small as 250 photos have already emerged, thus changing the dynamic and facilitating deepfake creation.[131] For example, Samsung's AI Center created a model that needs only a few images to transform it into a realistic "talking head" video.[132]

Mobile phone apps have also transformed the process of deepfake creation, enabling users to create deepfakes with simple apps. Examples of this are FakeApp,[133] the first app that greatly contributed to deepfake hype in January 2018, FaceApp,[134] a Russian face-aging app that went viral in 2019 but quickly encountered privacy policy issues that hampered its success, and finally Zao,[135] a Chinese app that enables users to swap their faces with film or TV characters on which the system has been trained. There are also several websites that offer deepfake creation upon request as a service.[136]

The growing commodification of tools, services, and apps are helping to serve as a workaround for the technical complexity of this technology. More importantly, it will lower the barrier for individuals or groups who are not experts to create deepfakes, increasing the possibilities for their malicious use in the near future.

The Current State of the Abuse of Deepfakes

As noted, early examples of deepfakes involved actors having their faces woven into pornographic videos. However, since then the use of deepfakes has expanded to include political leaders, comedians, entertainers, and so on.[137]

Apart from their use for the purposes of non-consensual pornography, there have been surprisingly few reported examples of the malicious use of deepfakes. In fact, some of the most prolific deepfake videos that have gone viral were not created for malicious intent, but rather for the purpose of raising awareness with the goal of highlighting the potential threat of this technology. This was the case with the deepfake video of former US President Barack Obama that used the voice of actor and director Jordan Peele,

* An example is available at https://deepfakesweb.com.

who collaborated with Buzzfeed in April 2018 to create the video.[138] Another example features Mark Zuckerberg, CEO of Facebook, where he appears to talk about Facebook's fake video policies.[139]

While there are not many reports of deepfakes being used for malicious or specifically criminal purposes, some examples of deepfake (or alleged deepfake) use have been identified. These examples have, in particular, been observed in the political scene. In 2019, a deepfake video that went viral in Malaysia involved a political aide who appeared to confess to having had homosexual relations with a cabinet minister. Additionally, the video included a call to have the minister investigated for alleged corruption.[140] While the motive behind the video (beyond character defamation) remains unclear, it succeeded to wreak havoc politically and destabilize the coalition government.

In late 2018, a video with Ali Bongo Ondiba, the President of Gabon, who had not been seen in public for several months and who had been believed to be in poor health or even possibly dead, was released. In effect, it sparked a national crisis.[141] The belief that the video was a deepfake seemed to confirm theories that the government was keeping the President's true condition under wraps. In reaction to this, the Gabonese military staged an unsuccessful coup d'état. Although it appears that the video was likely not fabricated, the potential political ramifications are clearly evident in this case.

Beyond the political realm, deepfakes have demonstrated the potential to cause damage in other areas. Recently, well-known Malaysian actor Zul Ariffin claimed that a viral sex video on social media with a man resembling him was a deepfake.[142] Aside from resulting defamation, the release of the video was alarming since the distribution or display of pornography in Malaysia is a criminal offense. That the video in question is a deepfake has, however, not been verified and thus it remains only an alleged case. This exemplifies one of the most damaging aspects of the spread of deepfakes — and fake news in general for that matter — which is that anyone can claim compromising audio-visual content to be fake.[143]

Shallowfakes have also been seen to present a political threat. For instance, a video of United States House of Representatives Speaker Nancy Pelosi was slowed down, giving the false impression that she was drunk or impaired. In this case, the shallowfake generated as much media frenzy and political controversy as its deepfake cousins.[144]

As already observed, deepfake technology can create convincing but entirely fictional photos from scratch. In 2019, a phantom LinkedIn profile under the name "Katie Jones" appeared online. Katie, a young professional in her thirties, worked at the Center for Strategic and International Studies and was connected with a number of US governmental officials. Experts who examined Katie's profile flagged it as fictitious and established that her profile photo was AI-generated, created most probably in an attempt to lure people of interest and collect information, possibly as part of an intelligence-gathering operation.[145]

One of the more prominent ways in which deepfakes can be used for major criminal financial gain involves deepfake audio rather than video or image. In March 2019, the chief of a UK-based energy firm transferred nearly 200,000 British pounds (approximately US$260,000 as of writing) to a Hungarian bank account after being phoned by an individual who used deepfake audio technology to impersonate the voice of the firm's CEO.[146] Meanwhile, Symantec reported that it saw three successful fake audio attacks (including the UK case) on private companies that used a call from the "CEO" to a senior financial officer requesting an urgent money transfer.[147] In another example from July 2020, NISOS discovered a similar attack targeting a tech company employee, with the aim of persuading them to send the attackers money.[148] Criminals also created a shallowfake audio of Europol's former executive director for major financial gain by using audio from a previous public interview and attempted to extort 10,000 euros (approximately US$11,790 as of writing) from a victim. The scheme was revealed for what it was, however, when the victim phoned Europol to inquire about the return of their money. Consequently, they were also advised not to send any more money to the fraudster.[149]

A deepfake audio was also recently submitted to a court in the UK in a custody battle over children. The fake audio clip, which allegedly demonstrated the father threatening another party, was used in an attempt to undermine the father's case to secure custody.[150] The fabrication of evidence is, of course, a criminal act. The unique developments in this case raise particularly serious concerns regarding the credibility and admissibility of audio-visual footage as electronic evidence before the courts and serves as an important warning for criminal investigators, courts, and legal professionals regarding audio-visual evidence in the era of deepfakes.

Another malicious abuse of deepfakes that received limited media attention but has serious security implications is the so-called morphed passport photos, also known as "morphing attack" or face-morphing trend. Face morphing allows a single passport to be used by two or more individuals through a process that merges their photos into one that retains a likeness to both or all individuals. The morphing process is made possible through the use of GANs that can trick both human ID examiners and the AI behind biometric facial recognition systems.[151] The morphed face photos are currently used in two ways, namely on Fraudulently Obtained Genuine (FOG) documents or falsified IDs.

In the first attack vector, criminals target the print application issuance process of an ID, which requires an applicant to submit a printed photo along with their application form (as opposed to the live enrolment process, in which a photo is taken of the applicant during the application process itself). However, it is possible to submit a fraudulent renewal application for an ID using a morphed photo of individuals who look similar. Since the size of the required photo is small, it is difficult to detect if it is a morphed photo. As a result, this falsification contributes to the issuance of FOG documents. Furthermore, since the process of generating FOG documents is in turn facilitated by false breeder documents (that is, documents that serve as a basis for obtaining other IDs such as birth certificates),[152] the identification of a fraudulent document at later stages of the application process has become increasingly challenging.

The second attack vector involves using a morphed photo during the falsification process of an ID. In this case, a forger can print the morphed photo over the genuine photo of the original ID. This type of falsification can also be done by using stolen IDs.

Unfortunately, current research indicates that both human examiners without special training and traditional solutions for identifying morphing attacks have low face-morphing detection rates. Both are also prone to error.[153, 154, 155, 156] It is worth emphasizing that this trend of using morph-based identity for fraud significantly undermines passport-issuing authorities, border security controls, and identity verification procedures.[157,158]

Potential Reasons for the Low Rate of Adoption of Deepfakes

Although examples of the malicious abuse of deepfakes such as the aforementioned have been witnessed, deepfake creation technology is not being employed on a large scale by criminals yet, compared with what might have been expected since its emergence in 2017. The main use of deepfakes still overwhelmingly appears to be for non-consensual pornographic purposes. Notably, in September 2019, the AI firm Deeptrace found 15,000 deepfake videos online, of which 96% were pornographic and 99% of which used mapped faces of female celebrities onto pornographic actors.[159] That there is not more evidence of the abuse of deepfakes by criminals or for criminal purposes raises the important question of why. Why, in spite of the potential for the malicious use of deepfakes, have there not been more instances of these? Here are some potential reasons:[160]

1. **Short life-span.** With high degrees of globalization and interconnectivity, fake content is generally flagged within a short period. In this regard, fake content is likely only to generate an immediate shock factor that would shortly wane afterward.

2. **Technical skills.** Although the technology is available in tools and packages that facilitate its implementation, it nevertheless requires a high degree of skill to produce very convincing videos. It is also plausible that criminals would struggle to find the necessary data scientist required to create a believable deepfake. Cheap fakes or shallowfakes, which do not require such skills, on the other hand, are generally more readily identifiable and thus would be a less effective tool for criminals.

3. **More effective “traditional” attacks.** Current criminal tools, technologies, and approaches are likely to be more effective and easier to employ. They would also be less costly and less time-consuming. Hence, if the objective is to generate profit from the use of deepfakes, there are easier ways to make money legally rather than illegally.

Although understanding these factors could shed light on why there have been relatively few instances involving the malicious abuse of deepfakes, it would be prudent to note that these factors might not always be applicable. As technology becomes more available to the general public and more convincing synthetic media becomes possible to generate, eventually criminals will no longer require such high levels of technical skill, and deepfakes might become a more effective means of attack.

Some Possible Future Threats of Deepfakes

The rise in the manipulation of images and voices for criminal purposes has been identified as a potential threat by many, including Europol[161, 162, 163] and UNICRI.[164] While the large-scale abuse of deepfakes by malicious actors and criminals in particular has not been witnessed to date, it is significant to note the potential of this technology when it is combined with the growing accessibility to and increasing quality of deepfake creation tools. As a result of these two conditions, there is merit in hypothesizing possible scenarios that speculate how criminals might soon use the technology.

The following have been identified as conceivable possibilities in the near future:

Scenario 1: Disinformation campaigns. Deepfakes could be used as part of mass audio-visual disinformation campaigns to create believable fake content that impersonates high-profile public officials in order to interfere with elections or incite violence or public unrest. For instance, deepfake videos of a politician taking a bribe or confessing complicity in a crime, or soldiers committing war crimes or a terrorist group brutally assaulting individuals from a minority, could likely cause domestic unrest, protests, and disruptions in elections. They could even lead to international conflicts if governments were to mobilize the military based on such fabricated videos.[165] Although deepfakes for disinformation purposes have a short-lived effective window, attackers can still leverage them if the window itself is kept in mind and built upon as part of the modi operandi. For instance, the short window of deepfakes can make them crucial instruments in situations where an immediate emotional response can be generated or in circumstances where time is of the essence, such as on the last day before an election.

Scenario 2: Securities fraud. Deepfakes could be instrumentalized as part of stock market manipulation or fraud schemes. For example, a deepfake showing a chief executive saying racist or misogynistic slurs, announcing a fake merger, or making false statements of financial losses or bankruptcy, could interfere with stock market values, and so could deepfakes that portray a chief executive seemingly committing a crime.[166, 167] Although the stock value can recover significantly once the video is claimed as fake, it could nevertheless trade at a much lower value than before, which would create a loss in the company that grows as concerns about the authenticity of the video linger globally. A deepfake of a senior company official could also be used to facilitate the theft of trade secrets or sensitive company data by deceiving company employees to send privileged information to criminals. In turn, this kind of stolen information can deeply affect the company's finance and stock value.

Scenario 3: Extortion. By threatening to release a fake video that would damage the reputation or credibility of an individual or legal entity, criminals could also use deepfakes to extort a ransom or privileged information. It is also possible that the use of deepfakes for extortion purposes could be combined with a ransomware attack.[168] Cyber-related extortion attacks oftentimes demand ransom to be paid in cryptocurrencies,[169] which further limits the investigative possibilities by the competent authorities. It is therefore plausible that criminals would combine the abuse of sophisticated audio-visual deepfakes with extortion attacks demanding alternative means of payment to avoid detection.

Scenario 4: Online crimes against children. Deepfake technology could be used to create a fake identity for the purposes of child grooming or to convert the face of an adult to that of a child or minor, thus raising concerns about entirely new challenges to preventing the online sexual exploitation of children.[170] Additionally, deepfakes might be used by criminals to facilitate online child sexual coercion and extortion[171] by generating sexually explicit material that depicts the victim or other children. Criminals could also use deepfakes to depict an offline sexual encounter and then demand payment by threatening to publicly release the deepfake nude photo or video.[172] This possibility is of particular concern due to the growing availability of self-generated explicit material (SGEM) online that is driven by the growing access of minors to high-quality smartphones and a lack of awareness of the risks associated with sharing sexually explicit content on online platforms.[173]

Scenario 5: Obstruction of justice. Deepfake audio-visual content could be mischievously submitted as "legitimate" evidence in an attempt to frustrate criminal investigations and judicial proceedings, thus casting doubt on audio-visual evidence as an entire category of evidence. This would create new legal hurdles for investigators and lawyers, not to mention undermine the credibility of proceedings, including the institutions and individuals participating in the same proceedings. Even the administration or justice system could be questioned as well. Moreover, existing vulnerabilities in widely-used technologies such as CCTV or police body cameras could be used in order to replace real-life footage with deepfakes. Ultimately, the possibility of hacking surveillance camera systems opens the door to conceal any kind of criminal and malicious practices.

Scenario 6: Cryptojacking. It is possible for deepfakes to become associated with the spread of malware and cryptojacking. Recently, researchers identified a website devoted to deepfakes that was being used as part of a cryptojacking scheme. Deepfake hobbyists, which refers to the category of individuals who tend to be more likely to have powerful computers than the average individual, are thus ideal targets for cryptojacking, a malicious attack that seems to exploit computing resources to mine valuable cryptocurrencies, such as Bitcoin.[174]

Scenario 7: Illicit markets. The proliferation of deepfake development toolkits and services for malicious purposes could become an emerging profitable underground market. Deepfake creation services are already offered in various underground markets and online forums for modest prices.[175] In combination with the CaaS business model,[176] the proliferation of tools and "for hire" services online are likely to grow in the near future.

Countering Deepfakes

While deepfakes present opportunities for those with malicious intent and could become a significant threat to society, political systems, businesses, and cybersecurity, there are still measures that can be taken to control and limit the impact of this technology. These include the development of technology for deepfake detection, content authentication, and deepfake prevention, as well as the adoption of policies, legislation and regulation, awareness-raising campaigns, capacity building, and the creation of channels for reporting deepfakes (including notice-and-takedown procedures).

Deepfake Detection

Furthermore, manual detection of deepfake videos and images, especially the less sophisticated ones, is still possible, particularly if content moderators or competent authorities are properly trained to detect scene inconsistencies (such as strange shadows and lighting, outdoor background noise in an indoor video, sudden changes in the background, and other remarkable details) as well as speaker inconsistencies (like lack of or irregular blinking, facial micro-expressions, a mismatch between facial and body skin tones, lip movement and audio mismatch, and other similar discrepancies). However, manual identification and human filtering of audio-visual inconsistencies is only possible for a small volume of suspicious files. It is neither scalable nor sustainable in light of the surge of deepfakes and the exponentially growing volume of audio-visual electronic evidence in today's investigations. Moreover, manual analysis could lead to mistakes as humans have a psychological predisposition to believe in audio-visual content and a truth-default tendency.[177]

In light of the various tools and techniques to create deepfakes, there is currently no technological solution that is tailor-fit to effectively detect all types of digital forgeries. Ironically, AI itself might be the answer to deepfake detection. The most popular technique is to use algorithms similar to the ones that were used to build the deepfakes in question. The goal of this technique is to recognize the patterns showing how these fake images and audio clips were created, as well as to pick up subtle inconsistencies. AI-enabled tools like FaceForensics++ have already shown a detection accuracy rate that exceeds the capacity of manual analysis.[178]

Detection is, however, not without its challenges. In 2018, US researchers discovered that deepfake video faces did not blink normally. Since the majority of internet images show people with their eyes open, the algorithms were not trained to blink.[179] Unfortunately, as soon as this research was published, new deepfakes appeared with an improved capability to blink. Indeed, the detection of deepfakes becomes harder as weaknesses are revealed and the technology to fix these are also discovered. Therefore, the automated systems that were developed to examine videos for errors (such as irregular blinking, inconsistent lighting patterns,[180] or distinctive facial expressions) soon became obsolete.[181] Another weakness of AI detection systems is that they work best for celebrities since most of the training occurs on open-source databases, and there are considerably more images of celebrities online than ordinary individuals.

An alternative technique consists of applying blockchain to verify the source of the media. With this technology, only videos from trusted sources would be approved, decreasing the spread of possibly harmful deepfake media.[182] Tools that leverage the open-source Ethereum blockchain and cryptographic authentication in the form of hashes[183] or mobile apps to facilitate a trusted chain of custody of captured images or videos[184] have also been set forth. Another solution proposes to use the Ethereum blockchain-based approach to prove the authenticity of digital content as well as to identify the original source of digital content. It could also be used to assess if the digital content came from a reputable source.[185] Other provenance-checking solutions such as Truepic have also emerged.[186]

As for deepfake audio calls, the most archaic yet effective way to verify the caller is to hang up and call them back. Unless a very sophisticated hacking group who can reroute phone calls is behind the act, it can be relatively easy to detect fake phone calls.[187] As with the use of AI to detect deepfakes, the same can be applied to distinguish real from fake audio clips. A deep neural network detector uses visual representations of audio clips called spectrograms that are also used to train speech synthesis models. While to the unsuspecting ear these clips sound basically identical,* spectrograms of real audio versus fake audio have different representations.[188]

It is noteworthy that several international organizations, governments, universities, and tech firms are funding research to develop tools that can detect deepfakes. For instance, UNICRI organized a challenge at the Hackathon for Peace, Justice and Security in 2019 to challenge research teams around the globe to search for a solution to deepfake detection. Microsoft, Facebook, and Amazon also launched a Deepfake Detection Challenge in 2019 in the hope of accelerating the technology for identifying manipulated content.[189] In September 2020, Microsoft released the Microsoft Video Authenticator, a new deepfake detection tool that was tested using the datasets of the Deepfake Detection Challenge. Although not publicly released, the Authenticator analyzes content and provides the user with a confidence score as to whether or not the video might have been manipulated.[190]

Furthermore, the social media industry can also play a key role in improving deepfake detection by providing publicly available data to properly train AI-enabled detection tools. For instance, Facebook has released an open-source database of 100,000 deepfakes to train AI models to spot manipulated videos.[191]

As the degree of sophistication for generating deepfake technology continues to improve, it will ultimately take a combination of techniques to enable the detection of synthetic or fake videos and images.

Deepfake Policies

The malicious potential of highly realistic deepfake videos has set off alarm bells in the digital world. As a result, several large social platforms have revised their user policies to forbid users from uploading or sharing deepfakes.

* Rita Singh's *Profiling Humans from Their Voice* provides an extensive coverage of the forensic aspects of the human voice and the difficulties of "faking it" at the micro level.

After the scandal involving several famous female personalities that birthed the deepfake phenomenon, Reddit banned "fake porn" — deepfake photos and videos that superimpose a subject's face over an explicit photo or video without the subject's permission — in early 2018.[192] More recently, in January 2020, the platform updated its policies to prohibit the impersonation of an individual or entity in a misleading or deceptive manner, encompassing deepfakes.[193] The new policies were made just days after Facebook updated its own platform policies to ban deepfake videos. Facebook's new policy does not, however, cover videos manipulated for the sake of parody or satire, nor does it include shallowfakes such as the controversial Pelosi video, a fact that has prompted criticism from political leaders and digital experts.[194] Shortly after, Twitter also decided to flag manipulated videos, including deepfakes and shallowfakes, as well as to ban accounts "deceptively" sharing edited videos with the intent of causing harm.[195] Still, one problem associated with the policy-making principles for deepfakes is that over time, it will become more difficult to draw the line between what causes harm and what is lawful parody.

Meanwhile at the national level, there have been discussions about the regulation of deepfakes, predominantly on the basis of concerns that the technology could be abused to interfere with political processes.

For instance, the Canadian Communications Security Establishment released a report stating that deepfakes could be used to interfere in Canadian politics, particularly to discredit politicians and influence voters. Although there are no specific Canadian laws that forbid deepfakes, it has been suggested that existent causes of action might be applicable to address the wrongs committed by a person's abuse of deepfake technology, namely: "copyright infringement, defamation, violation of privacy, appropriation of personality, criminal code, human rights complaint, intentional infliction of mental suffering, and harassment."[196]

In November 2019, Chinese regulators announced new rules governing video and audio content online, including a ban on the publishing and distribution of false information or deepfakes online without proper disclosure that the post in question was created with AI or virtual reality (VR) technology. The Chinese government stated that beginning in 2020, failure to disclose the use of such technology to create a post will be considered a criminal offense.[197]

In the US, several bills that deal with the regulation of deepfakes have been introduced before federal lawmakers, although none of these bills has become law.[198, 199]

Recommendations and Considerations for Further Research

As solutions for deepfake detection evolve, the individuals and groups behind the abuse of deepfakes are equally expected to adapt their mod operandi with the aim of evading detection and training their models to follow counter-detection measures. Deepfakes can, in this regard, become a significant challenge to the current forensic audio-visual analysis and authentication techniques employed by industries, competent authorities, media professionals, and civil society.

It is necessary not only to make significant investments to develop new screening technology to detect tampering and irregularities in visual and audio content,[200] but also to ensure that such tools are kept up to date to reflect evolving deepfake creation technology, including current and possible future malicious abuse of such technology. Effective detection algorithms should also build upon existing solutions and digital forensic practices in photo, audio, and video analyses. In order to optimize the efforts and address the current gaps, the development of systems to combat deepfakes should be done in a collaborative manner between industry and end-users from competent authorities (such as police investigators, crime analysts, forensic experts and examiners, and other similar agents) and combined with training and capacity building for these competent authorities, content moderators, and other relevant professionals.

New policy and legislation measures should likewise be employed.[201] In particular, those policies should be technology-agnostic in order to be effective in the long run and to avoid having to review and replace these on a regular basis as the technology behind the creation and abuse of deepfakes evolves. Nevertheless, such measures should also avoid obstructing the positive applications of GANs.

Consideration should also be given to include the abuse of deepfakes in current and future initiatives at the international level to tackle illicit content online by focusing on public-private partnerships, notice-and-takedown procedures, proactive use of detection technology, and closer cooperation with competent authorities.[202]

Understanding the creation technology behind deepfakes at an early stage will significantly improve the opportunities for both the detection and development of effective countermeasures.[203, 204] Thus, performing a technology watch function is crucial. The role of the academic and research community is also vital, as there is a growing need for research on deepfakes. Industry and law enforcement professionals should also continuously develop forward-looking threat assessments on the current and future abuse of deepfakes by malicious actors. Better understanding of the threat can also inform potential policy and legislative measures to address the common challenges to investigating and countering deepfake abuse.

In spite of the increased popularity of deepfakes, research indicates that many people remain largely unaware of the existence of such digital forgeries.[205] Enhanced efforts to raise awareness on deepfakes and their associated threats are therefore crucial for empowering the public, different industries, and competent authorities to identify fake images, videos, and audio files. In the long run, heightened awareness will further contribute to limit the potential abuse of deepfakes.

Conclusion

This joint report by Trend Micro, UNICRI, and Europol was initiated by a simple question: "Has anyone witnessed any examples of criminals abusing artificial intelligence?" In the pursuit of answering this inquiry, the report managed to provide a collective understanding of current and future malicious uses and abuses of AI.

AI promises greater efficiency and higher levels of automation and autonomy. A subfield of computer science with many cross relationships with other disciplines, AI is intrinsically a dual-use technology at the heart of the so-called fourth industrial revolution. As a result of this duality, while it can bring enormous benefits to society and help solve some of the biggest challenges we currently face, AI could also enable a range of digital, physical, and political threats. Therefore, the risks and potential criminal abuse of AI systems need to be well-understood in order to protect not only society but also critical industries and infrastructures from malicious actors.

Based on available insights, research, and a structured open-source analysis, this report covered the present state of malicious uses and abuses of AI, including AI malware, AI-supported password guessing, and AI-aided encryption and social engineering attacks. It also described concrete future scenarios ranging from automated content generation and parsing, AI-aided reconnaissance, smart and connected technologies such as drones and autonomous cars, to AI-enabled stock market manipulation, as well as methods for AI-based detection and defense systems.

Using one of the most visible malicious uses of AI — the phenomenon of so-called deepfakes — the report further detailed a case study on the use of AI techniques to manipulate or generate visual and audio content that would be difficult for humans or even technological solutions to immediately distinguish from authentic ones.

As speculated on in this paper, criminals are likely to make use of AI to facilitate and improve their attacks by maximizing opportunities for profit within a shorter period, exploiting more victims, and creating new, innovative criminal business models — all the while reducing their chances of being caught. Consequently, as "AI-as-a-Service"[206] becomes more widespread, it will also lower the barrier to entry by reducing the skills and technical expertise required to facilitate attacks. In short, this further exacerbates the potential for AI to be abused by criminals and for it to become a driver of future crimes.

Although the attacks detailed here are mostly theoretical, crafted as proofs of concept at this stage, and although the use of AI to improve the effectiveness of malware is still in its infancy, it is plausible that malware developers are already using AI in more obfuscated ways without being detected by researchers and analysts. For instance, malware developers could already be relying on AI-based methods to bypass spam filters, escape the detection features of antivirus software, and frustrate the analysis of malware. In fact, DeepLocker, a tool recently introduced by IBM and discussed in this paper, already demonstrates these attack abilities that would be difficult for a defender to stop.

To add, AI could also enhance traditional hacking techniques by introducing new ways of performing attacks that would be difficult for humans to predict. These could include fully automated penetration testing, improved password-guessing methods, tools to break CAPTCHA security systems, or improved social engineering attacks. With respect to open-source tools providing such functionalities, the paper discussed some that have already been introduced, such as DeepHack, DeepExploit, and XEvil.

The widespread use of AI assistants, meanwhile, also creates opportunities for criminals who could exploit the presence of these assistants in households. For instance, criminals could break into a smart home by hijacking an automation system through exposed audio devices.

Given the discussions in this paper, it is safe to assume that cybercriminals will progressively integrate AI techniques to enhance the scope and scale of their attacks, thereby exploiting AI both as an attack vector and an attack surface, additionally powered by the previously mentioned service-based criminal business model. Consequently, the second part of the report explored plausible future scenarios in which criminals might abuse AI technologies to facilitate their activities and increase the success of their attacks.

For instance, it is expected that AI will help criminals conduct social engineering at scale by automating the first steps of an attack through AI-aided content generation, improving business intelligence gathering, and speeding up the detection rate of both potential victims and business process compromise attacks.

Moreover, as defense systems increasingly employ AI-based approaches, criminals are expected to either directly attack those systems, or use AI for user behavior emulation as well.

The increasing automation of everyday life's numerous aspects using AI-based systems also inevitably brings with it a possible loss of control over the same aspects. This would not only broaden the attack surface, but also create new types of attacks. One novel example is AI-enabled stock market manipulation that takes advantage of the use of AI-based algorithms in the area of high frequency trading.

Meanwhile, current and upcoming technologies like 5G, in combination with AI, will shape industry and further drive automation and smart technologies at scale. Presumably, criminals are likely to target or manipulate these technologies as well.

In response to the different scenarios described here, the report identified possible countermeasures, ways of mitigating risks, and several other recommendations.

It is also worth emphasizing that close cooperation with industry and academia is a must in order to develop a body of knowledge and raise awareness on the potential use and misuse of AI by criminals. Not only will such cooperation anticipate malicious and criminal activities facilitated by AI, it will also prevent, respond to, or mitigate the effects of these attacks in a proactive manner.

Lastly, despite this report's focus on its misuse, it is undeniable that AI holds a plethora of potential for positive applications — including critical support for investigating crimes of all types and the fulfilment of important projects by organizations with global impact. The capacity for these applications, however, is best explored in detail in another paper, and we look forward to reading and learning from the future work that will be contributed by both industry and academia about the endless possibilities that AI holds for the benefit of all.

Indeed, an understanding of the capabilities, scenarios, and attack vectors is key to enhancing preparedness, increasing resilience, and ensuring the positive use of AI to unlock the potential that it holds.

Appendix

YARA Rules for AI-Powered Malware Detection

```
rule tensorflow

{
    strings:
        $CopyrightString  = /Copyright 201\d The TensorFlow Authors\./
        $PathString1      = /tensorflow\/c\/c_api[^.]+\.h/
        $PathString2      = “tensorflow/core/framework/”
        $PathString3      = “tensorflow/core/lib/”
        $DomainString     = “tensorflow.org”

    condition:
        (is_py or is_pe or is_elf)
        and any of them
        and positives > 2
}

rule pytorch

{
    strings:
        $ImportString     = “import torch”
    condition:
        is_py and $ImportString and positives > 2

}

rule rapidminer {
    strings:
        $a = “rapidminer” ascii nocase
    condition:
        is_jar and $a and positives > 2

}
```

References

1 European Commission. (Apr. 25, 2018). *European Commission*. "Communication from the Commission to the European Parliament, the European Council, the European Economic and Social Committee, and the Committee of the Regions." Accessed Jul. 16, 2020, at https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe.

2 High-Level Expert Group on Artificial Intelligence (AI HLEG). (Apr. 8, 2019). *European Commission*. "Ethics Guidelines for Trustworthy AI." Accessed on Jul. 16, 2020, at https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top.

3 Joint Research Centre (JRC). (2020). *European Commission*. "AI Watch Defining Artificial Intelligence." Accessed on Jul. 16, 2020, at https://ec.europa.eu/jrc/en/publication/ai-watch-defining-artificial-intelligence.

4 Europol. (Jul. 18, 2019). *Europol*. "Do Criminals Dream of Electric Sheep?" Accessed on Oct. 14, 2020, at https://www.europol.europa.eu/newsroom/news/do-criminals-dream-of-electric-sheep-how-technology-shapes-future-of-crime-and-law-enforcement.

5 Europol. (Oct. 5, 2020). *Europol*. "Internet Organised Crime Threat Assessment (IOCTA) 2020." Accessed on Oct. 14, 2020, at https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2020.

6 High-Level Expert Group on Artificial Intelligence (AI HLEG). (Apr. 8, 2019). *European Commission*. "Ethics Guidelines for Trustworthy AI." Accessed on Jul. 16, 2020, at https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top.

7 Europol. (Feb. 28, 2017). *Europol*. "European Union Serious and Organised Crime Threat Assessment (SOCTA)." Accessed on Jul. 20, 2020, at https://www.europol.europa.eu/activities-services/main-reports/european-union-serious-and-organised-crime-threat-assessment-2017.

8 Europol. (Sep. 29, 2014). *Europol*. "The Internet Organised Crime Threat Assessment (IOCTA) 2014." Accessed on Jul. 20, 2020, at https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2014.

9 Europol. (Feb. 28, 2017). *Europol*. "European Union Serious and Organised Crime Threat Assessment (SOCTA)." Accessed on Jul. 20, 2020, at https://www.europol.europa.eu/activities-services/main-reports/european-union-serious-and-organised-crime-threat-assessment-2017.

10 Europol and Eurojust. (July 5, 2019). *Europol and Eurojust*. "Common Challenges in Combating Cybercrime." Accessed on Oct. 14, 2020, at https://www.europol.europa.eu/publications-documents/common-challenges-in-combating-cybercrime.

11 Europol. (Jul. 18, 2019). *Europol*. "Do Criminals Dream of Electric Sheep?" Accessed on Oct. 14, 2020, at https://www.europol.europa.eu/newsroom/news/do-criminals-dream-of-electric-sheep-how-technology-shapes-future-of-crime-and-law-enforcement.

12 Europol. (n.d.). *Europol*. "Joint Cybercrime Action Taskforce (J-CAT)." Accessed on Jul. 20, 2020, at https://www.europol.europa.eu/activities-services/services-support/joint-cybercrime-action-taskforce.

13 International Crime Court. (n.d.). *International Crime Court*. "About." Accessed on Jul. 20, 2020, at https://www.icc-cpi.int/about.

14 Europol. (n.d.). *Europol*. "EC3 Partners." Accessed on Jul. 20, 2020, at https://www.europol.europa.eu/about-europol/european-cybercrime-centre-ec3/ec3-partners.

15 European Commission. (Feb. 19, 2020). *European Commission*. "White Paper on Artificial Intelligence – a European approach to excellence and trust." Accessed on Oct. 1, 2020, at https://ec.europa.eu/digital-single-market/en/news/white-paper-artificial-intelligence-public-consultation-towards-european-approach-excellence.

16 Sean Palka and Damian McCoy. (n.d.). *Usenix*. "Fuzzin E-mail Filters with Generative Grammars and N-Gram Analysis." Accessed on Jul. 20, 2020, at https://www.usenix.org/system/files/conference/woot15/woot15-paper-palka.pdf.

17 Ankit Singh and Vijay Thaware. (Jul. 26, 2017). *Black Hat USA*. "Wire Me Through Machine Learning." Accessed on Jul. 20, 2020, at https://www.blackhat.com/us-17/briefings/schedule/#wire-me-through-machine-learning-6749.

18 Chanil Jeon, et al. (Jul. 27, 2019). *Black Hat USA*. "AVPASS: Leaking and Bypassing Antivirus Detection Model Automatically." Accessed on Jul. 20, 2020, at https://www.blackhat.com/us-17/briefings/schedule/#avpass-leaking-and-bypassing-antivirus-detection-model-automatically-7354.

19 Hyrum Anderson. (Jul. 27, 2017). *Black Hat USA*. "Bot vs. Bot for Evading Machine Learning Malware Detection." Accessed on Jul. 20, 2020, at https://www.blackhat.com/us-17/briefings/schedule/#bot-vs-bot-for-evading-machine-learning-malware-detection-6461.

20 Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. (May 1, 1996). *Journal of Artificial Intelligence Research*. "Reinforcement Learning: A Survey." Accessed on Oct. 13, 2020, at https://www.jair.org/index.php/jair/article/view/10166.

21 Dan Petro and Ben Morris. (2017). *DefCon*. “Weaponizing Machine Learning: Humanity was Overrated Anyway.” Accessed on Jul. 20, 2020, at https://www.defcon.org/html/defcon-25/dc-25-speakers.html#Petro.

22 Do Son. (Apr. 7, 2018). *Penetration Testing*. “Deep Exploit: Fully automatic penetration test tool using Machine Learning.” Accessed on Jul. 20, 2020, at https://securityonline.info/deep-exploit/.

23 Arthur Juliani. (Dec. 17, 2016). *Medium*. “Simple Reinforcement Learning with Tensorflow Part 8: Asynchronous Actor-Critic Agents (A3C).” Accessed on Jul. 20, 2020, at https://medium.com/emergent-future/simple-reinforcement-learning-with-tensorflow-part-8-asynchronous-actor-critic-agents-a3c-c88f72a5e9f2.

24 Dhilung Kirat, Jiyong Jang, and Marc Ph. Stoecklin. (Aug. 9, 2018). *Black Hat USA*. “DeepLocker — Concealing Targeted Attacks with AI Locksmithing.” Accessed on Jul. 20, 2020, at https://www.blackhat.com/us-18/briefings/schedule/index.html#deeplocker---concealing-targeted-attacks-with-ai-locksmithing-11549.

25 Virus Total. (n.d.). *VirusTotal*. “YARA — The pattern matching swiss knife for malware researchers (and everyone else).” Accessed on Jul. 20, 2020, at https://virustotal.github.io/yara/.

26 Stephen Hilt. (Aug. 27, 2017). *Trend Micro*. “The Sound of a Targeted Attack.” Accessed on Jul. 20, 2020, at https://documents.trendmicro.com/assets/pdf/The-Sound-of-a-Targeted-Attack.pdf.

27 Craig S. Smith. (May 10, 2018). *The New York Times*. “Alexa and Siri Can Hear This Hidden Command. You Can’t.” Accessed on Jul. 20, 2020, at https://www.nytimes.com/2018/05/10/technology/alexa-siri-hidden-command-audio-attacks.html.

28 Stephen Hilt, et al. (Mar. 5, 2019). *Trend Micro*. “Cybersecurity Risks in Complex IoT Environments: Threats to Smart Homes, Buildings and Other Structures.” Accessed on Jul. 20, 2020, at https://documents.trendmicro.com/assets/white_papers/wp-cybersecurity-risks-in-complex-iot-environments.pdf?_ga=2.20408120.1686390182.1575332885-267397288.1557129106.

29 Hashcat (n.d.). *Hashcat*. “Hashcat Advanced Password Recovery.” Accessed on Aug. 26, 2020, at https://hashcat.net/hashcat/.

30 GitHub. (n.d.) *GitHub, Inc*. “openwall/john.” Accessed on Aug. 26, 2020, at https://github.com/openwall/john.

31 Jason Brownlee. (Jul. 12, 2019). *Machine Learning Mastery*. “Impressive Applications of Generative Adversarial Networks (GANs).” Accessed on Jul. 20, 2020, at https://machinelearningmastery.com/impressive-applications-of-generative-adversarial-networks/.

32 Briland Hitaj, et al. (Feb. 14, 2019). *arXiv*. “PassGAN: A Deep Learning Approach for Password Guessing.” Accessed on Jul. 20, 2020, at https://arxiv.org/pdf/1709.00440.pdf.

33 Botmaster Labs. (n.d.). *Botmaster Labs*. “XEvil.” Accessed on Aug. 26, 2020, at https://xevil.net/en/.

34 Europol (Jan. 18, 2019). *Europol*. “First report of the observatory function on encryption.” Accessed on Jul. 20, 2020, at https://www.europol.europa.eu/publications-documents/first-report-of-observatory-function-encryption.

35 Martín Abadi and David G. Andersen. (Oct. 24, 2016). *arXiv*. “Learning to Protect Communications with Adversarial Neural Cryptography.” Accessed on Jul. 20, 2020, at https://arxiv.org/pdf/1610.06918.pdf.

36 Europol and Eurojust. (Feb. 18, 2020). *Europol and Eurojust*. “Second report of the observatory function on encryption.” Accessed on Oct. 15, 2020, at https://www.europol.europa.eu/publications-documents/second-report-of-observatory-function-encryption.

37 Criminal Use of Information Hiding (CUING). (n.d.). *Europol European Cybercrime Centre (EC3)*. “Steganography to cybercriminals exploitation.” Accessed on Jul. 20, 2020, at https://CUING.org/.

38 Europol and Eurojust. (Feb. 18, 2020). *Europol and Eurojust*. “Second report of the observatory function on encryption.” Accessed on Oct. 15, 2020, at https://www.europol.europa.eu/publications-documents/second-report-of-observatory-function-encryption.

39 Mayra Rosario Fuentes and Fernando Mercês. (Oct. 29, 2019). *Trend Micro*. “Cheats, Hacks, and Cyberattacks: Threats to the Esports Industry in 2019 and Beyond.” Accessed on Jul. 20, 2020, at https://documents.trendmicro.com/assets/white_papers/wp-threats-to-the-esports-industry-in-2019-and-beyond.pdf.

40 Jennifer Carole. (Dec. 21, 2018). *Bromium*. “Laundering Via In-Game Currency and Goods is on the Rise | Part 2.” Accessed on Jul. 20, 2020, at https://threatresearch.ext.hp.com/laundering-via-gaming-currency-and-goods-part-2/.

41 Zachary Baker. (Feb. 28, 2020). *Journal of Law and International Affairs*. "Gaming the System: Money Laundering through Microtransactions and In-Game Currencies." Accessed on Jul. 20, 2020, at https://sites.psu.edu/jlia/gaming-the-system-money-laundering-through-microtransactions-and-in-game-currencies/.

42 Anton Moiseienko and Kayla Izenman. (Oct. 11, 2019). *RUSI Registered Charity*. "Gaming the System: Money Laundering Through Online Games." Accessed on Jul. 20, 2020, at https://rusi.org/publication/rusi-newsbrief/gaming-system-money-laundering-through-online-games.

43 OliverAi. (Nov. 26, 2019). *Youtube*. "LeagueAI: An AI Playing League of Legends using Deep Learning." Accessed on Jul. 20, 2020, at https://www.youtube.com/watch?time_continue=92&v=iB4PoNJuXzc.

44 Europol. (Oct. 5, 2020). *Europol*. "Internet Organised Crime Threat Assessment (IOCTA) 2020." Accessed on Oct. 14, 2020, at https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2020.

45 GitHub. (n.d.) *GitHub, Inc*. "ThoughtfulDev/EagleEye." Accessed on Jul. 20, 2020, at https://github.com/ThoughtfulDev/EagleEye.

46 GitHub. (n.d.). *GitHub, Inc*. "CorentinJ/Real-Time-Voice-Cloning." Accessed on Jul. 20, 2020, at https://github.com/CorentinJ/Real-Time-Voice-Cloning.

47 Europol's European Cybercrime Centre (EC3). (2019). *2019 Predictions — The Year of Artificial Intelligence*. Europol.

48 Erik Zouave, et al. (Mar. 4, 2020). *Swedish Defence Research Agency (FOI)*. "Artificially Intelligent Cyberattacks." Accessed on Jul. 20, 2020, at https://www.statsvet.uu.se/digitalAssets/769/c_769530-l_3-k_rapport-foi-vt20.pdf.

49 Tom Simonite. (Jun. 11, 2020). *Wired*. "OpenAI's Text Generator Is Going Commercial." Accessed on Jul. 20, 2020, at https://www.wired.com/story/openai-text-generator-going-commercial/.

50 OpenAI. (n.d.). *OpenAI*. Accessed on Jul 10, 2020 at https://openai.com/.

51 ESET. (n.d.). *ESET*. "Can Artificial Intelligence Power Future Malware?" Accessed on Jul. 20, 2020, at https://www.eset.com/me/whitepapers/can-artificial-intelligence-power-future-malware/.

52 ESET. (n.d.). *ESET*. "Can Artificial Intelligence Power Future Malware?" Accessed on Jul. 20, 2020, at https://www.eset.com/me/whitepapers/can-artificial-intelligence-power-future-malware/.

53 Han Zhang, et al. (Dec. 10, 2016). *arXiv*. "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks." Accessed on Jul. 20, 2020, at https://arxiv.org/abs/1612.03242.

54 Jason Brownlee. (Jul. 12, 2019). *Machine Learning Mastery*. "Impressive Applications of Generative Adversarial Networks (GANs)." Accessed on Jul. 20, 2020, at https://machinelearningmastery.com/impressive-applications-of-generative-adversarial-networks/.

55 TypingDNA. (n.d.). *TypingDNA*. Accessed on Aug. 26, 2020, at https://www.typingdna.com/.

56 Aniello Castiglione, et al. (2014). *ScienceDirect*. "A botnet-based command and control approach relying on swarm intelligence." Accessed on Aug. 26, 2020, at https://doi.org/10.1016/j.jnca.2013.05.002.

57 Aniello Castiglione et al. (2014). *ScienceDirect*. "A botnet-based command and control approach relying on swarm intelligence." Accessed on Aug. 26, 2020, at https://doi.org/10.1016/j.jnca.2013.05.002.

58 European Cybercrime Centre (EC3). (May 7, 2015). *Europol*. "Cybercrime Dependencies Map." Accessed on Jul. 20, 2020, at https://www.europol.europa.eu/publications-documents/cybercrime-dependencies-map.

59 James Bridle. (Mar. 14, 2017). *James Bridle*. "Autonomous Trap 001." Accessed on Sep. 16, 2020, at https://jamesbridle.com/works/autonomous-trap-001.

60 James Bridle. (Mar. 14, 2017). *James Bridle*. "Autonomous Trap 001." Accessed on Sep. 16, 2020, at https://jamesbridle.com/works/autonomous-trap-001.

61 Katyanna Quach. (Feb. 20, 2020). *The Register*. "Researchers trick Tesla into massively breaking the speed limit by sticking a 2-inch piece of electrical tape on a sign." Accessed on Jul. 20, 2020, at https://www.theregister.com/2020/02/20/tesla_ai_tricked_85_mph/.

62 European Network and Information Security Agency (ENISA). (Nov. 25, 2019). *European Network and Information Security Agency (ENISA)*. "ENISA good practices for security of smart cars." Accessed on Jul. 20, 2020, at https://www.ENISA.europa.eu/publications/smart-cars.

63 SAE International (Jun. 15, 2018) *SAE International*. "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles." Accessed on Jul. 20, 2020, at https://www.sae.org/standards/content/j3016_201806/.

64 Evelyn Cheng. (Nov. 22, 2019). *CNBC*. "Self-driving trucks likely to hit the roads before passenger cars." Accessed on Jul. 20, 2020, at https://www.cnbc.com/2019/11/22/self-driving-trucks-likely-to-hit-the-roads-before-passenger-cars.html.

65 Bernard Marr. (Jun. 11, 2019). *Forbes*. "The Incredible Autonomous Ships of the Future: Run By Artificial Intelligence Rather than a Crew." Accessed on Jul. 20, 2020, at https://www.forbes.com/sites/bernardmarr/2019/06/05/the-incredible-autonomous-ships-of-the-future-run-by-artificial-intelligence-rather-than-a-crew/#2481dafe6fbf.

66 Tiffany Blake. (Ed.). (Feb. 11, 2020). *National Aeronautics and Space Administration (NASA)*. "What is Unmanned Aircraft Systems Traffic Management?" Accessed on Aug. 27, 2020, at https://www.nasa.gov/ames/utm/.

67 GSMA. (Nov. 23, 2018). *GSMA*. "Using Mobile Networks to Coordinate Unmanned Aircraft Traffic." Accessed on Jul. 20, 2020, at https://www.gsma.com/iot/wp-content/uploads/2018/11/Mobile-Networks-enabling-UTM-v5NG.pdf.

68 Lily Hay Newman. (Jun. 20, 2019). *Wired*. "The Global Hawk Drone Iran Shot Down Was a $220M Surveillance Monster." Accessed on Jul. 20, 2020, at https://www.wired.com/story/iran-global-hawk-drone-surveillance/.

69 Europol. (n.d.). *Europol*. "Sim Swapping — A Mobile Phone Scam." Accessed on Jul. 20, 2020, at https://www.europol.europa.eu/activities-services/public-awareness-and-prevention-guides/sim-swapping-%E2%80%93-mobile-phone-scam.

70 Frederic Lardinois. (Jun. 5, 2019). *TechCrunch*. "A first look at Amazon's new delivery drone." Accessed on Jul. 20, 2020, at https://techcrunch.com/2019/06/05/a-first-look-at-amazons-new-delivery-drone/.

71 Marco Mancini. (Jan. 8, 2015). *Hackster.io*. "Wifi Hacking Drone." Accessed on Oct. 26, 2020, at https://www.hackster.io/MakerMark/wifi-hacking-drone-045041.

72 European Union. (May 24, 2019). *EUR-Lex*. "Consolidated text: Commission Implementing Regulation (EU) 2019/947 of 24 May 2019 on the rules and procedures for the operation of unmanned aircraft (Text with EEA relevance)." Accessed on Aug. 27, 2020, at https://eur-lex.europa.eu/eli/reg_impl/2019/947/.

73 European Union. (Jun. 4, 2020). *EUR-Lex*. "Commission Implementing Regulation (EU) 2020/746 of 4 June 2020 amending Implementing Regulation (EU) 2019/947 as regards postponing dates of application of certain measures in the context of the COVID-19 pandemic (Text with EEA relevance)." Accessed on Aug. 27, 2020, at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32020R0746.

74 Dominik Gerstner and Dietrich Oberwittler. (Jan. 14, 2018). *SAGE Journals*. "Who's hanging out and what's happening? A look at the interplay between unstructured socializing, crime propensity and delinquent peers using social network data." Accessed on Jul. 20, 2020, at https://doi.org/10.1177/1477370817732194.

75 Thomas Brewster. (Jan. 29, 2020). *Forbes*. "Remember FindFace? The Russian Facial Recognition Company Just Turned On A Massive, Multimillion-Dollar Moscow Surveillance System." Accessed on Oct. 26, 2020, at https://www.forbes.com/sites/thomasbrewster/2020/01/29/findface-rolls-out-huge-facial-recognition-surveillance-in-moscow-russia/#4309c5a3463b.

76 Lauren M. Johnson. (Aug. 27, 2019). *CNN*. "A drone was caught on camera delivering contraband to an Ohio prison yard." Accessed on Jul. 20, 2020, at https://edition.cnn.com/2019/09/26/us/contraband-delivered-by-drone-trnd/index.html.

77 Nicole Hensley. (Jun. 18, 2020). *Daily News*. "Drones being deployed to South Africa mines equipped with lasers, paintball guns." Accessed on Jul. 20, 2020, at https://www.nydailynews.com/news/world/drones-deployed-south-africa-mines-article-1.1834649.

78 Sandrine Amiel. (May 3, 2019). *Euronews*. "Yellow vests, Black Blocs, casseurs: what's the difference? Euronews explains." Accessed on Aug. 27, 2020, at https://www.euronews.com/2019/05/02/yellow-vests-black-blocs-casseurs-what-s-the-difference-euronews-explains.

79 James Pasley. (Oct. 8, 2019). *Business Insider*. "12 US states and 7 countries that have barred protesters from wearing masks." Accessed on Aug. 27, 2020, at https://www.businessinsider.com/countries-states-where-protesters-cant-wear-masks-2019-10.

80 Ming Lee Newcomb. (Jul. 9, 2019). *Consequence of Sound*. "It turns out that Juggalo makeup blocks facial recognition technology." Accessed on Jul. 20, 2020, at https://consequenceofsound.net/2019/07/juggalo-makeup-facial-recognition/.

81 Geordie Gray. (Jul. 5, 2018). *Tone Deaf*. "Facial recognition technology doesn't recognise Juggalo Makeup." Accessed on Jul. 20, 2020, at https://tonedeaf.thebrag.com/facial-recognition-technology-juggalo/.

82 Ian O'Neill. (Jul. 2, 2018). *Twitter*. "for anyone wondering why some face changes evade facial recognition and others don't, here's a visualization of how landmarks are placed on a few examples." Accessed on Sep. 22, 2020, at https://twitter.com/tahkion/status/1013568373622362113.

83 Ian O'Neill. (Jul. 2, 2018). *Twitter*. "for anyone wondering why some face changes evade facial recognition and others don't, here's a visualization of how landmarks are placed on a few examples." Accessed Sep. 22, 2020, at https://twitter.com/tahkion/status/1013568373622362113.

84 David Cardinal. (Sep. 14, 2017). *ExtremeTech*. "How Apple's iPhone X TrueDepth Camera Works." Accessed on Sep. 4, 2020, at https://www.extremetech.com/mobile/255771-apple-iphone-x-truedepth-camera-works.

85 theworacle. (Jul. 16, 2009). *Youtube*. "US Air Force Flapping Wing Micro Air Vehicle." Accessed on Jul. 20, 2020, at https://www.youtube.com/watch?list=RDCMUCZDh1CUNrvzefaLU0n6qZoA&v=_5YkQ9w3PJ4&feature=emb_rel_end.

86 Cloudera. (n.d.). *Cloudera*. Accessed on Jul. 10, 2020, at https://www.cloudera.com.

87 Tirthajyoti Sarkar. (Oct. 26, 2018). *Towards Data Science*. "What Is Benford's Law and why is it important for data science?" Accessed on Jul. 20, 2020, at https://towardsdatascience.com/what-is-benfords-law-and-why-is-it-important-for-data-science-312cb8b61048.

88 Algorithm&Blues. (Oct. 15, 2010). *Algorithm&Blues*. "Update: Day Traders Crack The Timber Hill Trading System." Accessed on Jul. 20, 2020, at https://algosandblues.wordpress.com/2010/08/17/norwegian-day-traders-crack-the-timber-hill-trading-system/.

89 Manceps. (Jul. 14, 2020). *Manceps*. "Could an Adversarial Bot Manipulate the Stock Market?" Accessed on Jul. 20, 2020, at https://www.manceps.com/articles/experimebnts/beat-the-bots.

90 Simon Weckert. (n.d.). *Simon Weckert*. "Google Maps Hacks." Accessed on Jul. 20, 2020, at http://www.simonweckert.com/googlemapshacks.html.

91 Moritz Ahlert. (2019). *Hamburger Journal für Kulturanthropologie*. "The Power of Virtual Maps." Accessed on Sep. 15, 2020, at https://journals.sub.uni-hamburg.de/hjk/article/view/1395.

92 Simon Weckert. (n.d.). *Simon Weckert*. "Google Maps Hacks." Accessed on Jul. 20, 2020, at http://www.simonweckert.com/googlemapshacks.html.

93 Strategy-Stocks. (n.d.). *Strategy-Stocks*. "What is micro-trading?" Accessed on Aug. 27, 2020, at https://www.strategystocks.co.uk/micro-trading.html.

94 CFM. (Jul. 5, 2018). *CFM*. "Artificial Intelligence: perspectives from the quant coalface." Accessed on Sep. 2, 2020, at https://www.cfm.fr/insights/artificial-intelligence-perspectives-from-the-quant-coalface/.

95 International Monetary Fund (IMF). (Sep. 2016). *International Monetary Fund (IMF)*. "Detailed Assessment Report on Anti-Money Laundering and Combating the Financing of Terrorism." Accessed on Jul. 20, 2020, at https://www.imf.org/external/pubs/ft/scr/2016/cr16294.pdf.

96 Europol. (n.d.) *Europol*. "WannaCry Ransomware." Accessed on Aug. 26, 2020, at https://www.europol.europa.eu/wannacry-ransomware.

97 Europol. (Jun. 28, 2017). *Europol*. "New Wave of Ransomware Affecting Businesses: What to Do?" Accessed on Aug. 26, 2020, at https://www.europol.europa.eu/newsroom/news/new-wave-of-ransomware-affecting-businesses-what-to-do.

98 Erik Zouave, et al. (Mar. 4, 2020). *Swedish Defence Research Agency (FOI)*. "Artificially intelligent cyberattacks." Accessed on Jul. 20, 2020, at https://www.statsvet.uu.se/digitalAssets/769/c_769530-l_3-k_rapport-foi-vt20.pdf.

99 FileCloud. (Aug. 27, 2018). *FileCloud*. "Machine Vs Machine: A Look at AI-Powered Ransomware." Accessed on Jul. 20, 2020, at https://www.getfilecloud.com/blog/2018/08/machine-vs-machine-a-look-at-ai-powered-ransomware/#.XuIHkkVLiUk.

100 Harshajit Sarmah. (Aug. 22, 2019). *Analytics India Magazine*. "What If Artificial Intelligence Becomes Ransomware's Sidekick?" Accessed on Jul. 20, 2020, at https://analyticsindiamag.com/what-if-artificial-intelligence-becomes-ransomwares-sidekick/.

101 ESET. (n.d.). *ESET*. "Can Artificial Intelligence Power Future Malware?" Accessed on Jul. 20, 2020, at https://www.eset.com/me/whitepapers/can-artificial-intelligence-power-future-malware/.

102 Scott Ferguson. (Apr. 28, 2020). *Dark Reading*. "Ransomware Attacks Target Public & Government Orgs with More Frequency, Ferocity." Accessed on Jul. 20, 2020, at https://www.darkreading.com/application-security/ransomware/ransomware-attacks-target-public-and-government-orgs-with-more-frequency-ferocity/a/d-id/746811.

103 Theo Douglas. (Apr. 11, 2018). *Government Technology*. "Phishing, Malware, Ransomware Among Top Public-Sector Threats, Reports Find." Accessed on Jul. 20, 2020, at https://www.govtech.com/security/Phishing-Malware-Ransomware-Among-Top-Public-Sector-Threats-Reports-Find.html.

104 David Canellos. (Apr. 23, 2020). *Threatpost*. "Public Sector Ransomware Attacks Rage On: Can Your Organization Repel Them?" Accessed on Jul. 20, 2020, at https://threatpost.com/public-sector-ransomware-attacks-rage/155086/.

105 Bill Siegel. (Nov. 1, 2019). *Security Boulevard*. “Ransomware Payments Rise as Public Sector is Targeted, New Variants Enter the Market.” Accessed on Jul. 20, 2020, at https://securityboulevard.com/2019/11/ransomware-payments-rise-as-public-sector-is-targeted-new-variants-enter-the-market/.

106 Trend Micro. (Apr. 26, 2017). *Trend Micro Security Intelligence Blog*. “Machine Learning and the Fight Against Ransomware.” Accessed on Jul. 20, 2020, at https://blog.trendmicro.com/machine-learning-and-the-fight-against-ransomware/.

107 ESET. (n.d.). *ESET*. “Can Artificial Intelligence Power Future Malware?” Accessed on Jul. 20, 2020, at https://www.eset.com/me/whitepapers/can-artificial-intelligence-power-future-malware/.

108 Pindrop. (n.d.). *Pindrop Security, Inc*. “Call Center Fraud Vectors & Fraudsters Analyzed: Summer Break Edition.” Accessed on Aug. 27, 2020, at https://www.pindrop.com/resources/video/webinar/call-center-fraud-vectors-fraudsters-analyzed-summer-break-edition/.

109 Rotem Shemesh. (Nov. 21, 2016). *NICE*. “4 Things You Must Have to Ensure Effective Voice Biometrics in a Contact Center.” Accessed on Aug. 27, 2020, at https://www.nice.com/engage/blog/4-things-you-must-have-to-ensure-effective-voice-biometrics-in-a-contact-center-2224/.

110 Descript. (n.d.). *Descript*. “Lyrebird AI.” Accessed on Aug. 27, 2020, at https://www.descript.com/lyrebird.

111 Nuance. (n.d.). *Nuance Communications, Inc*. “Index.” Accessed on Aug. 27, 2020, at https://www.nuance.com/index.html.

112 AI for Good Summit. (n.d.) *International Telecommunication Union*. “AI for Good Summit.” Accessed on Sep. 24, 2020, at https://aiforgood.itu.int/.

113 Europol. (Oct. 24 – 25, 2019). *Europol*. “3rd ENISA — Europol IoT Security Conference.” Accessed on Sep. 24, 2020, at https://www.europol.europa.eu/events/3rd-ENISA-europol-iot-security-conference.

114 Europol. (Oct. 11, 2019). *Europol*. “Fighting Cybercrime in a Connected Future.” Accessed on Sep. 24, 2020, at https://www.europol.europa.eu/newsroom/news/fighting-cybercrime-in-connected-future.

115 United Nations Interregional Crime and Justice Research Institute (UNICRI). (Jul. 3, 2019). *United Nations Interregional Crime and Justice Research Institute (UNICRI)*. “2nd INTERPOL – UNICRI Global Meeting on Artificial Intelligence for Law Enforcement.” Accessed on Sep. 24, 2020, at http://www.unicri.it/news/article/ai_UNICRI_INTERPOL_law_enforcement.

116 European Commission. (Jun. 2018). *European Commission*. “Ethics Guidelines for Trustworthy AI.” Accessed on Sep. 24, 2020, at https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top.

117 European Telecommunications Standards Institute (ETSI). (n.d.). *European Telecommunications Standards Institute (ETSI)*. “Industry Specification Group (ISG) Securing Artificial Intelligence (SAI).” Accessed on Sep. 24, 2020, at https://www.etsi.org/committee/1640-sai.

118 European Network and Information Security Agency (ENISA). (n.d.). *European Network and Information Security Agency (ENISA)*. “Ad-Hoc Working Group on Artificial Intelligence Cybersecurity.” Accessed on Sep. 24, 2020, at https://www.ENISA.europa.eu/topics/iot-and-smart-infrastructures/artificial_intelligence/adhoc_wg_calls/.

119 Europol. (Sep. 25, 2013). *Europol*. “Facing Future Cyber Threats.” Accessed on Sep. 24, 2020, at https://www.europol.europa.eu/newsroom/news/facing-future-cyber-threats.

120 Tonya Riley. (Jun. 13, 2019). *The Washington Post*. “The Technology 202: It's time for Congress to address deepfakes, experts say.” Accessed on Jun. 19, 2020, at https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2019/06/13/the-technology-202-it-s-time-for-congress-to-address-deepfakes-experts-say/5d01687aa7a0a4586bb2da85/.

121 Donie O’Sullivan. (Aug. 10, 2019). *CNN Business*. “The Democratic Party deepfaked its own chairman to highlight 2020 concerns.” Accessed on Jun. 19, 2020, at https://edition.cnn.com/2019/08/09/tech/deepfake-tom-perez-dnc-defcon/index.html.

122 Oscar Schwartz. (Nov. 12, 2018). *The Guardian*. “You thought fake news was bad? Deep fakes are where truth goes to die.” Accessed on Jun. 19, 2020, at https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth.

123 Haya R. Hasan and Khaled Salah. (Mar. 18, 2019). *IEEE Xplore*. “Combating Deepfake Videos Using Blockchain and Smart Contracts.” Accessed on Jun. 19, 2020, at https://doi.org/10.1109/ACCESS.2019.2905689.

124 Centre for Data Ethics and Innovation. (Sep. 2019). *Centre for Data Ethics and Innovation*. “Deepfakes and Audio-visual Disinformation.” Accessed on Jul. 20, 2020, at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831179/Snapshot_Paper_-_Deepfakes_and_Audiovisual_Disinformation.pdf.

125 Ian J. Goodfellow, et al. (Jun. 10, 2014). *arXiv*. “Generative Adversarial Networks.” Accessed on Jul. 1, 2020, at https://arxiv.org/abs/1406.2661.

126 Oscar Schwartz. (Nov. 12, 2018). *The Guardian*. “You thought fake news was bad? Deep fakes are where truth goes to die.” Accessed on Jun. 19, 2020, at https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth.

127 Joe Andrews. (Jul. 12, 2019). *CNBC*. “Fake news is real — AI is going to make it much worse.” Accessed on Jun 26, 2020, at https://www.cnbc.com/2019/07/12/fake-news-is-real-ai-is-going-to-make-it-much-worse.html.

128 Than Thi Nguyen, et al. (Jul. 28, 2020). *arXiv*. “Deep Learning for Deepfakes Creation and Detection: A Survey.” Accessed on Jul 3, 2020, at https://arxiv.org/abs/1909.11573.

129 Than Thi Nguyen, et al. (Jul. 28, 2020). *arXiv*. “Deep Learning for Deepfakes Creation and Detection: A Survey.” Accessed on Jul 3, 2020, at https://arxiv.org/abs/1909.11573.

130 Samantha Cole. (Jan. 24, 2020). *Vice*. “New Deepfake Method Can Put Words in Anyone's Mouth.” Accessed on Jun. 27, 2020, at https://www.vice.com/en_us/article/g5xvk7/researchers-created-a-way-to-make-realistic-deepfakes-from-audio-clips.

131 Henry Ajder. (Sep. 2019). *Deeptrace*. “The State of Deepfakes: Landscape, Threats, and Impact.” Accessed on Jun. 20, 2020, at https://deeptracelabs.com/archive/.

132 Campbell Kwan. (May 23, 2019). *ZDNet*. “Samsung uses AI to transform photos into talking head videos.” Accessed on Jun. 23, 2020, at https://www.zdnet.com/article/samsung-uses-ai-to-transform-photos-into-talking-head-videos/.

133 Adi Robertson. (2018, Feb 11). *The Verge*. “I’m using AI to face-Swap Elon Musk and Jeff Bezos, and I'm really bad at it.” Accessed on Jun. 29, 2020, at https://www.theverge.com/2018/2/11/16992986/fakeapp-deepfakes-ai-face-swapping.

134 Ashley Carman. (Jul. 17, 2019). *The Verge*. “FaceApp is back and so are privacy concerns.” Accessed on Jun 20, 2020, at https://www.theverge.com/2019/7/17/20697771/faceapp-privacy-concerns-ios-android-old-age-filter-russia.

135 Agence France-Presse in Shanghai. (Sep. 2, 2019). *The Guardian*. “Chinese deepfake app Zao sparks privacy row after going viral.” Accessed on Jun. 24, 2020, at https://www.theguardian.com/technology/2019/sep/02/chinese-face-swap-app-zao-triggers-privacy-fears-viral.

136 Descript. (n.d.). *Descript*. “Overdub.” Accessed on Jul. 20, 2020, at https://www.descript.com/overdub?lyrebird=true.

137 Haya R. Hasan and Khaled Salah. (Mar. 18, 2019). *IEEE Xplore*. “Combating Deepfake Videos Using Blockchain and Smart Contracts.” Accessed on Jun. 19, 2020, at https://doi.org/10.1109/ACCESS.2019.2905689.

138 Aja Romano (Apr. 18, 2018). *Vox*. “Jordan Peele's simulated Obama PSA is a double-edged warning against fake news." Accessed on Jun. 23, 2020, at https://www.vox.com/2018/4/18/17252410/jordan-peele-obama-deepfake-buzzfeed.

139 Samantha Cole. (Jun. 12, 2019). *Vice*. “This Deepfake of Mark Zuckerberg Tests Facebook’s Fake Video Policies.” Accessed on Jul. 3, 2020, at https://www.vice.com/en_us/article/ywyxex/deepfake-of-mark-zuckerberg-facebook-fake-video-policy.

140 Nic Ker. (Jun. 12, 2019). *Malay Mail*. “Is the political vide Viral sex video confession real or a Deepfake?” Accessed on Jul. 1, 2020, at https://www.malaymail.com/news/malaysia/2019/06/12/is-the-political-aide-viral-sex-video-confession-real-or-a-deepfake/1761422.

141 Sarah Cahlan. (Feb. 13, 2020). *The Washington Post*. “How misinformation helped spark an attempted coup in Gabon.” Accessed on Jul. 1, 2020, at https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/.

142 Angelin Yeoh. (Dec. 12, 2019). *AsiaOne*. “Malaysian actor Zul Ariffin makes ‘deepfake’ claims after sex video goes viral on social media.” Accessed on Jul. 1, 2020, at https://www.asiaone.com/digital/malaysian-actor-zul-ariffin-makes-deepfake-claims-after-sex-video-goes-viral-social-media.

143 Oscar Schwartz. (Nov. 12, 2018). *The Guardian*. “You thought fake news was bad? Deep fakes are where truth goes to die.” Accessed on Jun. 20, 2020, at https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth.

144 Beatrice Dupuy (May 24, 2019). *AP News*. “NOT REAL NEWS: Altered video makes Pelosi seem to slur words.” Accessed on Jun. 22, 2020, at https://apnews.com/4841d0ebcc704524a38b1c8e213764d0.

145 Raphael Satter. (Jun. 13, 2019). *ABC News Network*. “Experts: Spy used AI-generated face to connect with targets.” Accessed on Jun. 26, 2020, at https://abcnews.go.com/Technology/wireStory/experts-spy-ai-generated-face-connect-targets-63674174.

146 Jesse Damiani (Nov. 9, 2019). *Forbes*. “A Voice Deepfake Was Used To Scam A CEO Out Of $243,000.” Accessed on Jun. 25, 2020, at https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/#5b10b1802241.

147 Kaveh Waddell and Jennifer A. Kingson. (Jul. 19, 2019). *Axios*. “The coming of deepfakes threat to businesses.” Accessed on Jul. 1, 2020, at https://www.axios.com/the-coming-deepfakes-threat-to-businesses-308432e8-f1d8-465e-b628-07498a7c1e2a.html?utm_source=twitter&utm_medium=social&utm_campaign=organic.

148 Lorenzo Franceschi-Bicchierai. (Jul. 24, 2020). *Vice*. “Listen to This Deepfake Audio Impersonating a CEO in Brazen Fraud Attempt.” Accessed on Jul. 3, 2020, at https://www.vice.com/en_us/article/pkyqvb/deepfake-audio-impersonating-ceo-fraud-attempt.

149 Maeve Sheehan. (Feb. 16, 2020). *Independent.ie*. “Cyber scam stole voice of Europol's ex-boss to get cash.” Accessed on Aug. 26, 2020, at https://www.independent.ie/irish-news/cyber-scam-stole-voice-of-europols-ex-boss-to-get-cash-38960406.html.

150 Gabriela Swerling. (Jan. 31, 2020). *The Telegraph*. “Doctored audio evidence used to damn father in custody battle.” Accessed on Jul. 3, 2020, at https://www.telegraph.co.uk/news/2020/01/31/deepfake-audio-used-custody-battle-lawyer-reveals-doctored-evidence/.

151 Naser Damer, PhD, et al. (Oct. 2018). *IEEE Xplore*. “MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network.” Accessed on Jul. 20, 2020, at https://doi.org/10.1109/BTAS.2018.8698563.

152 Arjo Solutions. “Digital technologies: the solution for breeder documents protection?” (Sep. 12, 2017). Arjo Solutions. Accessed on Aug. 26, 2020, at https://www.arjo-solutions.com/en/2017/09/12/breeder-documents-protection/.

153 University of Lincoln. (Aug. 1, 2019). *ScienceDaily*. “Two fraudsters, one passport: Computers more accurate than humans at detecting fraudulent identity photos.” Accessed on Jul. 20, 2020, at www.sciencedaily.com/releases/2019/08/190801104038.htm.

154 Naser Damer, PhD. (n.d.). *Fraunhofer IGD*. “Face morphing: a new threat?” Accessed on Jul. 20, 2020, at https://www.igd.fraunhofer.de/en/press/annual-reports/2018/face-morphing-a-new-threat.

155 David J. Robertson, et al. (Jun. 27, 2018). *Springer Nature*. “Detecting morphed passport photos: a training and individual differences approach.” Accessed on Jul. 20, 2020, at https://cognitiveresearchjournal.springeropen.com/articles/10.1186/s41235-018-0113-8.

156 Robin S.S. Kramer, et al. (Jul. 29, 2019). *Springer Nature*. “Face morphing attacks: Investigating detection with humans and computers.” Accessed on Jul. 20, 2020, at https://link.springer.com/article/10.1186/s41235-019-0181-4?error=cookies_not_supported&code=bffbdaef-92aa-47c2-b4c1-37a7a51dd341.

157 David J. Robertson, et al. (Jun. 27, 2018). *Springer Nature*. “Detecting morphed passport photos: a training and individual differences approach.” Accessed on Jul. 20, 2020, at https://cognitiveresearchjournal.springeropen.com/articles/10.1186/s41235-018-0113-8.

158 Matteo Ferrara, Annalisa Franco, and Davide Maltoni. (Dec. 2014). *ResearchGate*. “The magic passport.” Accessed on Jul. 20, 2020, at https://www.researchgate.net/publication/283473746_The_magic_passport.

159 Giorgio Patrini. (Oct. 7, 2019) *DeepTrace*. “Mapping the Deepfake Landscape.” Accessed on Aug. 26, 2020, at https://deeptracelabs.com/mapping-the-deepfake-landscape/.

160 Europol, Trend Micro, and UNICRI. (Mar. 3, 2020). Criminal Misuse of AI and Deepfakes, The Hague, The Netherlands.

161 Europol European Cybercrime Centre (EC3). (2019). *2019 Predictions — The Year of Artificial Intelligence*. Europol.

162 Europol. (Oct. 9, 2019). *Europol*. “Internet Organised Crime Threat Assessment (IOCTA).” Accessed on Jul. 20, 2020, at https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2019.

163 Europol. (Jul. 18, 2019). *Europol*. “Do Criminals Dream of Electric Sheep?” Accessed on Oct. 14, 2020, at https://www.europol.europa.eu/newsroom/news/do-criminals-dream-of-electric-sheep-how-technology-shapes-future-of-crime-and-law-enforcement.

164 United Nations Interregional Crime and Justice Research Institute (UNICRI). (Apr. 2, 2019). *United Nations Interregional Crime and Justice Research Institute (UNICRI)*. “High-Level Event: Artificial Intelligence and Robotics — Reshaping the Future of Crime, Terrorism and Security.” Accessed on Aug. 26, 2020, at http://www.UNICRI.it/news/article/AI_Robotics_Crime_Terrorism_Security.

165 Hany Farid. (Jun. 18, 2019). *Fox News*. “Hany Farid: Deepfakes give new meaning to the concept of ‘fake news,’ and they're here to stay.” Accessed on Jun. 20, 2020, at https://www.foxnews.com/opinion/hany-farid-deep-fakes.

166 Tom Taulli. (Jun. 15, 2019). *Forbes*. “Deepfake: What You Need To Know.” Accessed on Jun. 20, 2020, at https://www.forbes.com/sites/tomtaulli/2019/06/15/deepfake-what-you-need-to-know/#52981563704d.

167 Drew Harwell. (Jun. 12, 2019). *The Washington Post*. “Top AI researchers race to detect ‘deepfake’ videos: ‘We are outgunned.’” Accessed on Jun. 20, 2020, at https://www.washingtonpost.com/technology/2019/06/12/top-ai-researchers-race-detect-deepfake-videos-we-are-outgunned/.

168 Mayra Rosario Fuentes. (May 26, 2020). *Trend Micro*. “Shifts in Underground Markets: Past, Present, and Future.” Accessed on Jul. 20, 2020, at https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/trading-in-the-dark.

169 Europol. (Jun. 14, 2018). *Europol*. “French Coder Who Helped Extort British Company Arrested in Thailand.” Accessed on Jul. 20, 2020, at https://www.europol.europa.eu/newsroom/news/french-coder-who-helped-extort-british-company-arrested-in-thailand.

170 Hany Farid. (Jun. 18, 2019). *Fox News*. “Hany Farid: Deepfakes give new meaning to the concept of ‘fake news,’ and they're here to stay.” Accessed on Jun. 20, 2020, at https://www.foxnews.com/opinion/hany-farid-deep-fakes.

171 Europol. (Jun. 19, 2017). *Europol*. “Online Sexual Coercion and Extortion as a Form of Crime Affecting Children: Law Enforcement Perspective.” Accessed on Jul. 20, 2020, at https://www.europol.europa.eu/publications-documents/online-sexual-coercion-and-extortion-form-of-crime-affecting-children-law-enforcement-perspective.

172 Europol. (Oct. 9, 2019). *Europol*. “Internet Organised Crime Threat Assessment (IOCTA).” Accessed on Jul. 20, 2020, at https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2019.

173 Europol. (Oct. 9, 2019). *Europol*. “Internet Organised Crime Threat Assessment (IOCTA).” Accessed on Jul. 20, 2020, at https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2019.

174 Laura Hautala. (Feb. 11, 2018). *InformationSecurity.Report*. “If you like deepfakes, you might be mining cryptocurrency.” Accessed on Jun. 23, 2020, at https://informationsecurity.report/news/if-you-like-deepfakes-you-might-be-mining-cryptocurrency/4825.

175 Mayra Rosario Fuentes. (May 26, 2020). *Trend Micro*. “Shifts in Underground Markets: Past, Present, and Future.” Accessed on Jul. 20, 2020, at https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/trading-in-the-dark.

176 Europol. (Oct. 9, 2019). *Europol*. “Internet Organised Crime Threat Assessment (IOCTA).” Accessed on Jul. 20, 2020, at https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2019.

177 Timothy Levine. (Aug. 11, 2014). *SAGE Publications*. “Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection.” Accessed on Aug. 26, 2020, at https://www.researchgate.net/publication/273593306_Truth-Default_Theory_TDT_A_Theory_of_Human_Deception_and_Deception_Detection.

178 Andreas Rössler, et al. (Aug. 26, 2019). *arXiv*. “FaceForensics++, Learning to Detect Manipulated Face Images.” Accessed on Aug. 26, 2020, at https://arxiv.org/abs/1901.08971v3.

179 Yuezun Li, Ming-Ching Chang, and Siwei Lyu. (Jun. 11, 2018). *arXiV*. “In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking.” Accessed on Aug. 26, 2020, at https://arxiv.org/abs/1806.02877.

180 Kara Manke. (Jun. 18, 2019). *Berkeley News*. “Researchers use facial quirks to unmask ‘deepfakes’.” Accessed on Aug. 26, 2020, at https://news.berkeley.edu/2019/06/18/researchers-use-facial-quirks-to-unmask-deepfakes/.

181 Jim Brenner. (Nov. 25, 2019). *iProov*. “The Deepfake Age: Who Gatekeeps Digital Trust?" Accessed on Aug. 26, 2020, https://www.iproov.com/newsroom/blog/the-deepfake-age-who-gatekeeps-digital-trust.

182 Antonio García Martínez. (Mar. 26, 2018). *Wired*. “The Blockchain Solution to Our Deepfake Problems.” Accessed on Jul. 20, 2020, at https://www.wired.com/story/the-blockchain-solution-to-our-deepfake-problems/.

183 Lily Hay Newman. (Feb. 11, 2019). *Wired*. “A New Tool Protects Videos From Deepfakes and Tampering.” Accessed on Jul. 20, 2020, at https://www.wired.com/story/amber-authenticate-video-validation-blockchain-tampering-deepfakes/.

184 eyeWitness. (n.d.). *eyeWitness*. “What we do.” Accessed on Jul. 20, 2020, at https://www.eyewitness.global/our-work.html.

185 Haya R. Hasan and Khaled Salah. (Mar. 18, 2019). *IEEE Xplore*. “Combating Deepfake Videos Using Blockchain and Smart Contracts.” Accessed on Jun. 19, 2020, at https://doi.org/10.1109/ACCESS.2019.2905689.

186 Centre for Data Ethics and Innovation. (Sep. 2019). *Centre for Data Ethics and Innovation*. “Deepfakes and Audio-visual Disinformation.” Accessed on Jul. 20, 2020, at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831179/Snapshot_Paper_-_Deepfakes_and_Audiovisual_Disinformation.pdf.

187 Kim Lyons. (Jan. 29, 2020). *The Verge*. "FTC says the tech behind audio deepfakes is getting better." Accessed on Aug. 26, 2020, at https://www.theverge.com/2020/1/29/21080553/ftc-deepfakes-audio-cloning-joe-rogan-phone-scams.

188 Dessa News. (Sep. 28, 2019). *Medium*. "Detecting Audio Deepfakes With AI." Accessed on Aug. 26, 2020, at https://medium.com/dessa-news/detecting-audio-deepfakes-f2edfd8e2b35.

189 Facebook AI. (Jun. 25, 2020). *Facebook AI*. "Deepfake Detection Challenge Dataset." Accessed on Aug. 26, 2020, at https://deepfakedetectionchallenge.ai.

190 Ivan Mehta. (Sep. 2, 2020). *The Next Web*. "Microsoft's new deepfake detection tool rates bogus videos with a confidence score." Accessed on Sep. 8, 2020, at https://thenextweb.com/neural/2020/09/02/microsofts-new-deepfake-detection-tool-rates-bogus-videos-with-a-confidence-score/.

191 Will Douglas Heaven. (Jun. 12, 2020). *MIT Technology Review*. "Facebook just released a database of 100,000 deepfakes to teach AI how to spot them." Accessed on Aug. 26, 2020, at https://www.technologyreview.com/2020/06/12/1003475/facebooks-deepfake-detection-challenge-neural-network-ai/.

192 Leo Kelion. (Feb. 7, 2018). *BBC News*. "Reddit bans deepfake porn videos." Accessed on Jun. 27, 2020, at https://www.bbc.com/news/technology-42984127.

193 Jay Peters. (Jan. 9, 2020). *The Verge*. "Reddit bans impersonation on its platform." Accessed on Jun. 27, 2020, at https://www.theverge.com/2020/1/9/21058803/reddit-account-ban-impersonation-policy-deepfakes-satire-rules.

194 Tony Romm, Drew Harwell, and Isaac Stanley-Becker. (Jan. 8, 2020). *The Washington Post*. "Facebook bans deepfakes, but new policy may not cover controversial Pelosi video." Accessed on Jun. 27, 2020, at https://www.washingtonpost.com/technology/2020/01/06/facebook-ban-deepfakes-sources-say-new-policy-may-not-cover-controversial-pelosi-video/.

195 Margi Murphy. (Feb. 4, 2020). *The Telegraph*. "Twitter bans 'deepfakes' and 'cheap fakes.'" Accessed on Jun. 27, 2020, at https://www.telegraph.co.uk/technology/2020/02/04/twitter-bans-deepfakes-cheap-fakes/.

196 Pablo Tseng. (Mar. 2018). *McMillan*. "What Can The Law Do About 'Deepfake'?" Accessed on Jun. 27, 2020, at https://mcmillan.ca/What-Can-The-Law-Do-About-Deepfake.

197 Nick Statt. (Nov. 29, 2019). *The Verge*. "China makes it a criminal offense to publish deepfakes or fake news without disclosure." Accessed on Jun. 27, 2020, at https://www.theverge.com/2019/11/29/20988363/china-deepfakes-ban-internet-rules-fake-news-disclosure-virtual-reality.

198 Ben Sasse. (Dec. 21, 2018). *Library of Congress*. "S.3805 – Malicious Deep Fake Prohibition Act of 2018." Accessed on Jun. 27, 2020, at https://www.congress.gov/bill/115th-congress/senate-bill/3805#:~:text=%2F21%2F2018)-,Malicious%20Deep%20Fake%20Prohibition%20Act%20of%202018,media%20records%20that%20appear%20realistic.

199 Yvette D. Clarke. (Jun. 12, 2019). *Library of Congress*. "H.R.3230 – Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019." Accessed on Jun. 27, 2020, at https://www.congress.gov/bill/116th-congress/house-bill/3230.

200 Centre for Data Ethics and Innovation. (Sep. 2019). *Centre for Data Ethics and Innovation*. "Deepfakes and Audio-visual Disinformation." Accessed on Jul. 20, 2020, at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831179/Snapshot_Paper_-_Deepfakes_and_Audiovisual_Disinformation.pdf.

201 Centre for Data Ethics and Innovation. (Sep. 2019). *Centre for Data Ethics and Innovation*. "Deepfakes and Audio-visual Disinformation." Accessed on Jul. 20, 2020, at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831179/Snapshot_Paper_-_Deepfakes_and_Audiovisual_Disinformation.pdf.

202 European Commission. (Feb. 4, 2020). *European Commission*. "Illegal content on online platforms." Accessed on Jul. 20, 2020, at https://ec.europa.eu/digital-single-market/en/illegal-content-online-platforms.

203 Europol European Cybercrime Centre (EC3). (2019). *DeepFake*. Europol Platform for Experts – EPE.

204 Europol. (Jul. 18, 2019). *Europol*. "Do Criminals Dream of Electric Sheep?" Accessed on Oct. 14, 2020, at https://www.europol.europa.eu/newsroom/news/do-criminals-dream-of-electric-sheep-how-technology-shapes-future-of-crime-and-law-enforcement.

205 Europol European Cybercrime Centre (EC3). (2019). *DeepFake*. Europol Platform for Experts – EPE.

206 Europol. (Oct. 5, 2020). *Europol*. "Internet Organised Crime Threat Assessment (IOCTA) 2020." Accessed on Oct. 14, 2020, at https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2020.

TREND MICRO™ RESEARCH

Trend Micro, a global leader in cybersecurity, helps to make the world safe for exchanging digital information.

Trend Micro Research is powered by experts who are passionate about discovering new threats, sharing key insights, and supporting efforts to stop cybercriminals. Our global team helps identify millions of threats daily, leads the industry in vulnerability disclosures, and publishes innovative research on new threat techniques. We continually work to anticipate new threats and deliver thought-provoking research.

www.trendmicro.com