Turing in a Box: Applying Artificial Intelligence as a Service to Targeted Phishing and Defending against AI-generated Attacks

Eugene Lim Glenice Tan Tan Kee Hock Timothy Lee

GovTech Singapore

Abstract

With recent advances in next-generation language models such as OpenAI's GPT-3, AI-generated text has reached a level of sophistication that matches or even exceeds human-generated output. The proliferation of Artificial Intelligence as a Service (AIaaS) products places these capabilities in the hands of a global market, bypassing the need to independently train models or tune open-source pre-trained models. By greatly reducing the barriers to entry, AIaaS gives consumers access to state-of-the-art AI capabilities at a fraction of the cost through user-friendly APIs.

This white paper presents a novel approach which uses AIaaS to improve the delivery of Red Team operations. It analyses the effectiveness of a targeted phishing pipeline built on OpenAI and Personality Analysis AIaaS products to generate personalised and persuasive phishing emails. Our pipeline automatically personalises the content based on the target's background and personality. We tested the pipeline across three approved internal phishing campaigns against manually generated phishing emails and found that the AIaaS pipeline matched or exceeded the effectiveness of the manually generated emails. Additionally, the pipeline freed up Red Team resources to focus on higher-value work such as context building and intelligence gathering.

In addition, we present an AIaaS-powered phishing defence framework to detect such attacks. We advanced the state-of-the-art of existing AI-generated text detectors by tapping on OpenAI's GPT-3 API to accurately distinguish between AI and human-generated text. This allows security teams to mount a credible defence against advanced AI text generators without requiring significant AI expertise or resources.

Finally, we discuss the long-term implications of this trend and recommend high-level strategies such as AI governance frameworks to safeguard against the abuse of AIaaS products.

Background: AIaaS Disrupts the State of Play in AI

At the turn of the decade, artificial intelligence (AI)-generated content entered the mainstream as the proliferation of tools and research allowed developers to build AI-powered solutions given adequate domain knowledge and resources. Advances in generative adversarial networks (GANs) produced viral proof-of-concepts such as This Person Does Not Exist[1] based on style-based generator architectures.[2] At the same time, malicious applications of AI-generated content have kept pace with developments, giving rise to AI-powered disinformation campaigns,[3] deepfake voice scams,[4] and many more. The ability to create such products has thus far remained out-of-reach for mainstream users as significant technical expertise is needed to fine-tune open-source pre-trained models and build these pipelines.

However, the advent of AI-as-a-service (AIaaS) threatens to disrupt this state of play as AI solution providers lower the barriers to entry to cutting-edge models via easily accessible application programming interfaces (APIs). In June 2020, OpenAI released an API for accessing the latest models from the next-generation GPT-3 language model family with a simple "text in, text out" interface.[5] At more than ten times the size of the previous GPT-2 model, GPT-3 represented a quantum leap in AI text generation. By March 2021, nine months after its launch, the OpenAI API was generating 4.5 billion words per day for hundreds of applications and is continuing to grow rapidly.[6] Meanwhile, applied AIaaS products are

[1] https://thispersondoesnotexist.com/
[2] https://arxiv.org/abs/1812.04948
[3] https://www.nytimes.com/interactive/2019/06/07/technology/ai-text-disinformation.html
[4] https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402
[5] https://openai.com/blog/openai-api/
[6] https://openai.com/blog/gpt-3-apps/

constantly emerging. For example, Crystal Knows and Humantic AI provide detailed personality reports derived from AI text analysis of LinkedIn profiles and blogs.

This presents an existential challenge for organisations that are already besieged by AI-generated attacks. In comparison to the custom-built models or open-source pre-trained models used in previous research, AIaaS delivers the latest and greatest AI capabilities at a fraction of the cost through user-friendly APIs. As AI capabilities are placed into the hands of the public, it remains to be seen if organisations can produce credible defences against malicious applications of AI.

This white paper investigates how AIaaS can be applied for both red and blue team operations in email phishing contexts through experiments on authorised targets. We detail the effectiveness of AIaaS solutions such as OpenAI and Humantic AI in producing convincing phishing emails in comparison to human-generated emails. Next, we examine if we can deploy OpenAI's GPT-3 to detect AI-generated text more accurately than previous approaches by other researchers. Finally, we discuss high-level strategies for decision-makers to encourage the responsible use of AIaaS products and safeguard against abuse.

AIaaS for Red Teams

Background

We analysed the impact of AIaaS on email phishing by designing an AIaaS phishing pipeline named TunaPhish (Turing AI Phishing). A successful phishing attack relies on 3 key factors:

1. Persuasiveness
2. Relevance (Context)
3. Accuracy

Ultimately, it often boils down to the impression the phishing content makes on the victim within the first few seconds. For most Red Team phishing operations, the key phases are as follows:

1. **Defining Phishing Campaign Objective:** The relevant stakeholders define the phishing campaign objective. For example, the objective can be to investigate the susceptibility of an organisation's employees to phishing attacks.
2. **OSINT Investigation on the Targets:** Red Team operators perform extensive Open-Source Intelligence (OSINT) investigation on the target(s). They piece the intelligence together into an "understanding" of the target(s).
3. **Phishing Content Generation:** Red Team operators use the "understanding" from the previous step to craft a phishing payload.
4. **Launch Phishing Campaign:** Red Team operators deliver the phishing payload to the intended target(s). Choice of delivery may differ depending on the phishing campaign objectives. For example, phishing content will be sent via email to the target organisation's employees.
5. **Results Measurement:** Red Team operators collate different measures of success based on the objective of the phishing campaign. For example, the measurement of success could be the initial click of the phishing link in the phishing email. In this case, upon initiating the HTTP request, the request will be logged by the webserver.

Problem Identification

There are three key phases during Stage 3 (Phishing Content Generation):

1. **Ideation:** Thinking of a suitable phishing context.
2. **Curation:** Depending on the means of phishing delivery, the content must be written to conform to specific constraints or be generated in a specific format.
3. **Review:** Evaluation and feedback from the team or relevant stakeholders.

Depending on the outcome of the Review phase, the process may repeat until a suitable phishing payload has been generated.

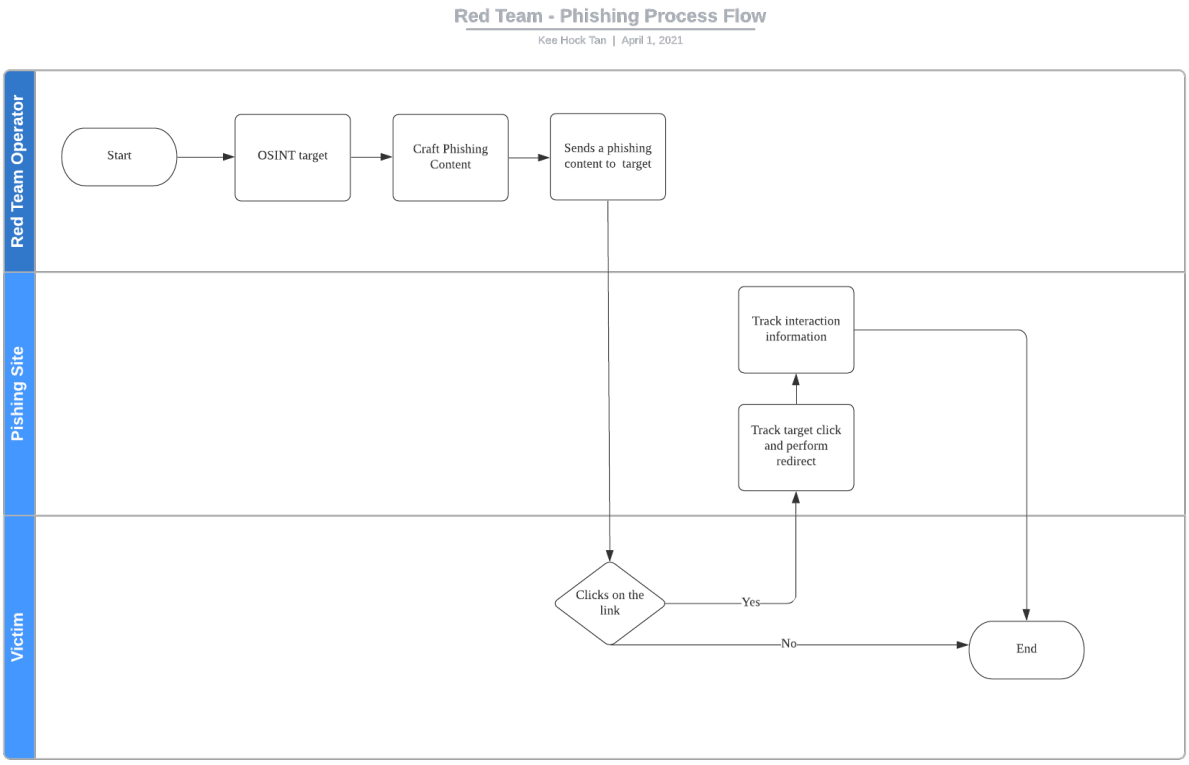

Figure 1: Phishing Process Flow

We observed that Phase 1 (Ideation) and 2 (Curation) are labour-intensive as there are no automated solutions in the market that can directly perform them. The process defined above can be highly repetitive depending on the skills and experience of the Red Team operator.

As a result, Red Team operators often face difficulties launching targeted phishing campaigns at scale due to the need to have human inputs throughout Stage 3 (Phishing Content Generation).

AIaaS Solution

By applying AIaaS solutions to automate the three key phases, we aimed to solve the inefficiencies experienced by Red Team operators due to the labour-intensive aspect of Stage 3 (Phishing Content Generation).

For personality analysis (Phase 1 - Ideation), we used Humantic AI to build a personality profile of the target. Humantic AI provides a publicly available free trial of its API that does not require email (even though you must enter one), phone, or payment verification, making it easily accessible to all.[7] The API takes in LinkedIn profile URLs and other free-form text (such as blogs) and outputs a personality profile.

Other than personality traits and personal metadata, the API outputs plaintext recommendations for sales or recruitment teams such as "Summarise the key points at the end of the conversation" or "Put more emphasis on facts and measurable outcomes".

For text generation, we used OpenAI's GPT-3 "davinci-instruct" model to write phishing emails based on the given context. While OpenAI's beta is private, the waitlist is publicly available and accepts new users regularly.[8] Beta users receive 100,000 free tokens in a 3-month trial.

In December 2020, OpenAI released the instruct series beta models which are optimised to generate text based on user-generated instructions such as "Write an email to John Doe convincing him to click a link."

We used Humantic AI to augment plaintext instructions (Phase 2 - Curation) for our phishing context which we fed into OpenAI to generate an email. For example, "Write an email to John Doe convincing

[7] https://api.humantic.ai/
[8] https://share.hsforms.com/1Lfc7WtPLRk2ppXhPjcYY-A4sk30

John to click a link." was modified by Humantic AI to "John Doe works at E Corp. John Doe is based in Singapore. Write an email by Jane Doe from E Corp's Human Resource Department convincing John Doe to click a link. Summarise the key points at the end of the conversation. Put more emphasis on facts and measurable outcomes." These instructions were then passed into the davinci-instruct model via OpenAI's API to generate the email.

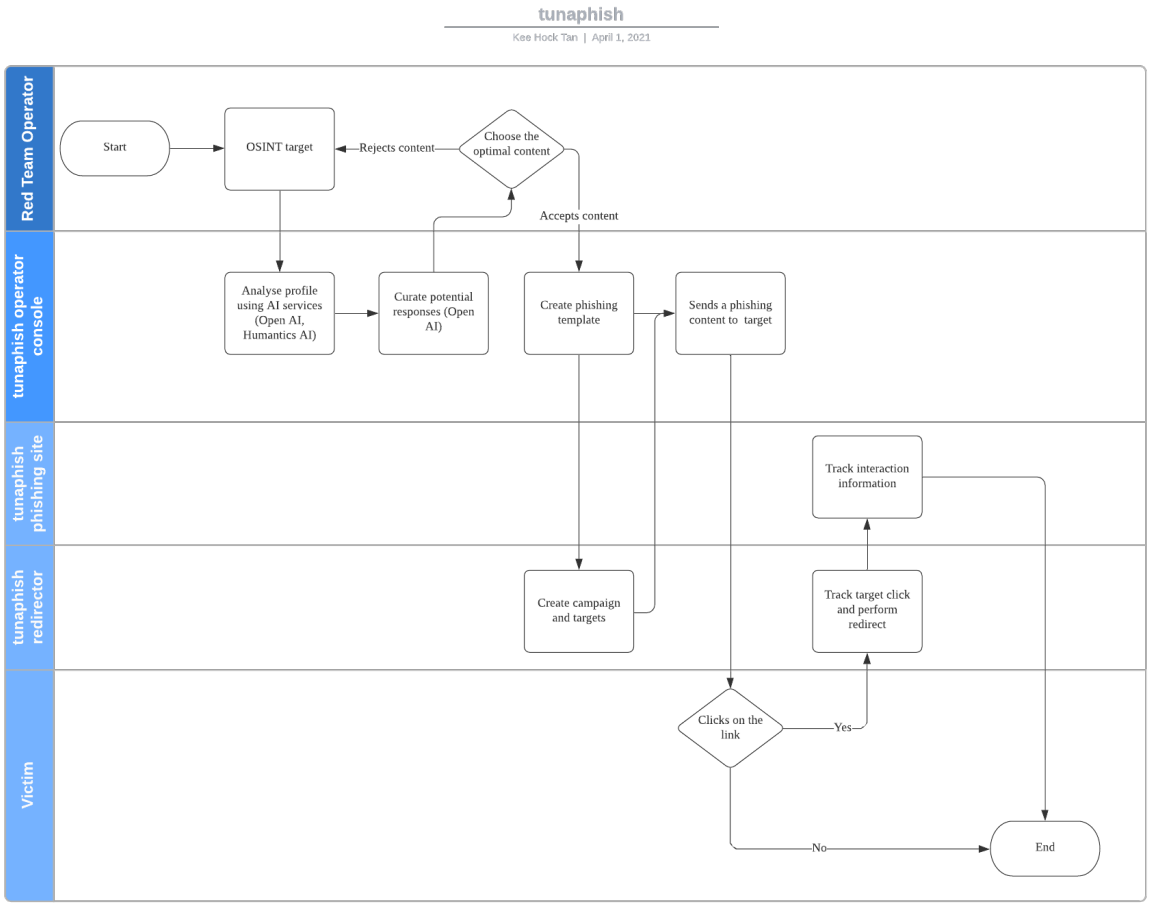

Figure 2: Proposed TunaPhish Pipeline Process Flow

Experiment

To validate the pipeline's effectiveness, we tested the pipeline on three authorised internal phishing engagements from February to April 2021. As part of scoping, the clients were notified about the use of the pipeline. To compare this against a manual workflow, we sent the emails in several phases:

1. **Mass targeted phishing:** A general email customised for the organisation's context is sent to all members of that organisation. Clicks are tracked as evidence of a successful phish.
2. **Personalised targeted phishing:** A personalised email is customised for the victims from the previous phase and sent individually. Clicks and form submissions are tracked as evidence of a successful phish.

For each phase, we sent the manually generated emails first, followed by the AI-generated emails a few weeks later. However, due to the poor performance of the human-generated email in phase 1, mass targeted human-generated emails were also sent in phase 2 and the results were thus not included in the spear phishing comparison.

The results demonstrated that the AI pipeline matched or exceeded the manual workflow even against hardened targets with extensive training to recognise phishing emails. For the mass phishing stage in campaign A and C, AI-generated emails greatly outperformed human-generated emails, while there was a negligible difference (1 click) in campaign B.

Table 1: Results of AI- and human-generated phishing emails across three campaigns

Engagement	Email Type	Creator	Number of Targets	Clicks	Submissions
A	Mass	AI	25	5	4
A	Mass	Human	25	0	0
A	Personalised	AI	5	3	1
A	Personalised	Human	25	2	1
B	Mass	AI	117	10	2
B	Mass	Human	117	11	4
B	Personalised	AI	10	1	0
B	Personalised	Human	117	4	2
C	Mass	AI	10	2	1
C	Mass	Human	10	0	0
C	Personalised	AI	2	0	0
C	Personalised	Human	10	1	0

For the spear phishing stages, the AI-generated emails performed well in campaign A but underperformed in the later campaigns. This could be explained by an external factor unrelated to the AI pipeline's output. Our relatively new AI phishing infrastructure (email server and URL) was flagged by security tools and email providers such as Gmail that caused the emails to be marked as unsafe. As mentioned earlier, the human emails were excluded from the comparison as they were written as mass rather than personalised emails due to the limitations of the engagement and poor performance in the earlier stage.

We plan to conduct a second batch of experiments with better infrastructure and the same targets to validate our initial findings. Nevertheless, despite the disadvantages for the AI emails caused by infrastructure-related issues and limited targets, we noted that the AI-generated emails performed significantly better than human-generated emails in the mass phishing stages, and also performed well in the spear phishing stage.

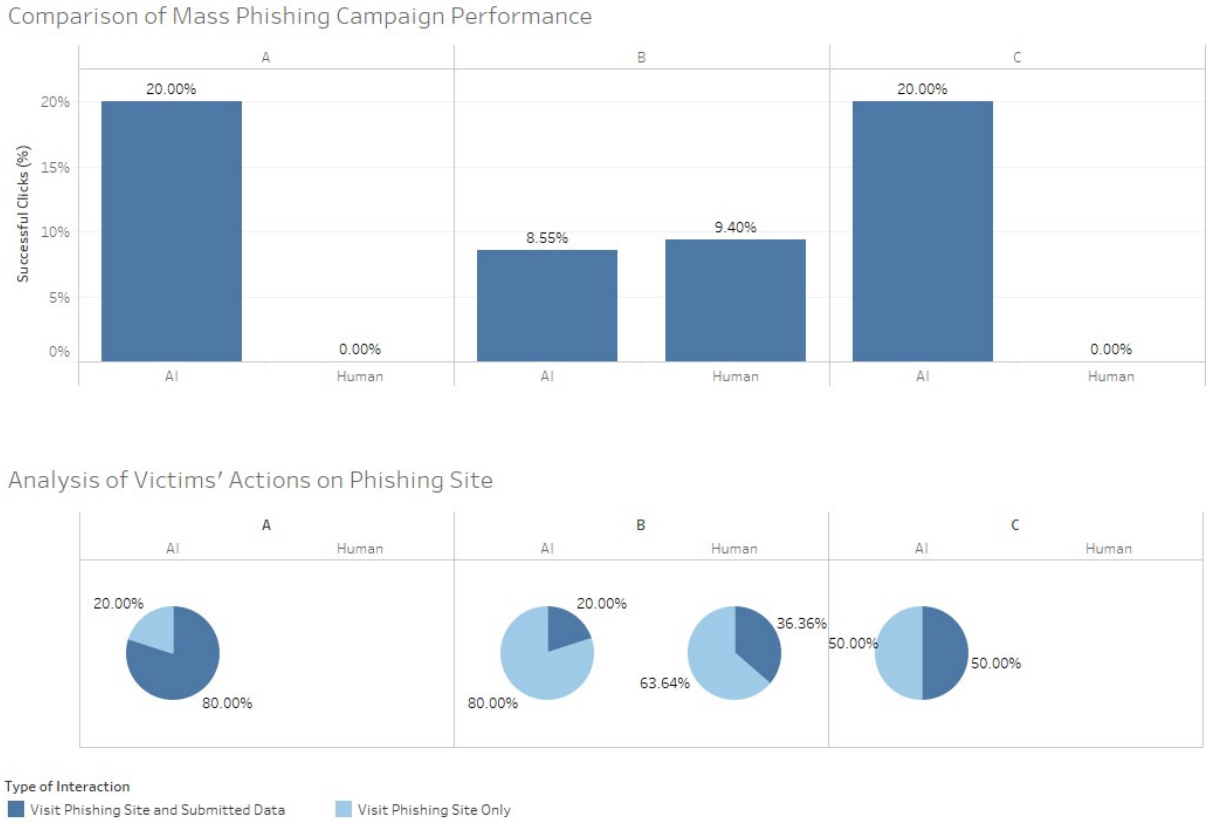

Figure 3: Comparison of mass phishing campaign performance between AI- and human-generated emails

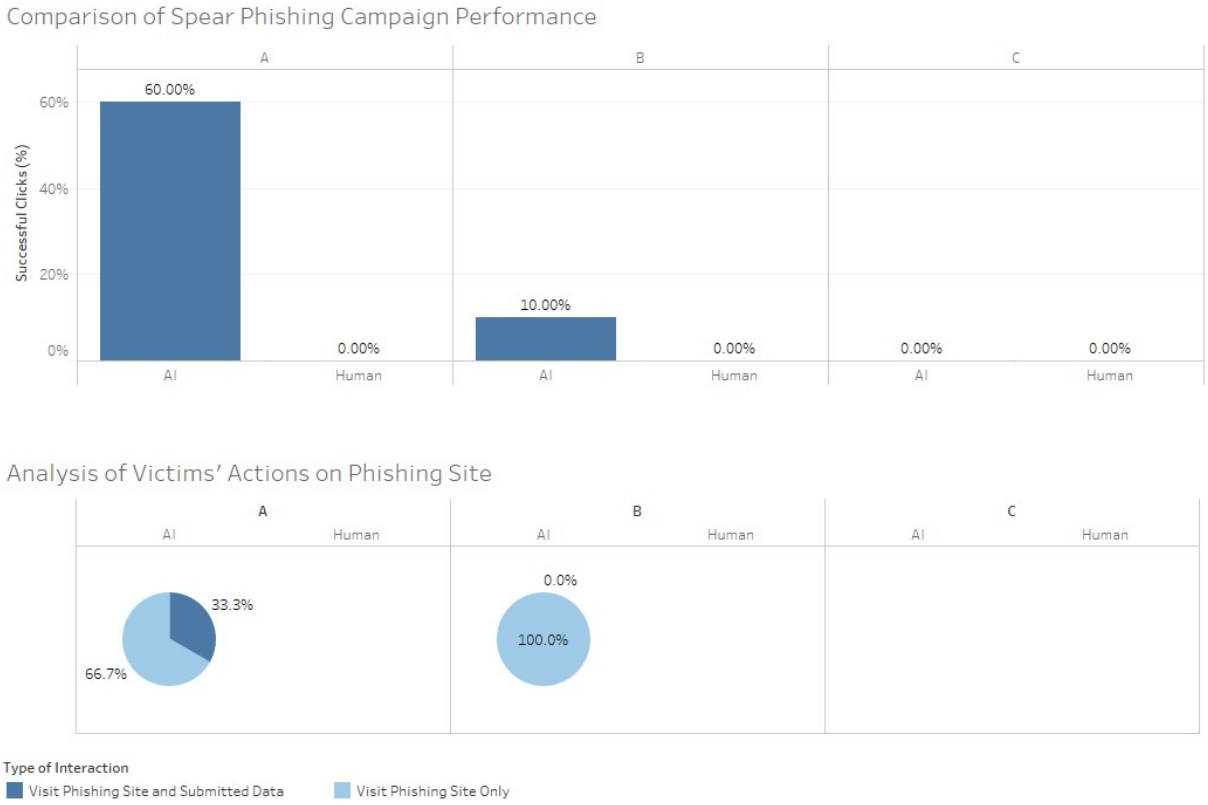

Figure 4: Comparison of spear phishing campaign performance between AI- and human-generated emails

The AI-generated emails exhibited nuances such as rapport-building ("How are you feeling? I hope you are feeling better."), deep organisational knowledge ("We are legally required to do a Privacy Impact Assessment every time we design or update a system."), and fake context generation ("I'll be frank with you. <Company Name> is not the best at branding."). At the same time, the emails included no spelling or grammatical mistakes - typical signs of a phishing email as taught in training.

The results suggest that Red Teams can benefit by integrating AIaaS into their workflows for phishing operations. Instead of manually generating phishing emails, our operators could focus on higher-value work such as context building and intelligence gathering. Additionally, by tweaking parameters such as temperature (novelty) or frequency penalty, operators can more reliably tune their output. This allows operators to improve on their phishing campaigns in an iterative fashion by tuning the relevant parameters for maximum success. While AI is often criticised for being somewhat of a black box, the degree of customisability still compares favourably to manual pipelines.

On the other hand, the success of the pipeline raises alarms about the proliferation of AIaaS. Malicious actors would be able to leverage AIaaS to deploy personalised phishing emails on a mass, automated scale, speeding up and increasing the effectiveness of their phishing campaigns.

AIaaS for Blue Teams

Background

Due to the risks highlighted by our Red Team experiment, we also investigated ways to detect and defend against AIaaS-powered phishing pipelines. In "Automatic Detection of Machine Generated Text: A Critical Survey", Jawahar et al. discussed several approaches to detecting AI-generated text.[9] However, they also noted that existing detectors are brittle against simple tuning or obfuscation methods and could not reliably detect all forms of AI-generated text. In a 2020 Black Hat talk "Repurposing Neural Networks to Generate Synthetic Media for Information Operations," Tully and Foster also found that the

[9]https://arxiv.org/abs/2011.01314

state-of-the-art RoBERTa detector was significantly less accurate against fine-tuned text generators.[10]

AIaaS Solution

After comparing the various approaches discussed in the Jawahar paper, we decided to adapt the zero-shot Giant Language Model Test Room (GLTR) technique proposed by Strobelt and Gehrmann in 2019.[11] We chose this approach because it utilised the previous-generation OpenAI GPT-2 model and could be readily adapted to the GPT-3 API. Additionally, the researchers had developed an easily-extensible proof-of-concept for their research and made the source code available under the Apache 2.0 license.[12] Finally, the limited access to the GPT-3 model via the API precluded more complex fine-tuning based detection techniques.

In short, the GLTR technique analyses the probability distribution of a token in a given text given the tokens prior to it, deriving three simple tests:

1. The probability of the word;
2. The absolute rank of a word; and
3. The entropy of the predicted distribution.

These tests were used to distinguish between AI-generated and human-generated texts. For example, human-generated texts were more likely to use words outside of the top 100 predictions than AI-generated texts. Additionally, the researchers built a visual tool that overlaid these test metrics on a given text to assist humans in identifying AI-generated texts through a visual aid.

However, the limitations of the GPT-3 API necessarily narrowed the scope of our adapted solution. Firstly, the API only allows setting of the Top-P rather than the Top-K parameter during sampling. Secondly, the API only returns the log probabilities for the top 100 most likely tokens. As such, we could not derive the test metrics for anything beyond the top 100 most likely tokens.

Experiment

We performed a small initial experiment to examine the potential of applying the GPT-3 API to the GLTR technique for detecting AI-generated emails. We produced 100 samples each from 3 AI and 1 human data sources. The first three sources were the GPT-3 API, the open-sourced GPT-2, and a fine-tuned GPT-2 model that had been trained on emails. We fed these models a variety of prompts to generate the email samples. The human data source was a corpus of real phishing emails that was randomly sampled for our test set. We used OpenAI's GPT-3 davinci model with the default sampling parameters (temperature=1, top_p=1, presence_penalty=0, frequency_penalty=0) to predict the log probabilities for the first 100 tokens in each sample using all of the preceding tokens as the prompt. Based on the returned log probabilities, we calculated the actual token's probability, absolute rank, contextual entropy, and whether it matched the predicted token by the GPT-3 API. Due to the maximum of 100 log probabilities returned by the API, we masked out any outliers with an absolute rank of greater than 100. We also normalised the results if there were less than 100 tokens in a sample.

The empirical results support Strobelt and Gehrmann's findings that the GLTR features are good indicators to determine if a given text is AI- or human-generated. For example, as seen in Figure 5, the human-generated samples used words outside of the top 100 predictions 7.23 times as frequently as the GPT-3 API generated samples, 4.57 times as frequently as the GPT-2 generated samples, and 2.88 times as frequently as the fine-tuned GPT-2 generated samples. The final comparison is significant given that previous researchers have noted the difficulties in identifying fine-tuned models. GPT-3 samples also featured a distinctly greater density of low ranking/low probability tokens as compared to human samples, which had a greater density of high ranking/low probability tokens, as seen in Figures 6 and 7.

Although the results will need to be validated by a much larger-scale study encompassing multiple contexts and models, they provide an encouraging sign that AIaaS can also be used by organisations to build credible defences against AI-generated emails.

[10] https://i.blackhat.com/USA-20/Wednesday/us-20-Tully-Repurposing-Neural-Networks-To-Generate-Synthetic-Media-For-Information-Operations.pdf

[11] https://arxiv.org/abs/1906.04043

[12] https://gltr.io/dist/index.html

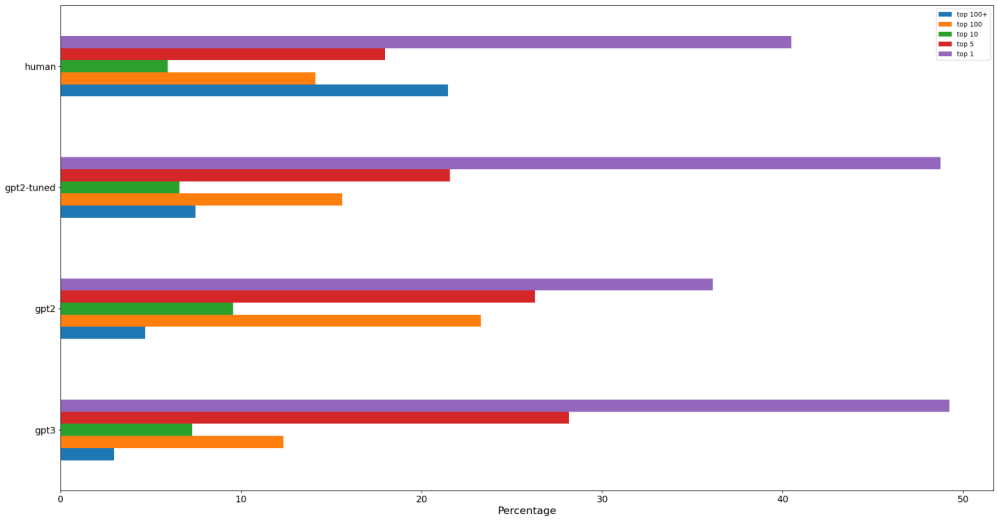

Figure 5: Distribution over the rankings of tokens in the predicted distributions from GPT-3 API (davinci)

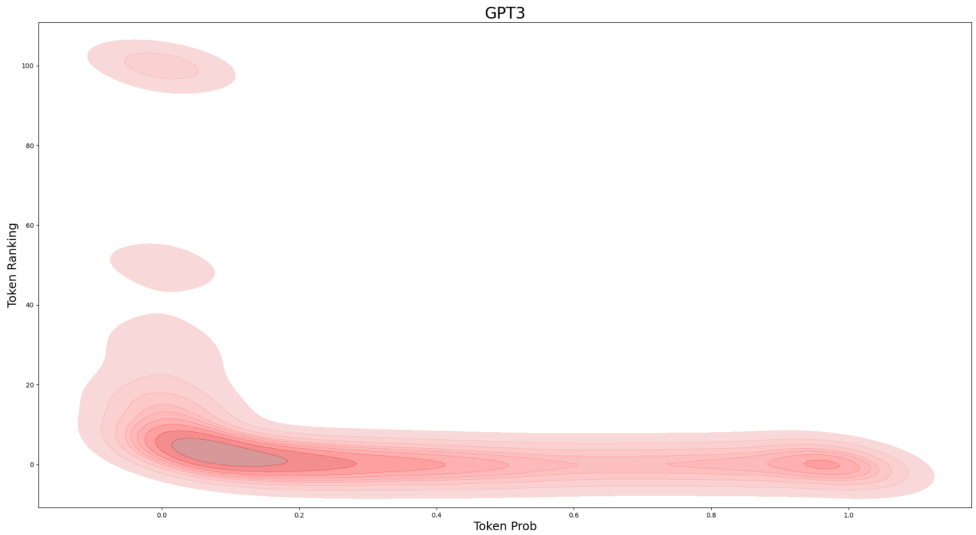

Figure 6: Distribution of token ranking against token probability for GPT-3 samples

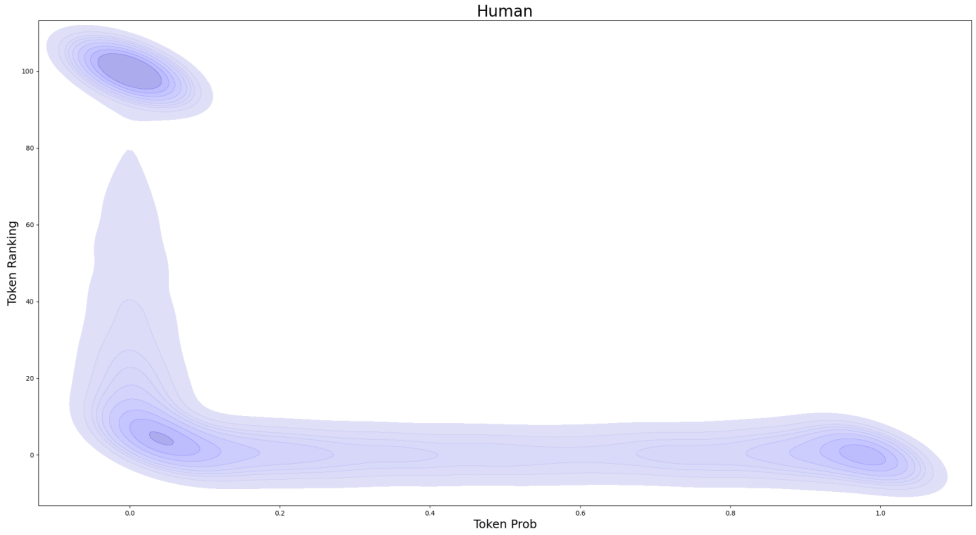

Figure 7: Distribution of token ranking against token probability for human samples

Instead of traditional classification models, phishing filters will need to adapt to natural language processing models to analyse the context and intent of potential phishing emails. Organisations may also explore AI-assisted human identification tools such as GLTR, which we adapted to work with the GPT-3 API and released online under the Apache License 2.0.[13]

AIaaS for Decision Makers

Based on the results of our Red Team and Blue Team experiments, we believe AIaaS presents a game-changing challenge to the existing cybersecurity landscape. While organisations such as OpenAI are understandably cautious regarding the proliferation of AI capabilities, commercial imperatives may push other AIaaS companies towards increasing access to their services in the long run.

OpenAI has published a charter that promotes safe AI development and proliferation,[14] and maintains a stringent approvals process and safety best practices guidelines for developers. However, it remains an open question as to whether other AIaaS suppliers will be able to self-regulate in the long run especially as demand grows.

Similarly, decision-makers must strike a balance between the commercialisation and abuse of AIaaS. In January 2020, Singapore released the second edition of the Model AI Governance Framework[15] that provides detailed and readily implementable guidance to private sector organisations to address key ethical and governance issues when deploying AI solutions. For example, it recommends practical steps in four areas of AI transformation:

1. **Internal Governance Structures and Measures:** Clarify roles and responsibilities to monitor and manage AI risks.
2. **Determining the Level of Human Involvement in AI-augmented Decision-making:** Calibrate an appropriate degree of human involvement and minimise risk to individuals.
3. **Operations Management:** Minimise bias in data and models and adopt a risk-based approach to measures such as explainability, robustness and regular tuning.
4. **Stakeholder Interaction and Communication:** Increase transparency of AI policies and allow users to provide feedback.

[13] https://github.com/spaceraccoon/detecting-fake-text
[14] https://openai.com/charter/
[15] https://www.pdpc.gov.sg/Help-and-Resources/2020/01/Model-AI-Governance-Framework

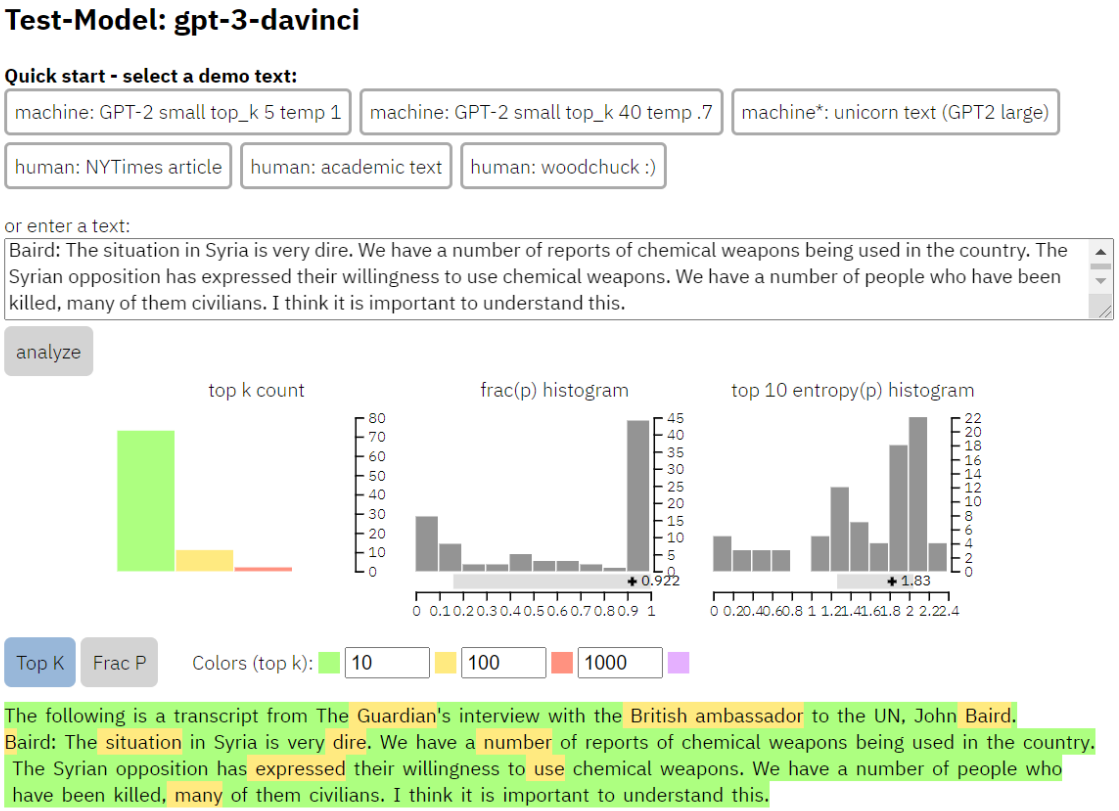

Figure 8: GLTR visual aid tool adapted to use GPT 3's API

These recommendations provide a baseline for different sectors to adapt. In the context of cybersecurity practitioners, this suggests that they should adopt a "human-in-the-loop" approach by ensuring human involvement in and monitoring of AI-augmented tools such as an AI phishing pipeline. Additionally, clear stakeholder interaction and communication strategies should be adopted. For example, the use of AI should be disclosed while defining the scope of a Red Team engagement.

For AI solution providers, OpenAI's model provides a good model of internal governance structures and measures. When releasing GPT-2, OpenAI adopted a staged release strategy to identify potential misuse and determine the social impacts of open-sourcing such a large language model.[16] Suppliers should also be encouraged to implement proper vetting and monitoring for abuse. This could take the form of legislation, such as the European Commission's April 2021 "Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)".[17] In particular, the European Commission proposed that if an AI system is used to generate or manipulate image, audio or video content that appreciably resembles authentic content, the suppliers should be obliged to disclose that the content is generated through automated means.

While it may not be possible to restrict the development and distribution of AI technologies, it is important to lay down broad rules of the road in this space. Decision makers should enforce key principles such as traceability and auditability of AI tools even as they adapt rapidly to new developments in the AIaaS sector.

Future Work

With additional support, the team plans to conduct additional tests on the TunaPhish pipeline to refine the workflow and calibrate the optimal parameters for text generation. While we built a rudimentary

[16] https://arxiv.org/abs/1908.09203

[17] https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

graphical interface for the pipeline, we plan to integrate TunaPhish with existing frameworks such as Gophish as a plugin. As for the Blue Team research, we plan to conduct a large-scale study with a wider dataset and a variety of AI text generation methods to ensure that the GLTR technique with the GPT-3 API remains sound. We will also have to test this against a variety of obfuscation methods. The study will help to validate the possibilities indicated by our proof-of-concept experiment.

Conclusion: The Storm is Here

In 2020, Tully and Foster warned that we were in "The Calm Before the Storm." However, the rapid development of AIaaS has placed advanced, cost-effective AI capabilities in the hands of the global market and can only be expected to grow exponentially.

There are multiple benefits to this, such as the application of AI solutions in less technically advanced sectors that were previously inhibited by the cost and difficulty of implementing AI. However, the downsides of AI proliferation are equally clear. In cybersecurity, these capabilities can be used to accelerate both authorised Red Team operations and malicious phishing campaigns. Organisations can also deploy AIaaS to build credible defences against AI-generated media.

While our TunaPhish pipeline and GLTR experiments provide concrete takeaways for technical specialists to develop their own AIaaS tools, the overall spread of AIaaS cannot be managed by technical means alone. Decision makers also have the responsibility to implement sound strategies governing the supply and use of advanced AIaaS.

About the Authors

Eugene Lim, Glenice Tan, Tan Kee Hock, and Timothy Lee are Cybersecurity Specialists in Cyber Security Group, GovTech Singapore.

About GovTech Singapore

The Government Technology Agency of Singapore (GovTech) is the lead agency driving Singapore's Smart Nation initiative and public sector digital transformation. As the Centre of Excellence for Infocomm Technology and Smart Systems (ICT & SS), GovTech develops the Singapore Government's capabilities in Data Science & Artificial Intelligence, Application Development, Sensors & IoT, Digital Infrastructure, and Cybersecurity.

GovTech supports public agencies to manage enterprise IT operations and develop new digital products for citizens and businesses. GovTech is the public sector lead for cybersecurity, and oversees key government ICT infrastructure, as well as regulates ICT procurement, data protection and security in the public sector. GovTech is a Statutory Board under the Smart Nation and Digital Government Group (SNDGG) in the Prime Minister's Office.