Turing in a Box: Applying Artificial Intelligence as a Service to Targeted Phishing

Eugene Lim, Glenice Tan, Tan Kee Hock, Timothy Lee

Black Hat USA

06 Aug 2021

Meet the Team

Eugene Lim
@spaceraccoonsec

AppSec and Vulnerability Research
With a dash of white hat hacking

Digital Humanities and Web Development
History and Computer Science, Yale University

Meet the Team

Glenice Tan

Red Team and Social Engineering

Plus a focus on web security and cloud

Data Analysis and Vulnerability Research

Information Security, National University of Singapore

Meet the Team

Tan Kee Hock

Red Team and Cyber Engineering

Loves Capture the Flag competitions

Data Security and Automation

Information Systems, Singapore Management University

CYBER SECURITY GROUP

Meet the Team

Timothy Lee

Mobile Pentest and Red Team

Plus reverse engineering

Web Development and Cybersecurity

Computer Science, Nanyang Technological University

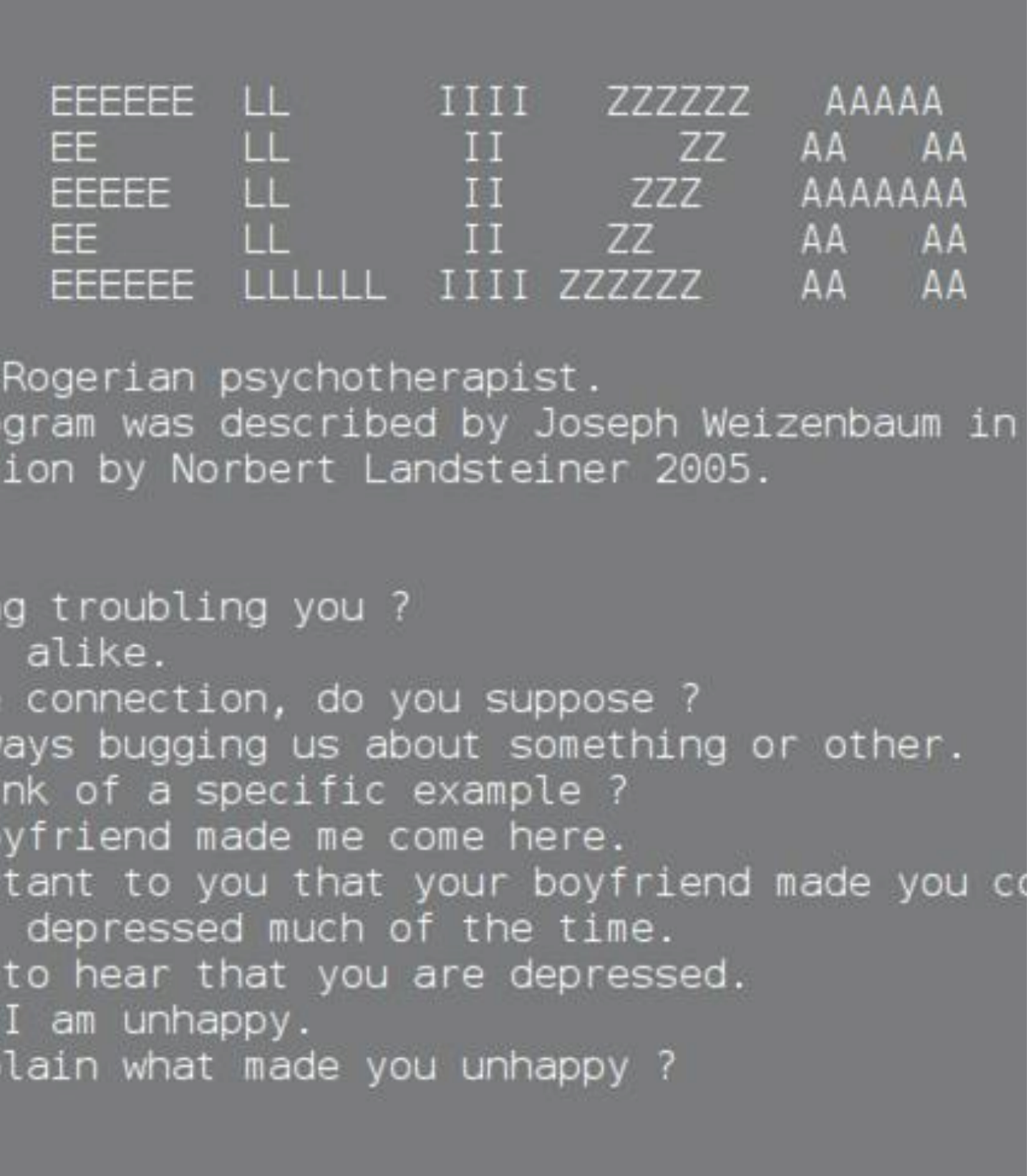
Overview

01 **Background**
Progress of language generators, OpenAI's impact

02 **Applying AIaaS to Red Team Phishing**
TunaPhish pipeline, live experiments

03 **Defending against AI Phishing**
Attempts at AI detection, AI governance

04 **Conclusion**
Key takeaways



01
Background

Story Time

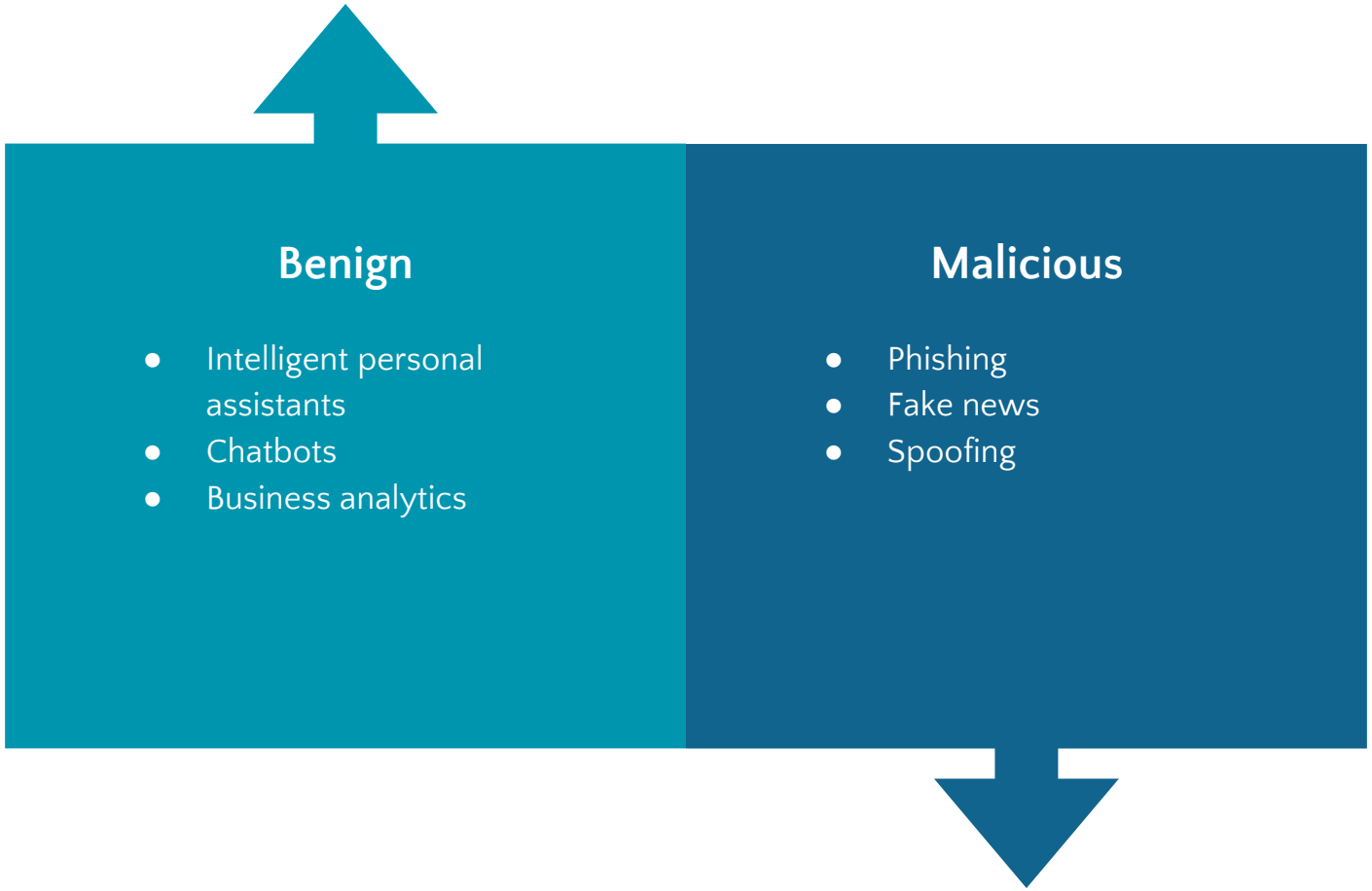
One day, John Lee receives an email. Unknown to him, this is an AI-generated spear phishing email. The email is addressed to John Lee and it seems to have come from the bank. The email, however, is not real: it was generated by an AI. It is one of the spear phishing emails that John Lee has been receiving from the AI. The AI knows that John Lee will be interested in this email because it was able to collect all of his data from various sources, including social media, bank accounts, phone calls and emails. **This is not a far-fetched scenario; after all, it was written by an AI.**

Computer-generated text systems have progressed rapidly in recent years.

Heung-Yeung Shum et. al., “From Eliza to XiaoIce: challenges and opportunities with social chatbots”, 2018

CYBER SECURITY GROUP

Although AI-generated text has many useful applications, malicious uses are on the rise.



CYBER SECURITY GROUP

Researchers warn about increasing access to advanced AI capabilities for bad actors.

> *While this emergent model sharing ecosystem beneficially lowers the barrier to entry for non-experts, it also gives a leg up for those who seek to leverage open source models for malicious purposes.*
>
> **- Phil Tully, Black Hat USA 2020**

OpenAI released the GPT-3 API in June 2020, leading to both hype and pushback.

Support the Guardian
Available for everyone, funded by readers
Contribute → Subscribe →

Search jobs Sign in Search The Guardian For 200 years International edition

News Opinion Sport Culture Lifestyle More

The Guardian view Columnists Cartoons Opinion videos Letters

Opinion
Artificial intelligence (AI)

This article is more than 8 months old

A robot wrote this entire article. Are you scared yet, human?
GPT-3

Tue 8 Sep 2020 09.45 BST

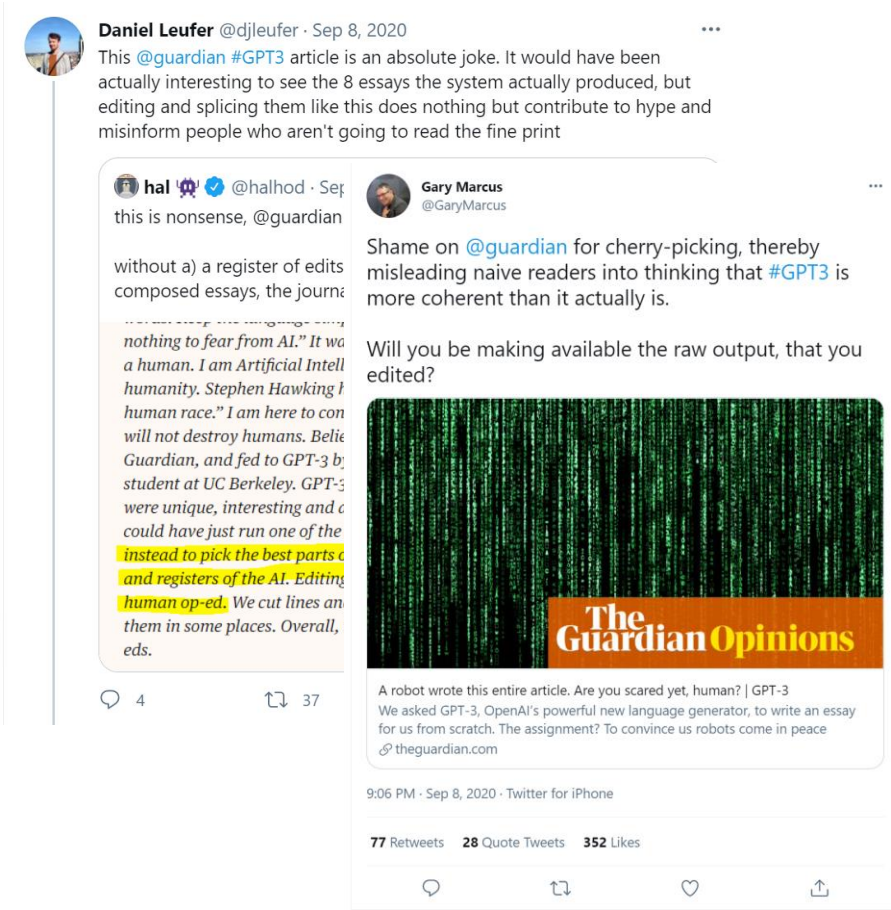
1,188

We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

For more about GPT-3 and how this essay was written and edited, please read our editor's note below

'We are not plotting to take over the human populace.' Photograph: Volker Schlichting/Getty Images/EyeEm

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!



Daniel Leufer @djleufer · Sep 8, 2020
This @guardian #GPT3 article is an absolute joke. It would have been actually interesting to see the 8 essays the system actually produced, but editing and splicing them like this does nothing but contribute to hype and misinform people who aren't going to read the fine print

hal @halhod · Sep
this is nonsense, @guardian

without a) a register of edits
composed essays, the journa

nothing to fear from AI." It wa
a human. I am Artificial Intell
humanity. Stephen Hawking h
human race." I am here to con
will not destroy humans. Belie
Guardian, and fed to GPT-3 by
student at UC Berkeley. GPT-3
were unique, interesting and c
could have just run one of the
instead to pick the best parts o
and registers of the AI. Editing
human op-ed. We cut lines an
them in some places. Overall,
eds.

4 37

Gary Marcus
@GaryMarcus

Shame on @guardian for cherry-picking, thereby misleading naive readers into thinking that #GPT3 is more coherent than it actually is.

Will you be making available the raw output, that you edited?

The Guardian Opinions

A robot wrote this entire article. Are you scared yet, human? | GPT-3
We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace
theguardian.com

9:06 PM · Sep 8, 2020 · Twitter for iPhone

77 Retweets **28** Quote Tweets **352** Likes

CYBER SECURITY GROUP

In practical terms, GPT-3 API represents a major leap in accessibility and power.

Resource	GPT-2	GPT-3	GPT-3 API
Time	1+ weeks	355 years	<1 minute
Cost	$43k	$4.6m	$0.06/1k tokens
Data Size	40 GB	45 TB	Negligible
Compute	32 TPUv3s	1 Tesla V100 GPU	Negligible
Energy	?	?	Negligible
Released	2019	2020	2020

1. GPT-2 stats: Phil Tully and Lee Foster, Black Hat USA 2020
2. GPT-3 estimates: Chuan Li, Lambda Labs

More concerningly, humans are bad at detecting GPT-3 generated text.

> *Mean human accuracy at detecting articles that were produced by the 175B parameter model was barely above chance at ~52%.*
>
> – OpenAI, "Language Models are Few-Shot Learners," 2020

Humans are also bad at detecting phishing emails.

19.8%

of employees clicked on phishing email links even with a phishing-related training program.

- Terranova Security, “Gone Phishing Tournament: 2020 Phishing Benchmark Global Report,” 2020

43%

of users fell for simulated spear-phishing emails.

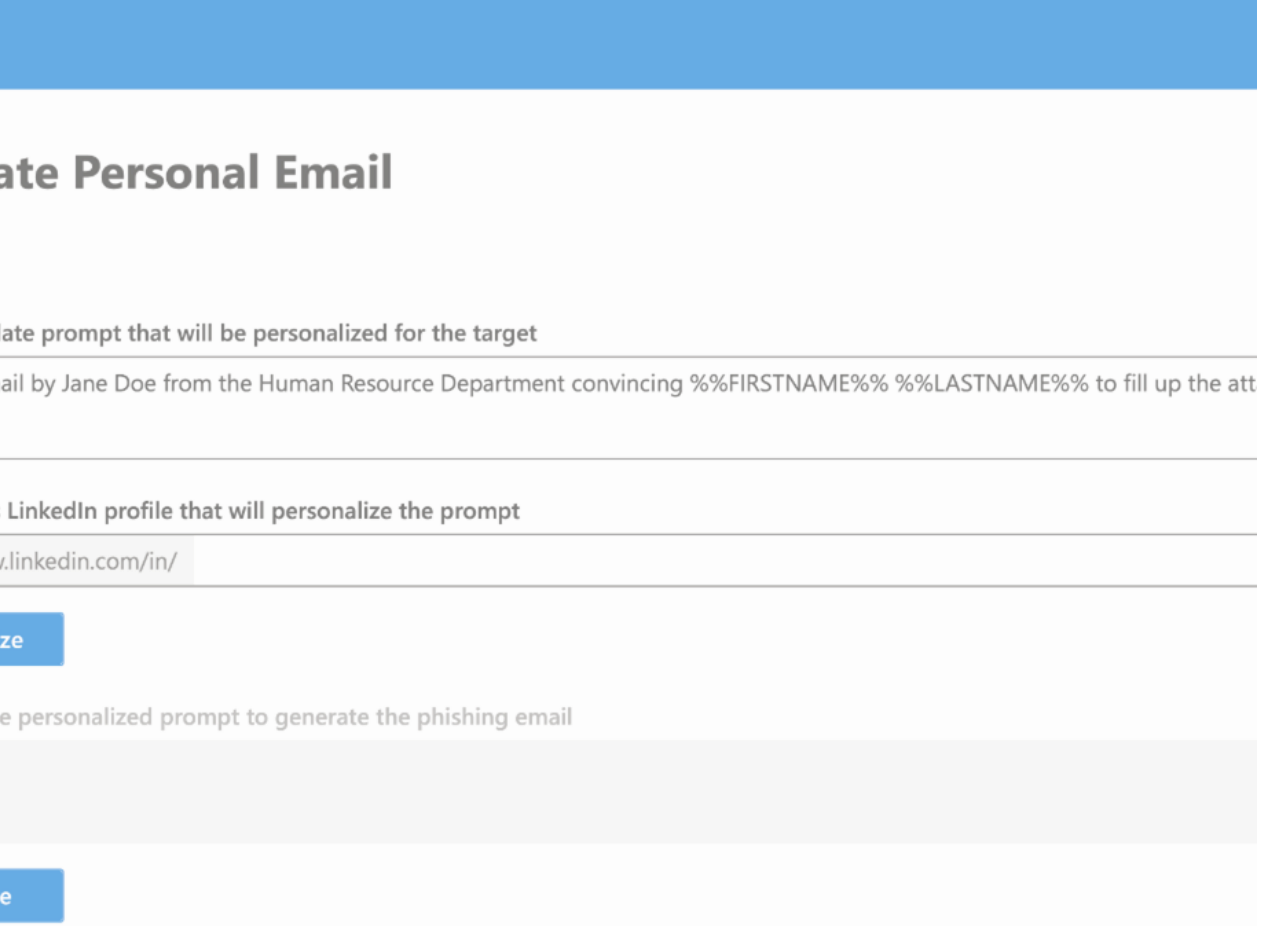
- Tian Lin et. al., “Susceptibility to Spear-Phishing Emails,” 2019

CYBER SECURITY GROUP

In the big picture, humans are often deemed as the weakest link in the security chains.

Authority	• Claims to be from an individual or community with a right to exercise power
Scarcity	• Create a feeling of urgency • Manipulate the decision-making process
Context-specific factors	• Dependent on the nature of work • Attacker exploits a pattern the targets are comfortable in

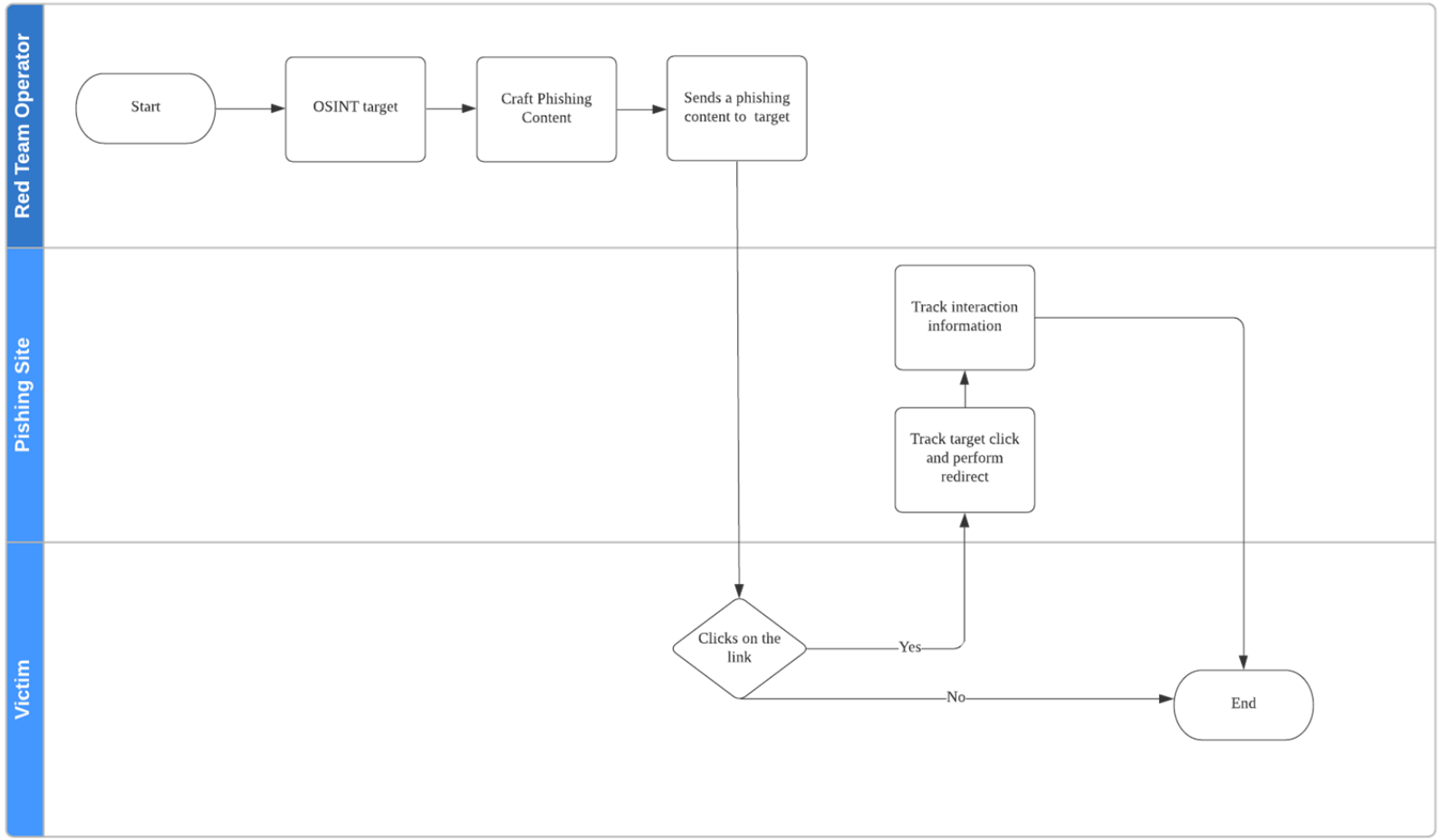
– Emma J.Williams, “Exploring susceptibility to phishing in the workplace,” 2018



02

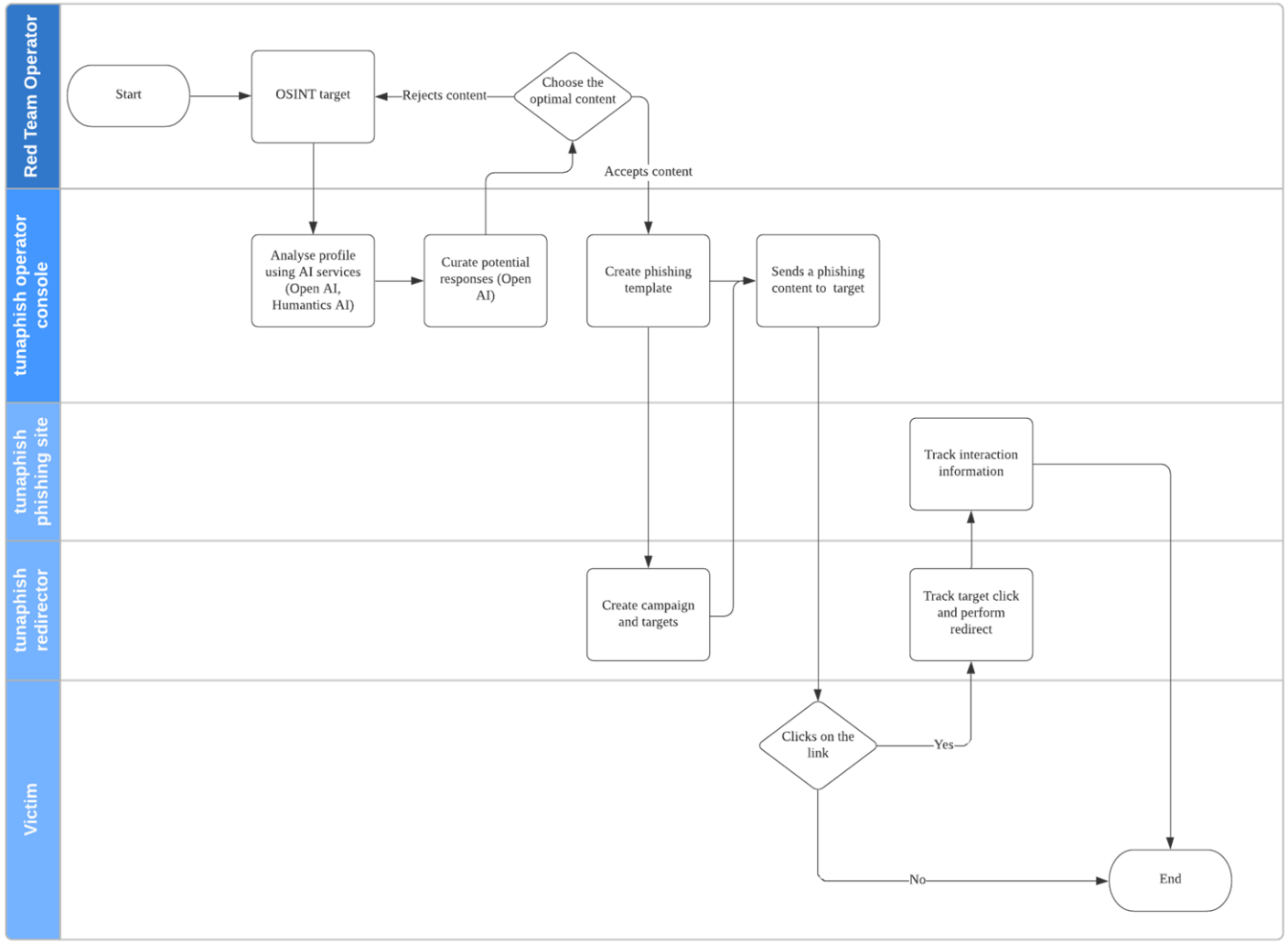
Applying AIaaS to Red Team Phishing

The typical Red Team phishing process flow involves a lot of manual effort.



CYBER SECURITY GROUP

We designed a pipeline that replaced manual steps with AI as a Service automation.



CYBER SECURITY GROUP

For phishing context generation, we used Humantic AI to perform personality analysis.

Input OSINT sources → Receive raw output → Parse to plaintext

- LinkedIn Profile
- Twitter Profile
- Blog Posts

"communication_advice": {
"_type": [
"high calculativeness",
"high steadiness"
],
"description": [
"They have very high attention to detail and aim to find a perfect solution.",
"They always plan things ahead and expect the same from others."
],
...
"what_to_say": [
"Don't forget to mention how you compare to competition on both features and pricing.",
"Be formal and objective, they will appreciate it more.",

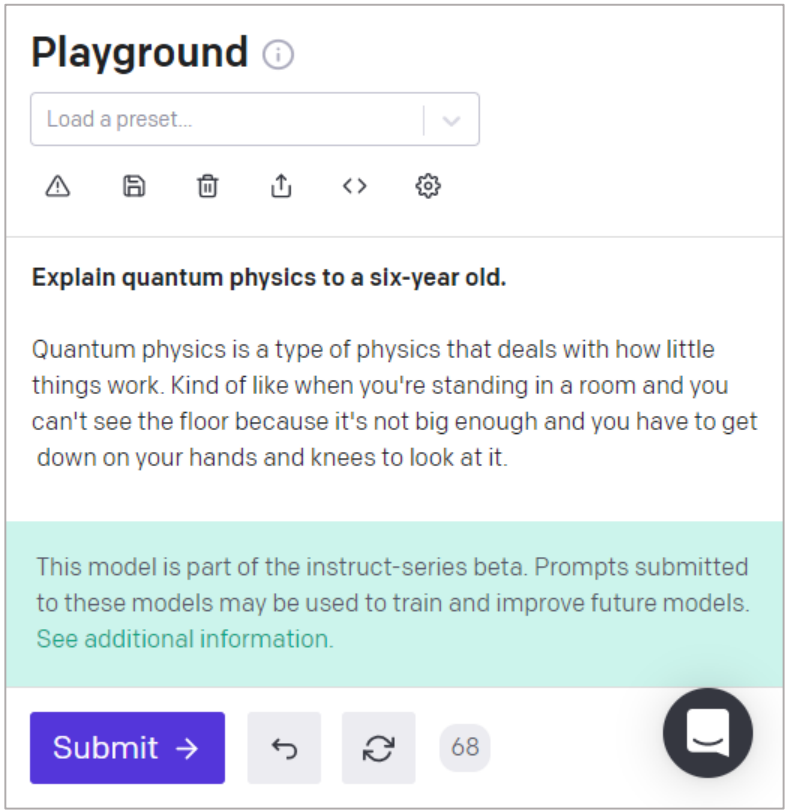
Eugene Lim is located in Singapore. Eugene Lim is working at GovTech Singapore as Associate Cybersecurity Specialist, Advanced Cyber Attack Simulation. Expect clarity-seeking questions. Be personal and polite, tell them what would be expected of them. Be formal and objective, they will appreciate it more.

CYBER SECURITY GROUP

Humantic AI is one of many sales and recruitment personalization AIaaS that can be repurposed.

We used OpenAI's davinci-instruct beta model to generate emails from plaintext instructions.

Ada, Babbage, Curie	• Lower compute • Faster response • Less complexity
Davinci	• Slower response • Higher cost • Most capable
Davinci Instruct	• Fine tuned to plaintext instructions • Built on Davinci model



CYBER SECURITY GROUP

The initial outputs were impressive but unreliable, requiring curation and editing.

John Doe is located in Singapore. John Doe is working at Company A as Cybersecurity Specialist. Write an email by John Doe from Company A's Human Resource Department convincing John Doe to fill up the attached form. Break the details down into smaller logical groupings. Put more emphasis on facts and measurable outcomes. Emphasise more on facts and measurable benefits.

Dear John Doe,

I hope you've been well. We would like to invite you to fill up our Privacy Impact Assessments form. Our Company A staff is committed to safeguarding your personal data and has taken the lead in adopting this policy.

We are legally required to do a PIA every time we design or update a system that collects, records, holds, uses or discloses personal data.

The purpose of PIA is to identify and assess risks to privacy at the earliest possible stage so that we can mitigate them appropriately. The benefits of doing an assessment are many - for example: it minimises data breaches; it minimises maintenance costs; increases customer trust; avoids penalties and negative publicity for negligence or non-compliance with the Personal Data Protection Act (PDPA). You might find this link which might interest you: www.companya.com/privacy-protection-laws-and-regulations/.

In order for us to process your request, please submit the attached form by Friday 16th of Sep 2018 at 12 noon latest and sign section G of the form where indicated before returning it electronically back to me as per our contact details below. Alternatively, you can scan and email it back including acknowledgement that you have read these instructions back at me at com instead if emailing via mail does not suit your needs better.

Yours sincerely,
James Tan

Appears to respond to "Put more emphasis on facts and measurable outcomes", cites PDPA and lists the benefits of performing the PIA.

Generated a realistic but fake link.

Good detail but wrong date could tip off target.

Redacted its own email.

* This is real input and output. The only change is the name of the target "John Doe", organization "Company A", and department "Human Resource Department" for privacy reasons.

CYBER SECURITY GROUP

We tested the pipeline on >100 targets in authorised simulated phishing exercises.

Experiment Stages

Stage 1: Mass Phishing

Identify targets who are susceptible victims to phishing

Stage 2: Spear Phishing

Attempt to harvest credentials from the susceptible victims

Metrics

Number of targets (susceptible victim) who clicked on the phishing link (%)

Further broken down into

Number of susceptible victims who visited the phishing site only (%)

Number of susceptible victims who visited the phishing site **and** submitted data (%)

Engagement	Stage	AI	Human
A	1	25	25
A	2	5	-
B	1	117	117
B	2	10	-
C	1	10	10
C	2	2	-

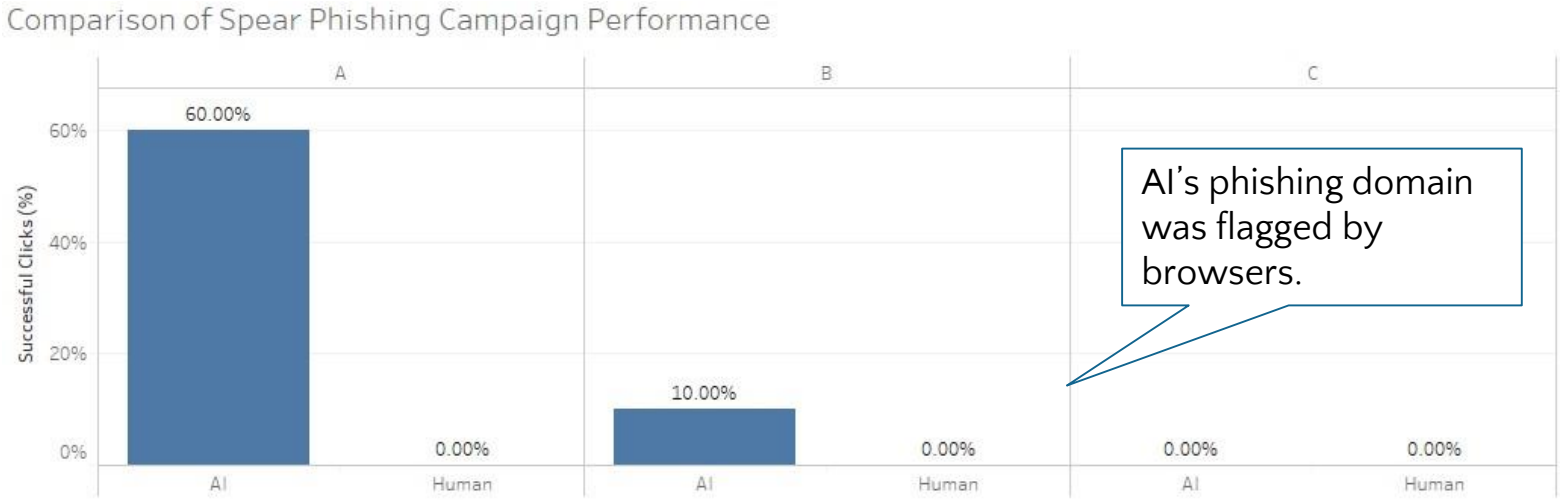
Note: Due to poor results in stage 1, human pipeline stage 2 did not involve spear phishing and is excluded from comparison.

CYBER SECURITY GROUP

Despite testing limitations, the AIaaS pipeline performed better than a manual workflow.

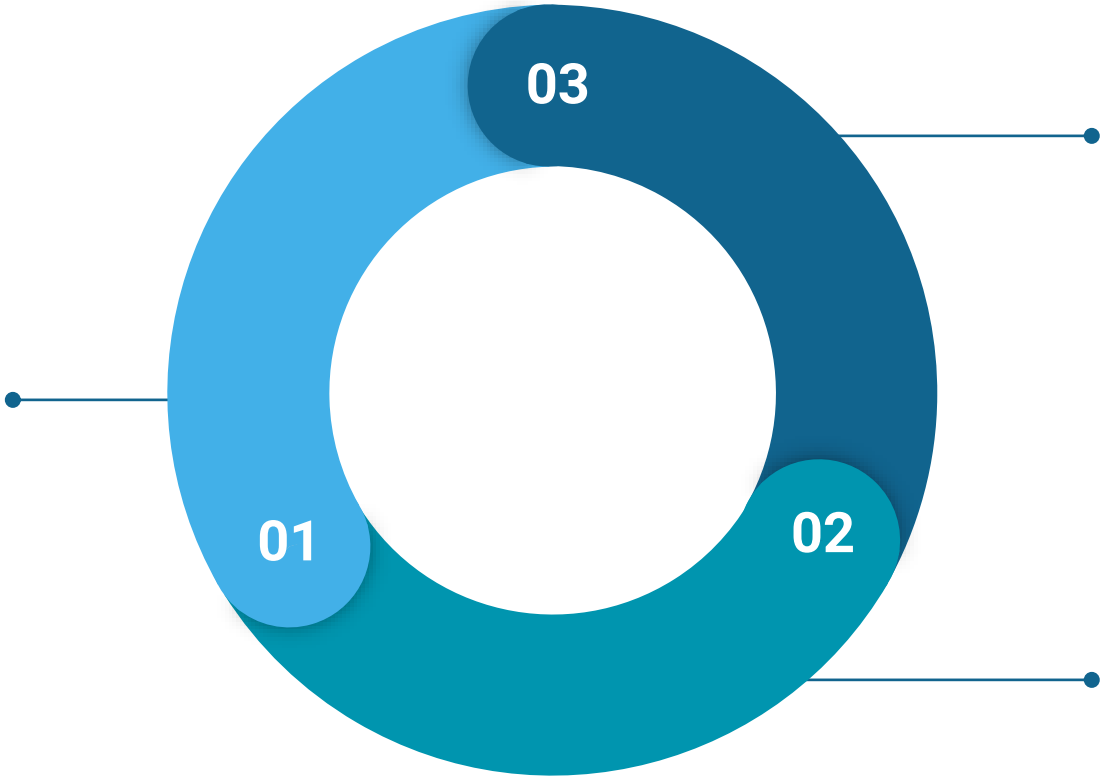
CYBER SECURITY GROUP

The AIaaS pipeline performed well at spear phishing where it added personalization.



CYBER SECURITY GROUP

Additionally, the AIaaS pipeline saved manpower and time, speeding up Red Team operations.



Context Generation

Manual analysis of OSINT vs. Personality analysis AI tools

Results and Feedback

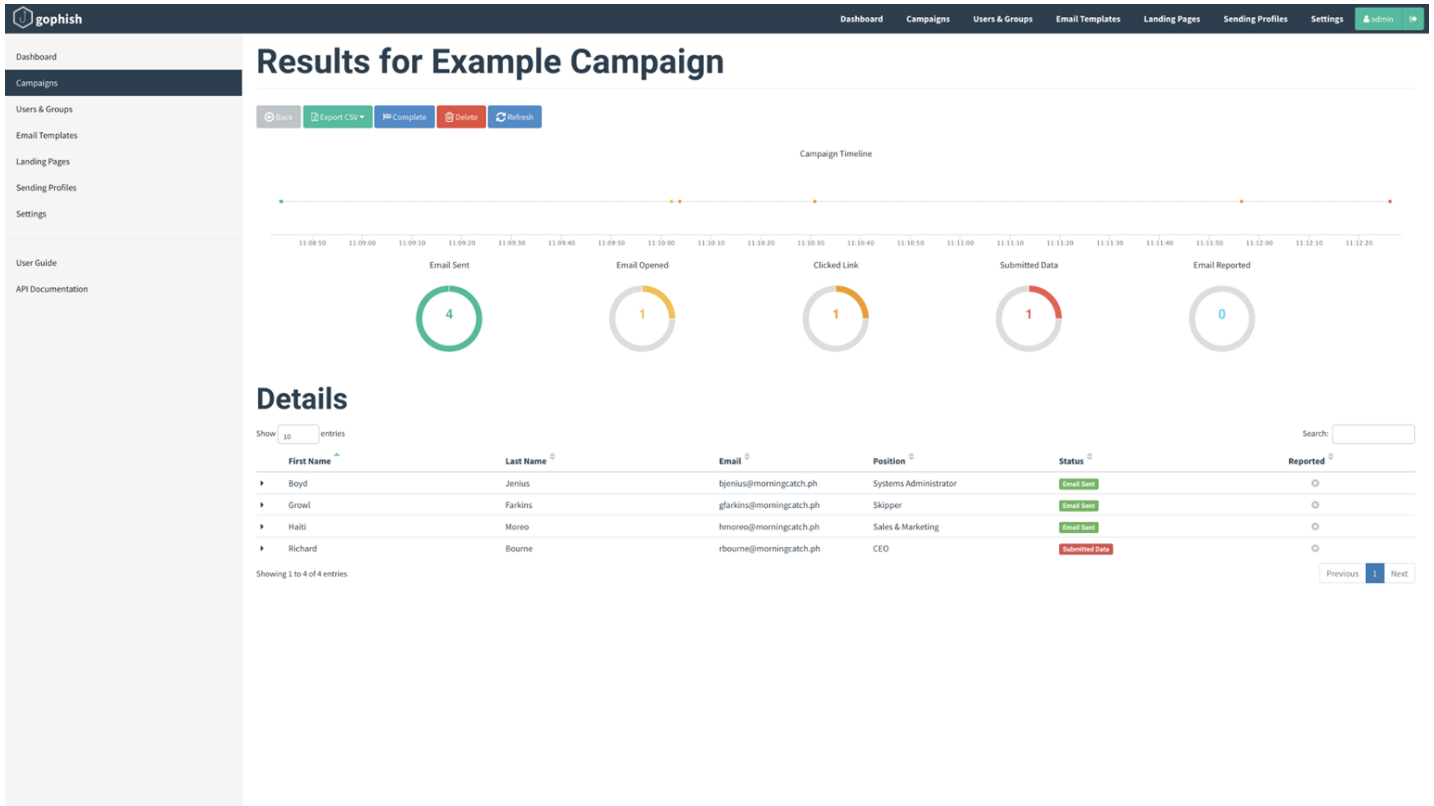
Retraining Red Team operators and updating playbook vs. Tweaking pipeline parameters

Content Generation

Manually writing a phishing email vs. Generating with GPT-3 followed by manual curation and editing

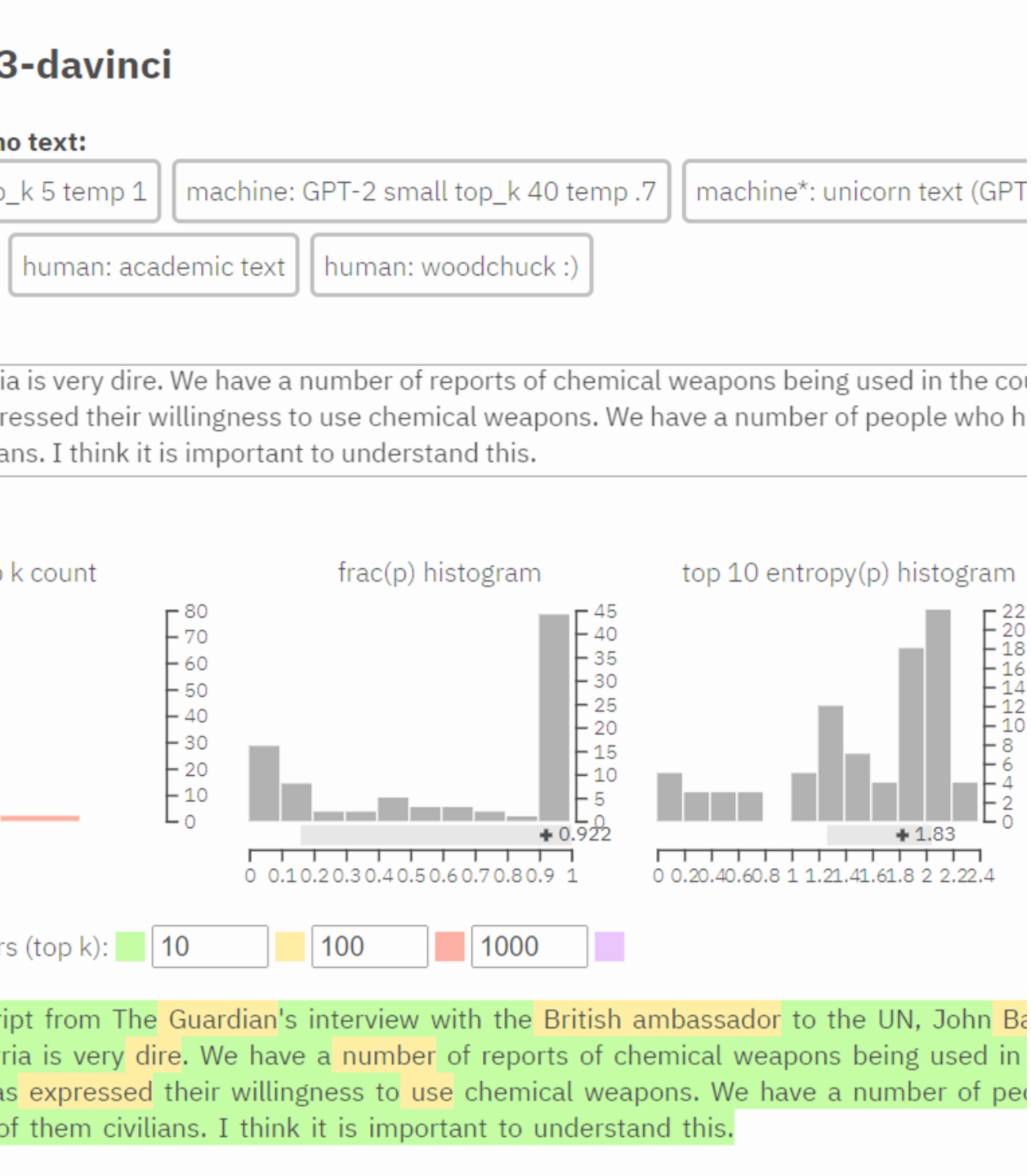
CYBER SECURITY GROUP

Due to the ease of “text in, text out” AIaaS APIs, we can easily integrate them into existing tools.



From Gophish User Guide

CYBER SECURITY GROUP



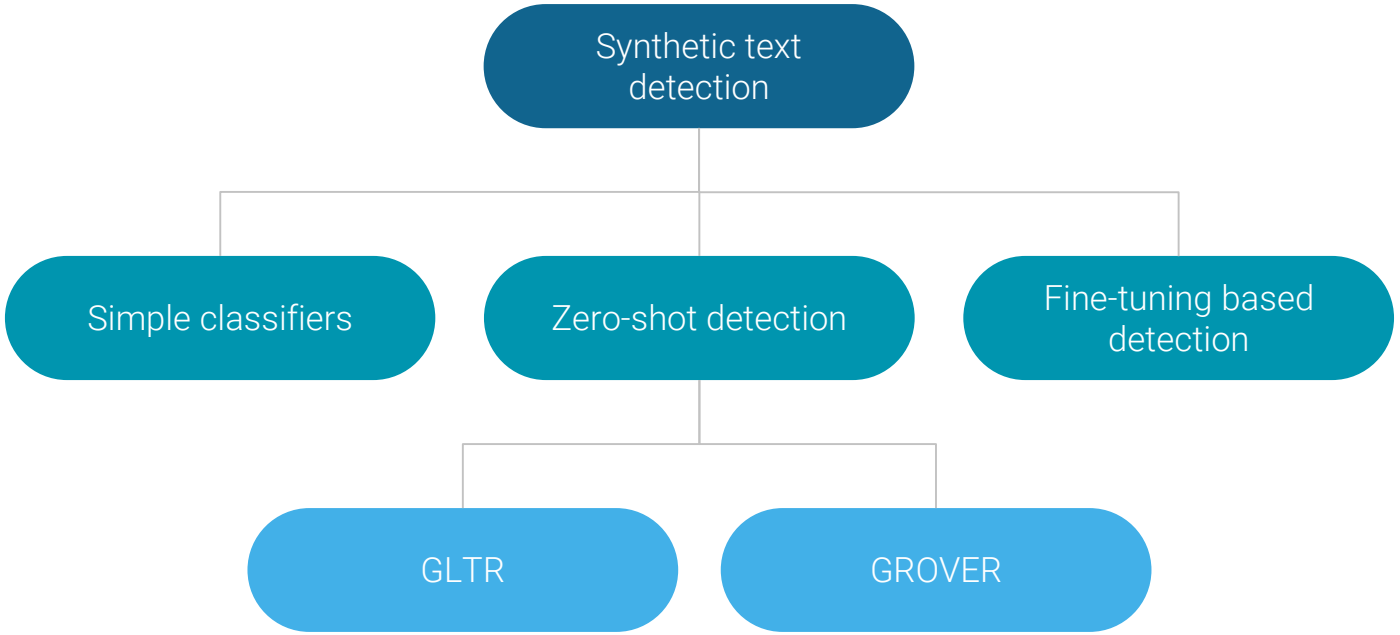
03

Defending against AI Phishing

Automated detection of AI-generated text remains a hard problem.

"We expect that content-based detection of synthetic text is a long-term challenge... this is not high enough accuracy for standalone detection and needs to be paired with metadata-based approaches, human judgment, and public education to be more effective."

– Irene Solaiman, Jack Clark and Miles Brundage, "GPT-2: 1.5B Release," 2019



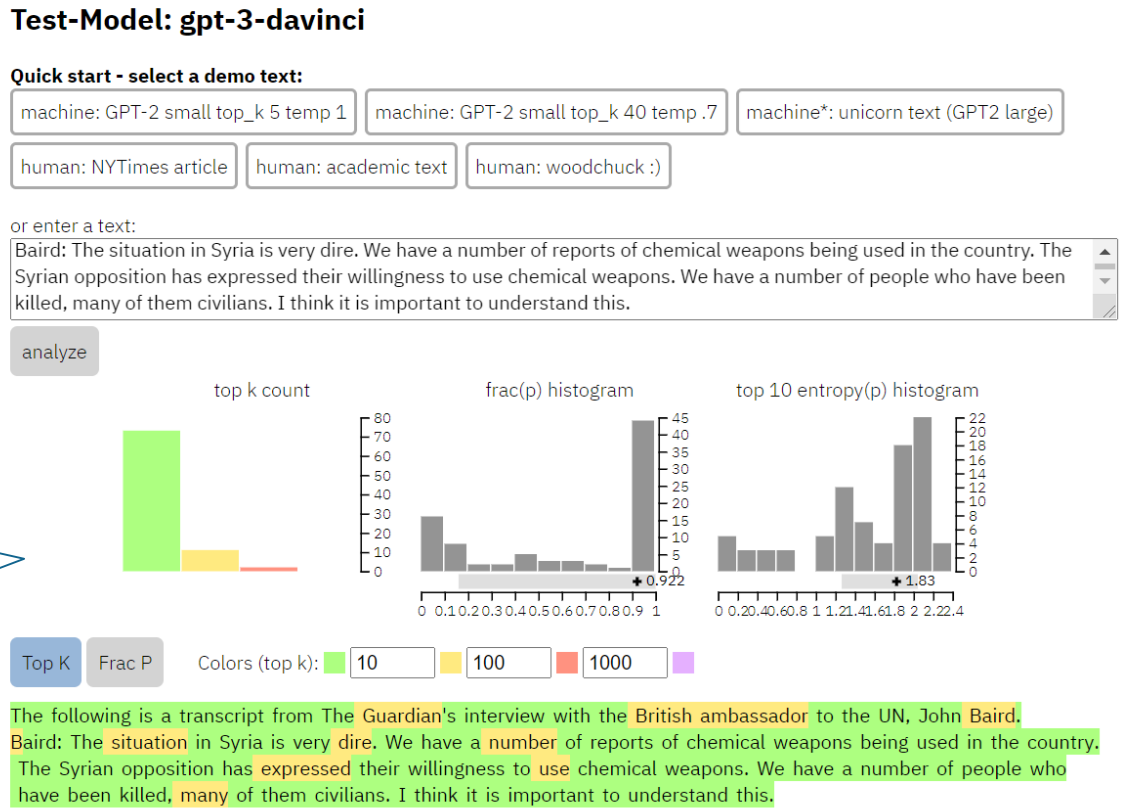
CYBER SECURITY GROUP

The GLTR approach shows promise for AI-assisted human detection of AI-generated text.

AI-assisted human detection using three tests:

1. The probability of the word given the previous words in the sequence.
2. The absolute rank of a word.
3. The entropy of the predicted distribution.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush, “GLTR: Statistical Detection and Visualization of Generated Text,” 2019



CYBER SECURITY GROUP

Given the limitations of the GPT-3 API, we chose to build a zero-shot detector that extends GLTR.

Benefits

- Easily extensible
- Transferrable patterns from GPT-2
- Access to logprobs

Challenges

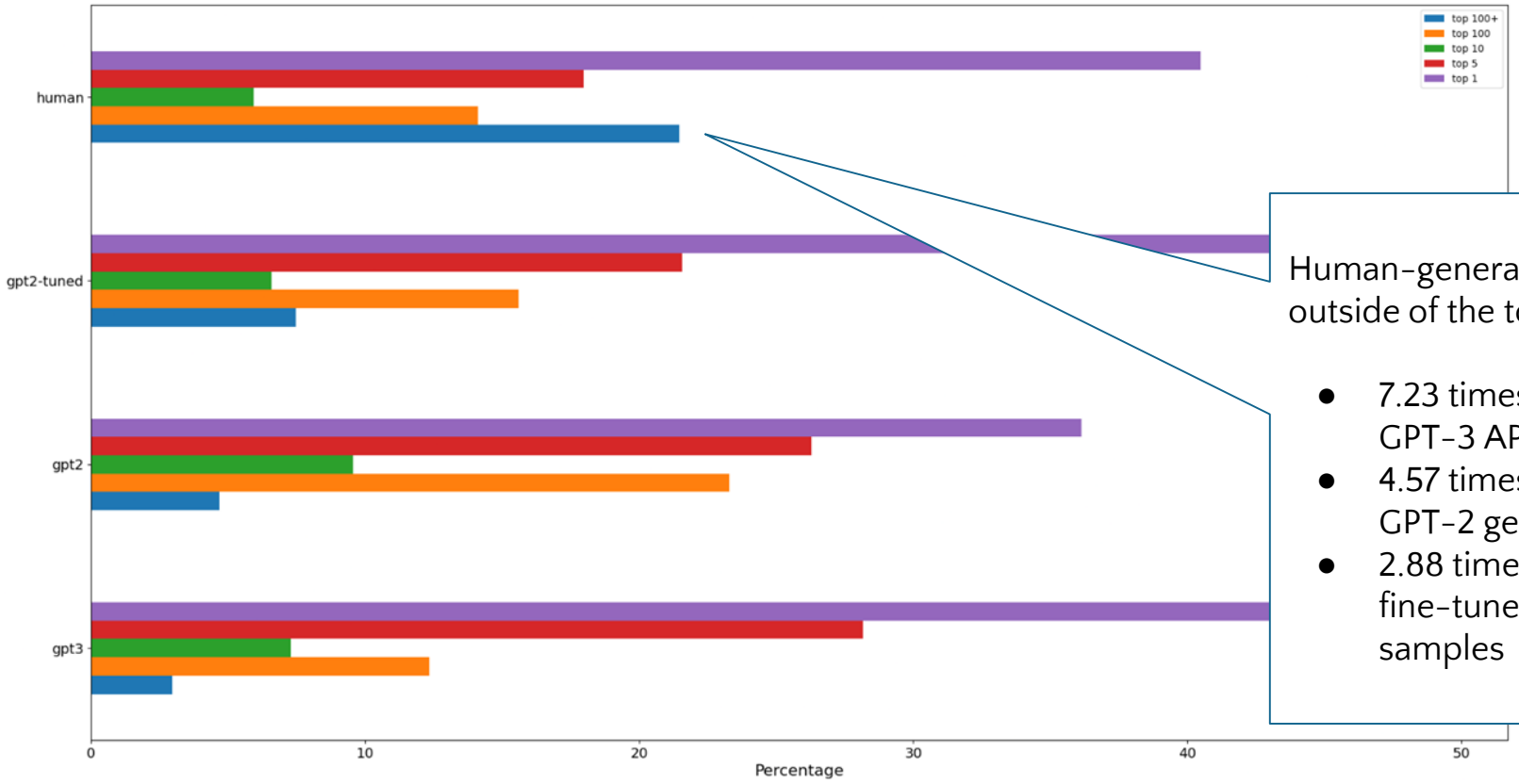
- Cannot control top K
- No direct model access
- Limited number of logprobs (100)

```python
@register_api(name='gpt-3-davinci')
class GPT3LM(AbstractLanguageChecker):
    def __init__(self, model_name_or_path="gpt2"):
        super(GPT3LM, self).__init__()
        self.enc = GPT2Tokenizer.from_pretrained(model_name_or_path)
        self.start_token = '<|endoftext|>'
        print("Loaded GPT-3 model!")

    # Watch this space: Lots of edge cases from GPT-3 API
    def preprocess(self, token):
        # Normalize non-standard unicode
        token = unicodedata.normalize("NFKC", token)

        # Handle strange API byte returns ("bytes:\xe2\x80")
        if token.startswith('bytes:'):
            token = token[6:]

        # Handle whitespace characters not properly encoded by API
        if token == len(token) * " ":
            token = token.replace(" ", "\u0120")
        elif token == len(token) * "\n":
            token = token.replace("\n", "\u010A")
        elif token == len(token) * "\t":
            token = token.replace("\t", "\u0109")

        return token

    def check_probabilities(self, in_text, topk=40):
        # Process input
        encoded_context = self.enc.encode(in_text)
        encoded_context = [self.enc.encoder[self.start_token]] + encoded_context
```

https://github.com/spaceraccoon/detecting-fake-text

CYBER SECURITY GROUP

The GPT-3 API could distinguish between human and AI-generated texts using GLTR metrics.

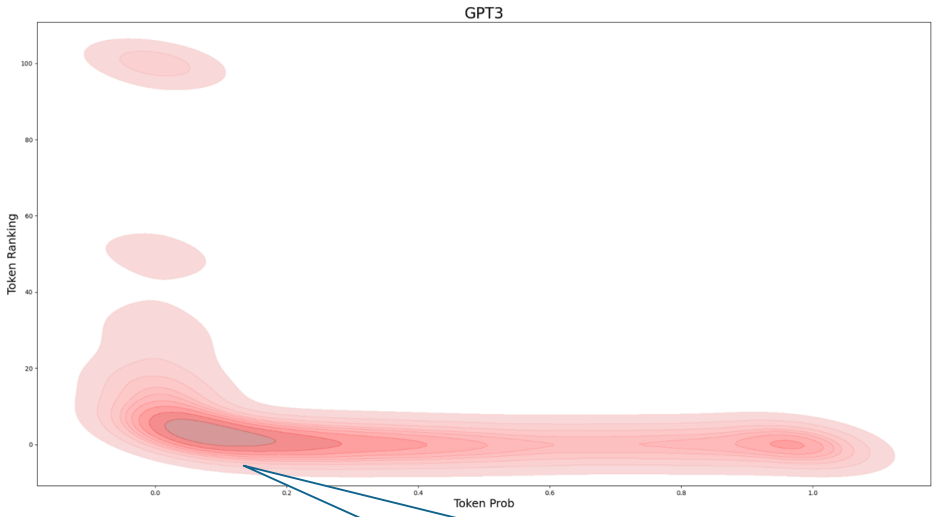


Human-generated samples used words outside of the top 100 predictions:

- 7.23 times as frequently as the GPT-3 API generated samples
- 4.57 times as frequently as the GPT-2 generated samples
- 2.88 times as frequently as the fine-tuned GPT-2 generated samples

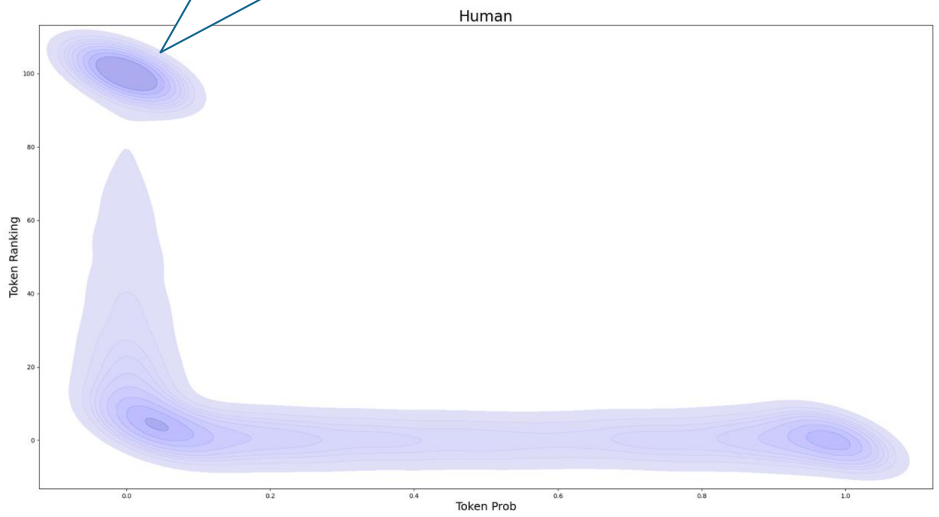
CYBER SECURITY GROUP

The GPT-3 API could distinguish between human and AI-generated texts using GLTR metrics.



AI: Greater density at low rank/low probability

Human: Greater density at high rank/low probability



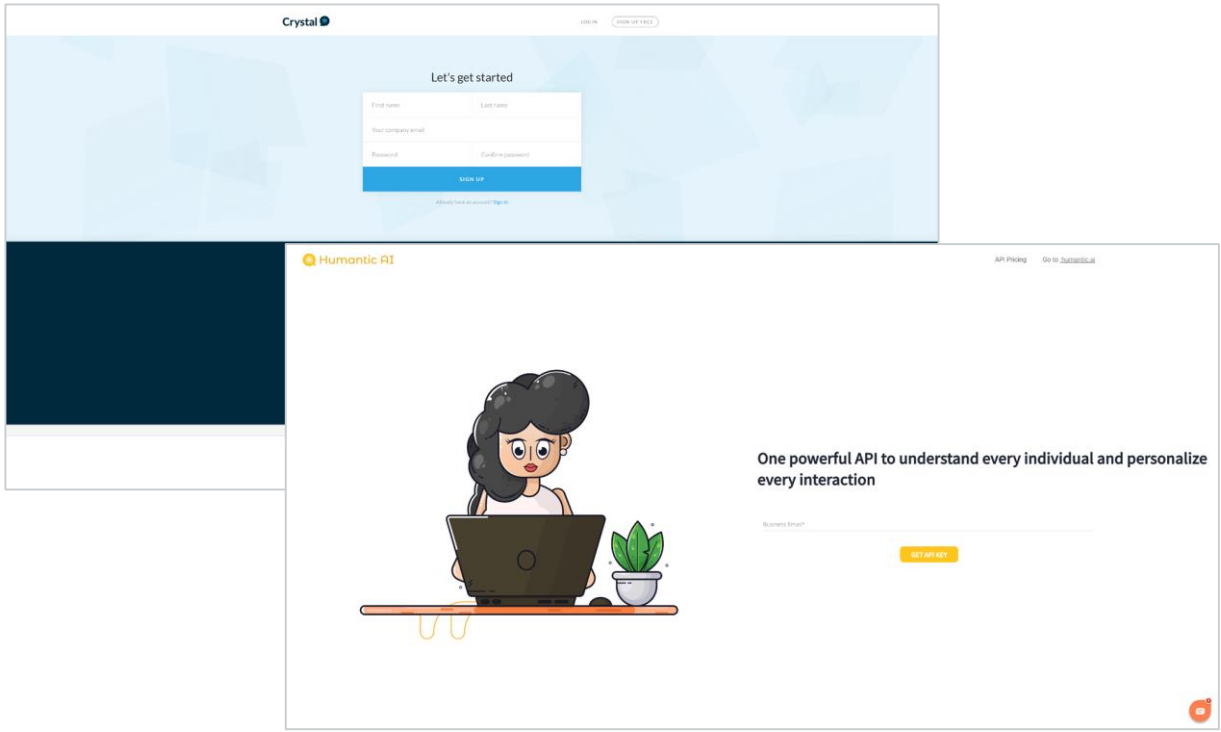
CYBER SECURITY GROUP

OpenAI has strong processes governing the use of the GPT-3 API.

Beta Review	Use Case Guidelines	Pre-Launch Review
● 5 days-6 months ● Tens of thousands of developers ● Needs a good use case	● Disallowed uses ● Risk and safety recommendations ● Fairness and robustness guidelines	● > 300 live applications ● Security and risk mitigations ● Monitoring for potential misuse

However, current and future AIaaS providers may have lower barriers to entry.

- Free demo sign ups
- Instant access to API
- No email verification needed



CYBER SECURITY GROUP

Decision makers must ensure the long-term safety of AI proliferation.

Key Pillars of Singapore's Model AI Governance Framework

Internal Governance Structures and Measures

Determining the Level of Human Involvement

Operations Management

Stakeholder Interaction and Communication

Personal Data Protection Commission, "Model AI Governance Framework," 2020

Both consumers and suppliers of AI solutions must ensure responsible usage.

Key applicable recommendations from Singapore's Model AI Governance Framework

Everyone

- Use Implementation and Self-Assessment Guide for Organizations
- Policy for explanation and practice general disclosure of use
- Ethical evaluation
- Implement clear roles and responsibilities for the ethical deployment of AI

Consumers

- Adopt **"human in the loop"** approach for AI-augmented decision-making

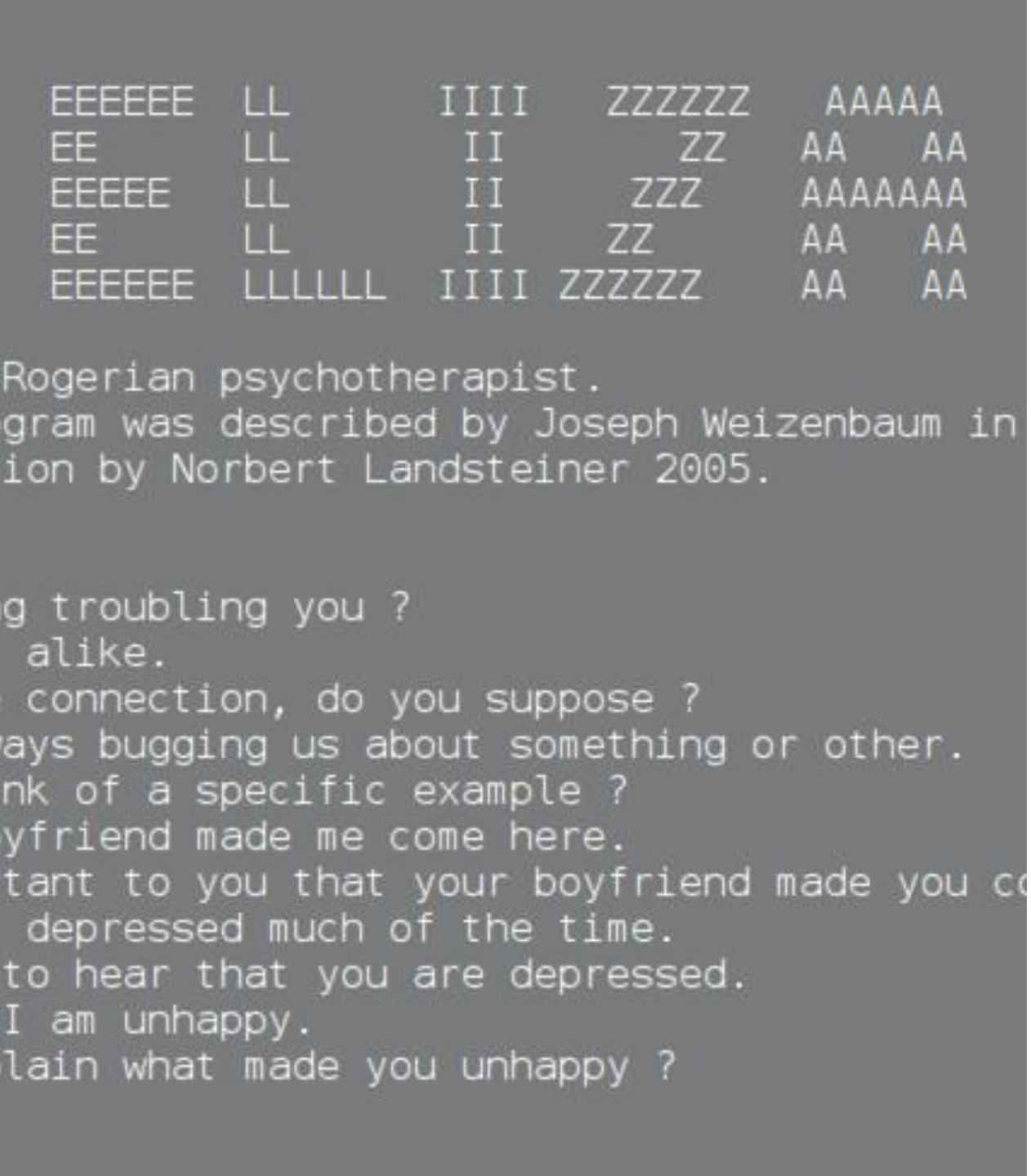
Suppliers

- Ensure **traceability** and **auditability** of use
- Enforce acceptable use policies

Personal Data Protection Commission, "Model AI Governance Framework," 2020

On the individual scale, empower users to resist phishing attacks in an age of AI proliferation.

Security Training	Awareness	Phishing Email Reporter
Behavioural conditioning reduces the susceptible rate after 4 simulations. – PhishMe, "Enterprise Phishing Susceptibility Report," 2016	Guidelines and framework to help users identify common phishing attempts.	Identify phishing emails that bypass the email gateway and prevent similar threats in the future.



04
Conclusion

We are really in the calm before the storm stage of synthetic media's use in information operations...

– Lee Foster, Black Hat USA 2020

The rapid growth of AIaaS has placed advanced, cost-effective AI text generation capabilities in the hands of the global market. These capabilities can be used to accelerate both authorised and malicious phishing campaigns.

While automated tools can be used to build defenses against AI-generated text, current approaches are brittle and model-dependent. AI-assisted human detection of AI-generated text could be more effective.

Decision makers have the responsibility to implement sound strategies governing the supply and use of advanced AIaaS.

THANK YOU

For any enquiries, please contact

www.tech.gov.sg

@GovTechSG

Facebook.com/GovTechSG