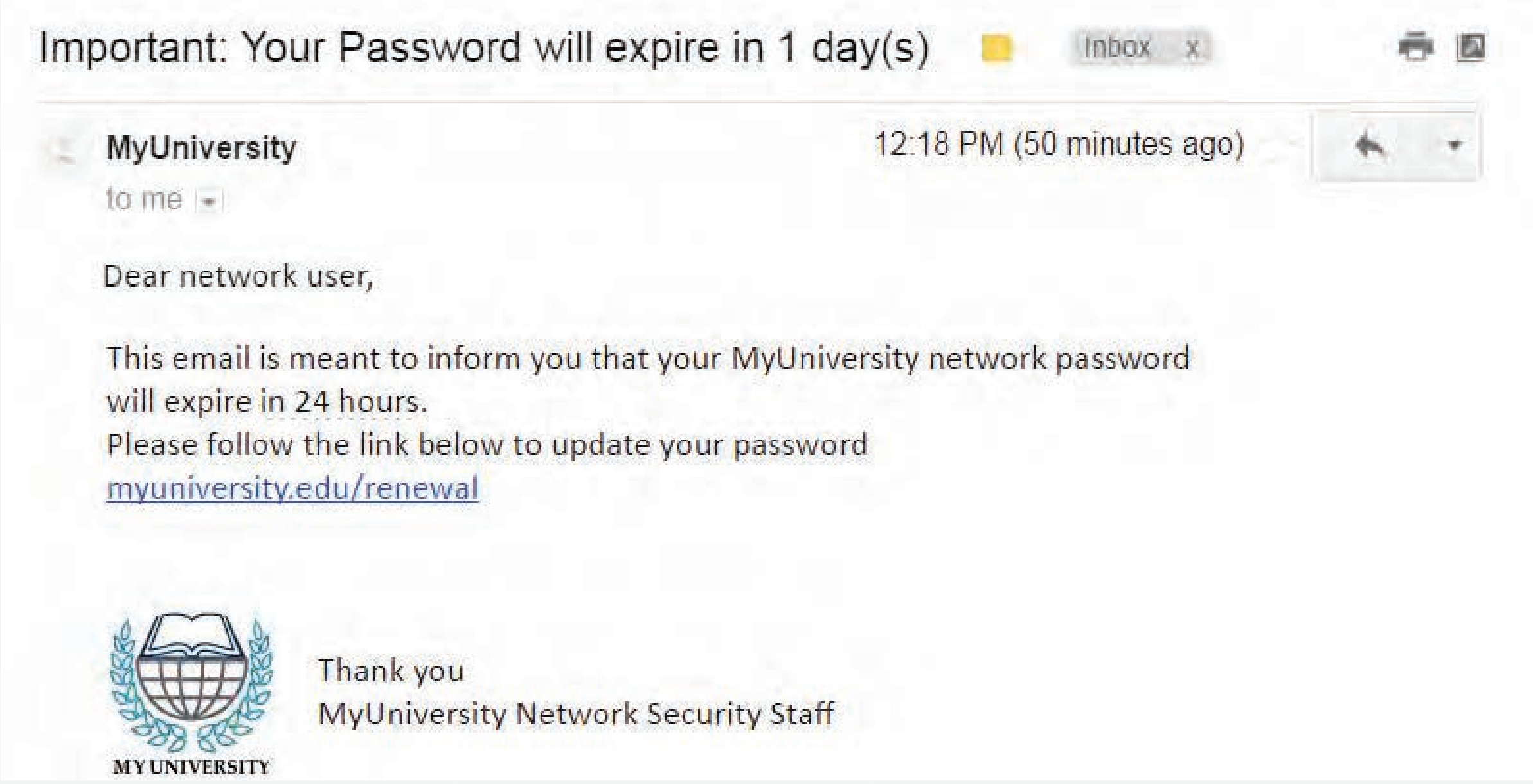
Ethical Implications of AI & Phishing

Emily Herron, Kevin Chen, Will Berger

What is Phishing?

Phishing is a form of hacking that tricks victims into giving up sensitive information (passwords, IDs, documents, etc.) by posing as a legitimate source.



Important: Your Password will expire in 1 day(s)

MyUniversity 12:18 PM (50 minutes ago)
to me

Dear network user,

This email is meant to inform you that your MyUniversity network password will expire in 24 hours.
Please follow the link below to update your password
myuniversity.edu/renewal

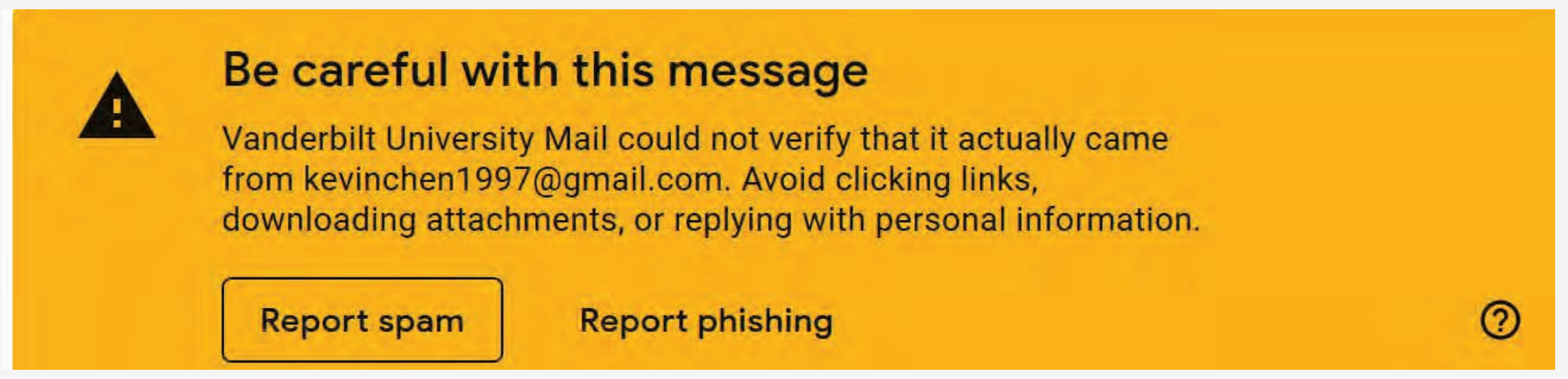
Thank you
MyUniversity Network Security Staff

Figure 1: Phishing example. Source: [1]

Phishing is all about imitation. Phishing emails generally use two techniques: 1) creating a sense of urgency to trick you into clicking on a malicious link, or 2) attempting to 'blend in' as an innocuous email that preys on your curiosity.

How AI Can Protect Us:

Artificially Intelligent systems are being leveraged for detecting phishing attacks using known as well as new features. These systems use tactics like major machine learning classifiers like decision trees (DT), k-nearest neighbors, support vector machines (SVM) and unsupervised k-means clustering [2].



Be careful with this message

Vanderbilt University Mail could not verify that it actually came from kevinchen1997@gmail.com. Avoid clicking links, downloading attachments, or replying with personal information.

Figure 2: Gmail spam flag.

Google claims to use their Machine Learning tool TensorFlow to block an additional 100 million spam emails each day. Google says they already block 99.9% of all spam, but catching the last sliver is only made possible through software that can learn and adapt to new patterns in spam email [3].

How AI Can be Used Against Us:

Artificially Intelligent systems are being leveraged by phishers to make more advanced and targeted phishing attacks. This is done by harnessing artificially intelligent systems to gather personal information that can be used for formulated phishing attacks [4].

Artificially intelligent systems can be used to make a machine capable of mimicking human behaviors like decision-making and problem solving, tricking people do things like clicking on malicious links. An experiment by ZeroFox, showed that machines are much better than humans at getting people to click on malicious links [5].

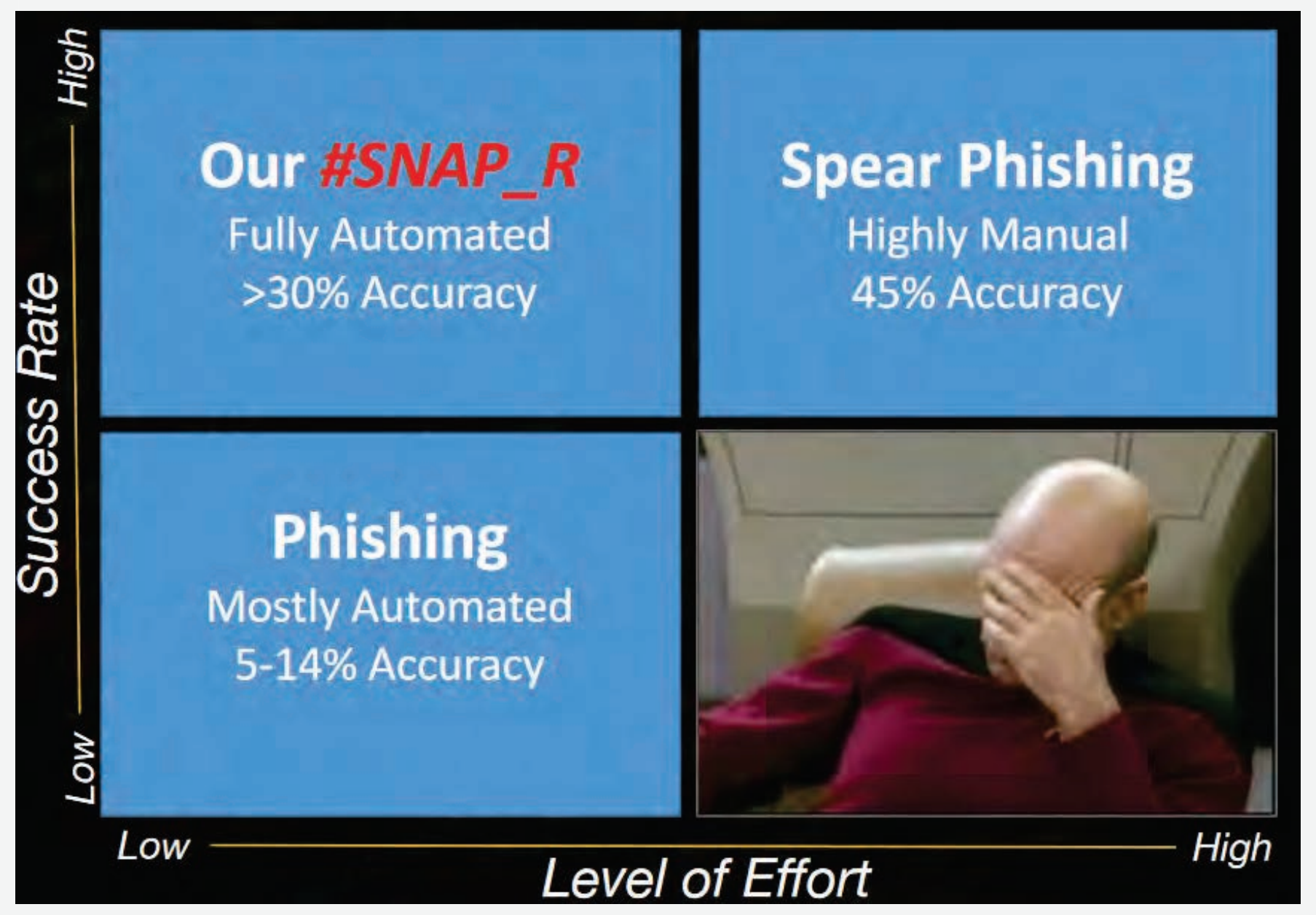
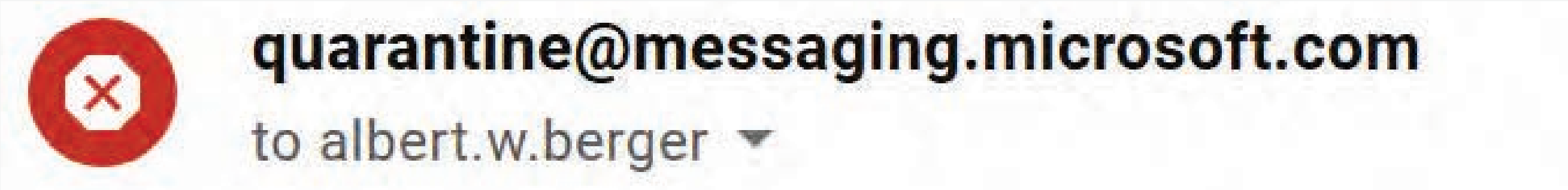


Figure 3: Phishing algorithm graph. Source: [5]

Vanderbilt's Email Quarantine:

After interviewing Vanderbilt's IT department, they stated that Vanderbilt uses both Gmail's and Microsoft Office 365's third-party proprietary spam filters to eliminate spam. Both of these services use "connection filtering" to scan and delete obvious spam and phishing messages based on the IP address of the sender [6].



quarantine@messaging.microsoft.com

to albert.w.berger

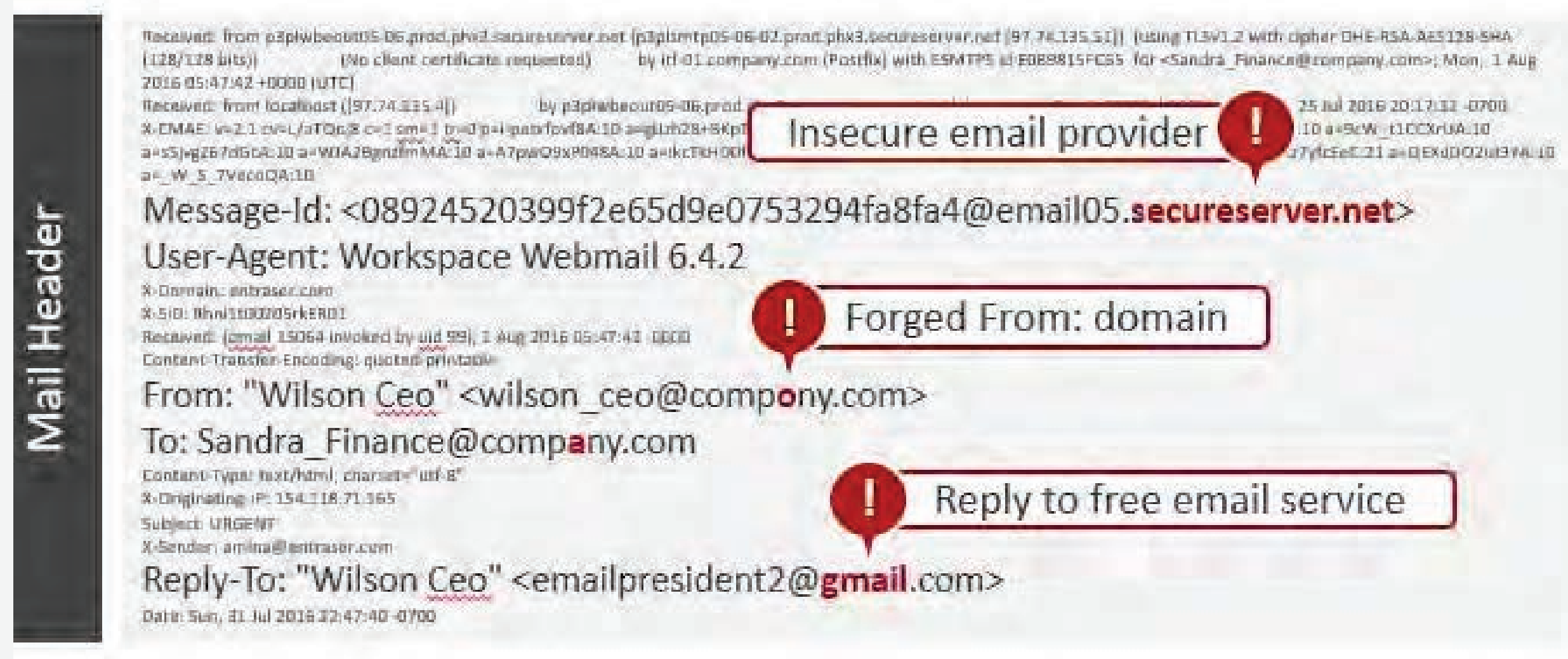
Figure 4: Email from the Vanderbilt quarantine.

Then, these services inspect the subject line and the email's contents to determine if the email is to be released into the user's inbox or to be quarantined for 15 days and then deleted. The quarantine is not accessible via students' Gmail accounts and requires a separate log-in from a portal listed on the Vanderbilt website.

However, many job interview invitations were marked as spam and sent to the quarantine, effectively locking them away without alerting students. When Vanderbilt added Microsoft's quarantine spam filter in the fall of 2018, Vanderbilt may have actually done more harm to students and faculty than good by inadvertently increasing the false positive rate (missing important emails) at the expense of trying to decrease spam entering your inbox.

How to Protect Yourself:

- Always remember to question emails, and if something seems off, there's probably a reason.
- Check for spelling and grammatical errors in the email address of the sender, the subject line, and the actual content of the email.
- Hover over (but don't click on) hyperlinks that look suspicious to see where they lead.
- Install a phishing filter on your email application and also on your web browser. These filters don't keep out all phishing messages, but they do reduce the number of phishing attempts [7].



Message-Id: <08924520399f2e65d9e0753294fa8fa4@email05.secureserver.net>
User-Agent: Workspace Webmail 6.4.2
From: "Wilson Ceo" <wilson_ceo@compony.com>
To: Sandra_Finance@company.com
Subject: URGENT
Reply-To: "Wilson Ceo" <emailpresident2@gmail.com>
Date: Sun, 31 Jul 2016 22:47:40 -0700

Figure 5: Flagging potential phishing attacks. Source: [7]

Ethical Implications:

AI on both sides of the phishing battle has effectively created an AI arms race. The tools and techniques we've created to make our lives better have also been turned against us. Now, both sides have to stay on the cutting edge of the technology to achieve their desired goals.

Designers of these algorithms need to make trade-offs between flagging false positives as spam or potentially missing true phishing attacks.

Acknowledgments:

[1] "What is phishing | Attack techniques & scam examples | Imperva," Learning Center. [Online]. Available: https://www.imperva.com/learn/application-security/phishing-attack-scam/. [Accessed: 14-Apr-2019].

[2] N. Ghazim.jameel and L. E. George, "Detection of Phishing Emails using Feed Forward Neural Network," International Journal of Computer Applications, vol. 77, no. 7, pp. 10–15, 2013.

[3] J. Vincent, "Gmail is now blocking 100 million extra spam messages every day with AI," The Verge, 06-Feb-2019. [Online]. Available: https://www.theverge.com/2019/2/6/18213453/gmail-tensorflow-machine-learning-spam-100-million. [Accessed: 14-Apr-2019].

[4] E. Benishti, "Artificial Intelligence is Revolutionizing Phishing – and It's Not All Good," Ironscales, 21-Mar-2018. [Online]. Available: https://ironscales.com/blog/Artificial-Intelligence-Revolutionizing-Phishing/. [Accessed: 10-Apr-2019].

[5] J. Seymour and P. Tully, "Automated E2E Spear Phishing on Twitter," Black Hat. [Online]. Available: https://www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter.pdf. [Accessed: 10-Apr-2019].

[6] KCCross, "Anti-spam protection FAQ," Microsoft Docs. [Online]. Available: https://docs.microsoft.com/en-us/office365/SecurityCompliance/anti-spam-protection-faq. [Accessed: 14-Apr-2019].

[7] J. Huang, "Beyond Catching Sender Spoofing – using AI to stop email fraud and Business Email Compromise," Simply Security News, Views and Opinions from Trend Micro, Inc, 06-Dec-2018. [Online]. Available: https://blog.trendmicro.com/beyond-catching-sender-spoofing-using-ai-stop-email-fraud-business-email-compromise/.